ENAR

ENAR 2020 SPRING MEETING

WITH IMS & SECTIONS OF ASA
March 22-25, 2020
JW Marriott Nashville
Nashville, TN

FINAL PROGRAM AND SCIENTIFIC PROGRAM

WELCOME

I am thrilled to welcome you to the ENAR 2020 Spring Meeting in Nashville! I would like to extend a special welcome to our first-time attendees and look forward to your return for future meetings and your future involvement in ENAR!

The ENAR 2020 Spring Meeting will be held at the JW Marriott Nashville. Nashville's foundation was built on music. It is the common thread connecting the life and soul of the city and its people. Nashville is home to the Grand Ole Opry, Ryman Auditorium, Country Music Hall of Fame, and many honkey tonks where you can enjoy great music, food and drink.

The four-day meeting, **March 22-25, 2020,** will host biostatistics students, researchers, and practitioners, from academia, government, and industry. The meeting will expose attendees to the latest developments in methods, software, and applications through the Scientific and Educational programs. The meetings also provide a great opportunity for professional networking, meeting new people, connecting job seekers with employers, and reconnecting with friends and colleagues. Our exhibitors and vendors will give you opportunities to check out the latest textbooks and see demonstrations of new software.

ENAR is committed to fostering a culture of inclusion, professionalism and civil discourse that cultivates an environment where ideas are exchanged openly and freely with mutual respect and trust. ENAR has adopted a **Meeting Conduct Policy** intended to guide all attendees at ENAR's annual Spring Meeting and attendees will be required to assent to the policy as part of registration.

The ENAR Spring Meeting is only possible through the efforts of many hard-working volunteers. Thanks to all of the volunteers for helping make the ENAR Spring Meeting a success!

SCIENTIFIC PROGRAM

Through the leadership of the Program Chair Juned Siddique (Northwestern University) and Associate Program Chair Chenguang Wang (Johns Hopkins University), and contributions from many of you, the Program Committee (consisting of 10 ASA section representatives and 4 at-large ENAR members) has assembled a diverse and exciting invited program. The sessions cover a wide range of topics, including modern graphical modeling, complex innovative clinical trial design, electronic health records data, machine learning, neuroimaging, wearable/mobile technology, data integration, causal inference, survival outcomes, spatial modeling, environmental health, and statistical modeling in Alzheimer's disease. The IMS Program Chair Sunduz Keles (University of Wisconsin, Madison) has also put together complementary sessions on causal inference with genetic data, statistical methods for single-cell omics analysis, microbiome data analysis, precision medicine, and asymmetrical statistical learning.

Poster sessions play a prominent role at the ENAR Spring Meeting, and continue to be a vital part of the program. In addition to contributed and invited posters, the ENAR 2020 Spring Meeting will continue contributed SPEED poster sessions, in which presenters give a two-minute elevator speech on the highlights of their posters. As in 2019, these speed sessions will utilize digital poster boards, giving presenters the opportunity for more interactive posters. Monday, March 23rd will feature the thematically grouped contributed speed poster sessions. Each session will feature two invited posters from well-known researchers and will run parallel with the rest of the sessions in the scientific program. As in previous years, the regular contributed posters will be featured during the Opening Mixer on Sunday evening. This year, poster presenters will be assigned one-hour slots to be available at their poster, giving everyone a chance to view the amazing research on display. Posters in this session will be eligible to win an award as part of the popular ENAR Regional Advisory Board's poster competition!

EDUCATIONAL PROGRAM

Our educational program provides many opportunities to learn new statistical techniques, to develop new computational skills, and to discuss the latest research or career development skills with leading experts. The Educational Advisory Committee has assembled an engaging suite of short courses, tutorials and roundtables covering a wide range of topics from renowned instructors.

Short course topics include design and analysis in platform & basket trials and in SMART, multivariate meta-analysis, using NIMBLE for MCMC, working with electronic health records (EHRs), implementing Bayesian adaptive designs, and statistical networks in biology. Tutorial topics include disease risk modelling and causal inference in R, methods for geometric functional data, difference in difference studies, integrating 'omics and imaging data, and creating R packages. Roundtable luncheons provide a more focused discussion with distinguished statisticians from academia, government and industry. Topics range from reviewing and writing grants, working in government and a medical school, publishing and reviewing manuscripts, data science, and mentoring. Be sure to take a look and sign up for something interesting!

I would like to extend a special thanks to the members of the Educational Advisory Committee – Lynn Eberly (University of Minnesota), Jason Roy (Rutgers), Veera Baladandayuthapani (University of Michigan), and Haoda Fu (Eli Lilly) for their support and guidance in helping to put together an outstanding educational program.

PRESIDENTIAL INVITED ADDRESS

I am thrilled to announce that the 2020 ENAR Presidential Invited Address will be given by **Dr. Sharon-Lise Normand.** Dr. Normand is a statistician whose work has made impactful contributions to health services and regulatory policy, particularly in the areas of cardiovascular disease and mental health. Methodologically, these contributions have been accomplished via Bayesian hierarchical models and Bayesian approaches for causal inference. Her contributions have been recognized in the statistics community (ASA fellow), the medical community (American College of Cardiology fellow), and the broader scientific community (AAAS fellow).

To learn more about Dr. Normand and her Invited Address, please see page 11.

ADDITIONAL MEETING ACTIVITIES

The ENAR 2020 Spring Meeting will feature several other activities in addition to the scientific and educational programs. On Sunday, March 22nd, there will be the **Fostering Diversity in Biostatistics Workshop,** organized by Felicia R. Simpson (Winston-Salem State University) and Loni Philip Tabb, (Drexel University). Dr. Adrian Coles (Eli Lilly) will serve as this year's keynote speaker. This workshop has been very popular and impactful and registration typically fills up quickly. Please be sure to register early if you are interested in attending!

Students, recent graduates, and other young professionals should plan to attend the Networking Mixer on Monday evening and the Tuesday luncheon event organized by the **Council for Emerging and New Statisticians (CENS).** These are great opportunities for our "younger" members to meet new people, learn about CENS and become more engaged with ENAR. Attendees seeking employment and prospective employers have the opportunity to connect via the **Career Placement Center.**

Tuesday evening will feature our second annual **ENAR Sponsor and Exhibitor Mixer.** You will be able to peruse the latest books and software while joining the sponsors and exhibitors for the reception in the exhibition area after the last Tuesday session. It will be a great opportunity to catch up with friends, collaborators, and colleagues. After the mixer, you will have time to walk around and dine in a great Nashville restaurant. The Local Arrangements Committee, chaired by Cindy Chen (Vanderbilt University), will provide recommendations for attendees.

We hope to see you in Nashville for the 2020 ENAR Spring Meeting!

Mike Daniels,
ENAR 2020 President

Advancing biological and life science through the development of quantitative theories and the application, development and dissemination of effective mathematical and statistical techniques.

TABLE OF CONTENTS

ENAR 2020 Spring Meeting
With IMS & Sections of ASA
March 22-25 | JW Marriott Nashville | Nashville, TN

ENAR

ACKNOWLEDGEMENTS

ENAR would like to acknowledge the generous support of the 2020 Local Arrangements Committee, chaired by Cindy Chen and our student volunteers.

Executive Committee – Officers

President	Michael Daniels
Past President	Sarah Ratcliffe
President-Elect	Brent Coull
Secretary (2019-2020)	Brisa Sánchez
Treasurer (2019-2020)	Renee Moore

Regional Committee (RECOM)

President (Chair):	Michael Daniels
RAB Chair:	Sarah Ratcliffe

Nine ordinary members (elected to 3-year terms):

2018-2020

Babette Brumback
Dean Follmann
Laura Hatfield

2020-2022

Emma Benn
Paul Rathouz
Pamela Shaw

2019-2021

Lynn Eberly
Peter Song
Alisa Stephens-Shields

Regional Members of the International Biometric Society Executive Board

Karen Bandeen-Roche, Joel Greenhouse and José Pinheiro

Regional Members of The Council of the International Biometric Society

Paul Alberts, Nandita Mitra, Dionne Price and Brisa Sánchez

Appointed Members of Regional Advisory Board (3-Year Terms)

Chair: Leslie McClure

2018-2020

Naomi Brownstein
Jan Hannig
Eric Lock
Mark Meyer
Taki Shinohara
Daniela Stores-Alvarez
Peng Wei
Zhenke Wu
Lin Zhang
Lili Zhao

2019-2021

Ashley Buchanan
Emily Butler
Ani Eloyan
Tanya Garcia
Joseph Kang
Sung Duk Kim
Benjamin Risk
Ana-Maria Staicu
Sameera Wijayawardana
Yize Zhao

2020-2022

Joey Antonelli
Arkendu Chatterjee
Lorin Crawford
Samson Ghebremariam
Ana Ortega-Villa
Harrison Quick
Sandra Safo
Briana Stephenson
Dandan Xu
Ni Zhao

CENS – Council for Emerging and New Statisticians

RAB Liaison: Elizabeth Handorf, Fox Chase Cancer Center

Steering Committee:

Alessandra Valcarcel, University of Pennsylvania
Donnie Herbert, The Emmes Company LLC
Alex Kaizer, University of Colorado
Jing Li, Eli Lilly and Company
Bryan Blette, University of North Carolina at Chapel Hill
Will Eagan, Purdue University
Fan (Frank) Li, Yale University
Nancy Murray, Emory University
Hannah Weeks, Vanderbilt University
Emily Zabor, Cleveland Clinic

PROGRAMS

2020 Spring Meeting—Nashville

Program Chair: Juned Siddique

Program Associate Chair: Chenguang Wang

2021 Spring Meeting—Baltimore

Program Chair: Howard Chang

Associate Program Chair: Yize Zhao

Digital Organizing Chair: Joshua Lukemire

2020 Joint Statistical Meeting

Jeremy Gaskins

Biometrics Executive Editor

Marie Davidian

Biometrics Co-Editors

Yi-Hau Chen
Michael J. Daniels
Jeanine Houwing-Duistermaat

Biometric Bulletin Editor

Ajit Sahai

JABES Editor

Brian Reich

ENAR Correspondent for the Biometric Bulletin

Jarcy Zee

ENAR Executive Director

Shannon Taylor

International Biometric Society Executive Director

Peter Doherty

Representatives

Committee of Presidents of Statistical Societies (COPSS)

ENAR Representatives

President: Michael J. Daniels

President-Elect: Brent Coull

Past President: Sarah Ratcliffe

ENAR Standing/Continuing Committees

Nominating Committee (2019)

Chair: Jeffrey S. Morris

Scarlett L. Bellamy, Drexel University

Babette Brumback

Leslie McClure

Jeff Goldsmith, Columbia University

Mary Sammel, University of Pennsylvania

ENAR Webinar Committee (2019)

Sameera Wijayawardana, Eli Lilly

Lili Zhao, University of Michigan

ENAR Social Media Committee

Rachel Carroll, need affiliation

Joe Koopmeiners, need affiliation

2020 Fostering Diversity in Biostatistics Workshop

Felicia R. Simpson (Co-Chair), Winston-Salem State University

Loni Philip Tabb, Drexel University

Danisha Baker, Naval Nuclear Lab

Scarlett Bellamy, Drexel University

Emma Benn, Mount Sinai

Portia Exum, SAS Institute Inc.

Vladimir Geneus, Eli Lilly and Company

Justine Herrera, Columbia University

Miguel Marino, Oregon Health & Science University

Renee Moore, Emory University

Knashawn Morales, University of Pennsylvania

Distinguished Student Paper Awards Committee

Chair: Jeffrey S. Morris

Dipankar Bandyopadhyay, Virginia Commonwealth University

Howard Chang, Emory University

Shuo Chen, University of Maryland

Haitao Chu, University of Minnesota

Ying Guo, Emory University

Brian Hobbs, University of Texas, Austin

Jian Kang, University of Michigan

Robert Krafty, University of Pittsburgh

Jenna Krall, George Mason University

Hulin Li, New York University Langone Health

Eric Lock, University of Minnesota

Wenbin Lu, North Carolina State University

Nandita Mitra, University of Pennsylvania

Hernando Ombao, King Abdullah University of Science of Technology

Xiao Son, University of Georgia

Andrew Spieker, Vanderbilt University Medical Center

Hong Tian, Johnson and Johnson

Xiaofei Wang, Duke University

Zhenke Wu, University of Michigan

Sharon Xie, University of Pennsylvania

Fei Zou, University of North Carolina

Ni Zhao, Johns Hopkins University

Van Ryzin Award Winner:

Zhe Sun, University of Connecticut

Distinguished Student Paper Award Winners

Rebecca Anthopolos, New York University

Lin Dong, Wells Fargo Bank

Teng Fei, Emory University

Guillermo Granados, King Abdullah University of Science and Technology

Yichen Jia, University of Pittsburgh

Yunchua Kong, Emory University

Yan Li, University of Connecticut

Wodan Ling, Fred Hutchinson Cancer Research Center

Yusha Liu, Rice University

Peng Liu, University of Pittsburgh

Lan Luo, University of Wisconsin, Madison

Xin Ma, Emory University

Stefani Mokalled, Clemson University

Arman Oganisian, University of Pennsylvania

Chan Park, University of Wisconsin, Madison

Bo Wei, Emory University

Guangyu Yang, University of Michigan, Ann Arbor

Bingxin Zhao, University of North Carolina at Chapel Hill

Jincheng Zhou, Amgen

Please visit the ENAR website (**www.enar.org**) as a resource of information on all ENAR activities.

We gratefully acknowledge the invaluable support and generosity of our Sponsors and Exhibitors.

SPONSORS

Abbvie
Emmes
Emory University
Drexel University
Incyte
Ely Lilly and Company
University of Michigan
University of Minnesota
SAS
StataCorp LLP
University of North Carolina Biostatistics Department
Vanderbilt University
Proctor & Gamble
Janssen, J&J
Takeda

EXHIBITORS

CENS
FDA
Springer
SAS
Stata
The Lotus Group
IBS

SPECIAL THANKS

Program Chair
Juned Siddique, Northwestern University

Associate Program Chair
Chenguang Wang, Johns Hopkins University

IMS Program Chair
Sunduz Keles, University of Wisconsin, Madison

Digital Program Coordinator
David Aaby, Northwestern University

Local Arrangements Chair
Cindy Chen, Vanderbilt University

ASA Section Representatives – Program Committee

Veronica Berrocal, University of California at Irvine
ASA Statistics in Epidemiology Section

Hongyuan Cao, University of Missouri
ASA Mental Health Statistics Section

Susmita Datta, University of Florida
ASA Statistics in Genomics and Genetics Section

Stephine Keeton, PPD
ASA Biopharmaceutical Statistics Section

Lei Liu, Washington University (St. Louis)
ASA Health Policy Statistics Section

Alexandra Schmidt, McGill University
ASA Statistics & the Environment Section

Anuj Srivastava, Florida State University
ASA Statistics in Imaging Section

Donatello Telesca, UCLA
ASA Section on Bayesian Statistics

Zheyu Wang, Johns Hopkins University
ASA Biometrics Section

Ying Wei, Columbia University,
ASA Statistical Learning and Data Mining Section

ENAR At-Large Members – Program Committee

Lauren Balmert, Northwestern University

Hwanhee Hong, Duke University

Jonathan Schildcrout, Vanderbilt University

Cory Zigler, University of Texas at Austin

ENAR Executive Committee

Michael Daniels, President

Brent Coull, President-Elect

Sarah Ratcliffe, Past President

Brisa Sánchez, Secretary

Renee Moore, Treasurer

RECOM Chair: Michael Daniels

RAB Chair: Leslie McClure

Educational Advisory Committee

Veera Baladandayuthapani, University of Michigan

Lynn Eberly, University of Minnesota

Haoda Fu, Eli Lilly

Jason Roy, Rutgers University

2020 ENAR Student Awards

Jeffrey S. Morris, University of Pennsylvania

ENAR Fostering Diversity in Biostatistics Workshop

Felicia R. Simpson, Winston-Salem State University

Loni Philip Tabb, Drexel University

ENAR Staff

Shannon Taylor, Executive Director

Leah Sibilia, Event Director

Tayler Kenney, Event Specialist

Amber Darcy, Event Specialist

Laura Stapleton, Administrative/Membership Manager

PRESIDENTIAL INVITED SPEAKER

Sharon-Lise Normand, Ph.D.

S. James Adelstein Professor of Health Care Policy (Biostatistics)
Department of Health Care Policy, Harvard Medical School
Department of Biostatistics, Harvard T.H. Chan School of Public Health

MEDICAL PRODUCT, HEALTHCARE DELIVERY, AND ROAD SAFETY POLICIES: SEEMINGLY UNRELATED REGULATORY QUESTIONS

The evaluations of medical product effectiveness and safety, the quality of hospital care, and the safety of U.S. roadways involve the use of large, complex observational data to make policy decisions. Careful design and analysis of such data are critical given the large populations impacted. While increasing access to data of increased size and type permit, in theory, richer evaluations, study design should assume a more prominent role. This talk will describe three different policy problems: the impact of the hospital readmission reduction program, the effectiveness of seemingly similar drug eluting coronary stents, and the safety of U.S. motor carriers. Statistical issues common across these problems, including clustered data, multiple treatments, multiple outcomes, high-dimensional data, and lack of randomization, are highlighted and solutions discussed.

BIOGRAPHY

Sharon-Lise Normand is the S. James Adelstein Professor of Health Care Policy (biostatistics) in the Department of Health Care Policy at Harvard Medical School and in the Department of Biostatistics at the Harvard Chan School of Public Health. Dr. Normand earned her BSc (1984) and MSc (1985) degrees in statistics from the University of Western Ontario and her PhD (1990) in biostatistics from the University of Toronto. Dr. Normand's research focuses on the development of statistical methods for health services and regulatory policy research, primarily using Bayesian and causal inference approaches, including assessment of quality of health care, provider profiling, diffusion of medical technologies, and regulatory science. She has developed a long line of research on methods for the analysis of patterns of treatment and quality of care for patients with cardiovascular disease and with mental disorders in particular.

Dr. Normand has developed analytical approaches for comparing hospitals and physicians using outcomes and process-based measures. Since 2002, she served as director of Mass-DAC, the data-coordinating center responsible for collecting, analyzing, and reporting on the quality of care for adults discharged following a cardiac procedure from all non-federal hospitals in Massachusetts. She is serves as the director of the Medical Device Epidemiology Network (MDEpiNet) Methodology Center, a public-private partnership aimed at medical device evaluation. MDEpiNet partners with the FDA's Center for Device and Radiological Health and the Science and Infrastructure Center at Weill Cornell Medical School. Her focus is on the development of statistical approaches to active medical device surveillance, valid inferences from distributed networks, and the improvement of causal inference in the presence of high dimensional data.

On the mental health side, Dr. Normand is leading an NIMH-funded study to estimate the value of publicly funded mental health care for patients with serious mental illness. She is also undertaking an observational study to estimate causal dose "outcomes" curves in the context of understanding weight gain associated with cumulative antipsychotic drug exposure among subjects with schizophrenia for numerous different antipsychotics.

Dr. Normand was elected fellow of the American Statistical Association, fellow of the American Association for the Advancement of Science, fellow of the American College of Cardiology, and Associate Member of the Society of Thoracic Surgeons. She served as the 2010 President of the Eastern North American Region of the International Biometrics Society; was inaugural co-chair of the PCORI Methodology Committee; co-chairs a Committee on National Statistics/National Academy of Sciences panel reviewing the Safety Measurement System of the Compliance, Safety, Accountability program run by the Federal Motor Carrier Safety Administration; and served on several National Academy of Sciences Committees, including the Committee of Applied and Theoretical Statistics (CATS) focusing on the intersections of statistics and computer science for big data. Dr. Normand received ASA's Health Policy Statistics Section Long Term Excellence Award, the Outstanding Lifetime Achievement Award from the American Heart Association, the L. Adrienne Cupples Award for Excellence in Teaching, Research, and Service in Biostatistics from Boston University, and the Mosteller Statistician of the Year from the Boston Chapter of the ASA.

WELCOME TO NASHVILLE!

Named for Francis Nash, a general of the Continental Army during the American Revolutionary War, the city was founded in 1779. The city grew quickly due to its strategic location as a port on the Cumberland River and, in the 19th century, a railroad center. Nashville seceded with Tennessee during the American Civil War; in 1862 it was the first state capital in the Confederacy to fall to Union troops. After the war, the city reclaimed its position and developed a manufacturing base. Today, Nashville is known as Music City.

If cities had soundtracks, Nashville's would be like no others. It would be a mix of music's past, present and future with cuts of country, bluegrass, rock, pop, Americana, gospel, classical, jazz and blues, all blending and overlapping in perfect harmony. Live music can be heard when walking through almost any neighborhood, with open mic nights featuring talent you'd expect to pay good money to hear. The city has experienced significant growth in the last few years, as the healthcare industry and the growing appeal of tourism have led to the development of new neighborhoods and the revitalization of old ones, as well as a booming food and beer scene. You won't have any trouble filling your trip with the sights, sounds, and tastes of Nashville.

LANDMARKS AND TOURS

RYMAN AUDITORIUM

When you walk through the doors of historic Ryman Auditorium, one thing becomes clear right away: this isn't just another nightly music venue, and it's so much more than a daytime tourist stop. This place is hallowed ground. This is the exact spot where bluegrass was born—where Johnny Cash met June Carter, where souls were saved and a slice of history was nearly lost. It was right here that country music found an audience beyond its own back porch, and countless careers took off as deals were signed on napkins and paper scraps backstage. Open daily for tours and shows, right in the heart of Music City.

TENNESSEE STATE CAPITOL AND LEGISLATIVE PLAZA

Designed by architect William Strickland and built in the Greek Revival architecture style that models a Greek temple, the Tennessee State Capitol is one of 12 capitol buildings in the U.S. that does not have a dome. The Capitol sits on the hilltop site once occupied by the Holy Rosary Cathedral, which was the first Roman Catholic cathedral in Nashville. On the grounds of the Capitol are two statues of U.S. presidents: Andrew Jackson and Andrew Johnson. President James K. Polk is buried in a tomb on the Capitol grounds, along with his wife, Sarah Childress Polk. Other monuments include a Tennessee Holocaust Memorial, the Sam Davis Memorial, and Sen. Edward Ward Carmack Memorial. Across the street, check out Legislative Plaza, where you'll find a statue dedicated to the Women of the Confederacy, a monument to Tennesseans, who served in the Korean War, and to the south Vietnam Veterans Park.

TENNESSEE STATE MUSEUM

Situated on the bottom floors of the James K. Polk building downtown is the Tennessee State Museum. It depicts the history of the state of Tennessee, starting from pre-colonization and going into the 20th century. With more than 120,000 square feet (11,148 square meters) of space among three floors, the museum includes both permanent and changing exhibits that display paintings, weapons, furniture, uniforms, and battle flags from the Civil War. Larger exhibits include a painting gallery, a reproduction of a historic print shop, and a grist mill. There's also a museum store where visitors can purchase handmade crafts and Tennessee memorabilia.

THE PARTHENON

Standing as the centerpiece in Nashville's Centennial Park, the Parthenon is a full-scale replica of the Parthenon in Athens, Greece. Come inside to see the 42-foot gilded sculpture of Athena, the permanent display of American paintings from the Cowan Collection, the history of the Nashville Parthenon dating back to the 1897 Tennessee Centennial Exposition, and a variety of temporary shows and exhibitions! The entrance is located on the ground level of the East side of the building.

THE GRAND OLE OPRY

Take a trip to the historic Grand Ole Opry, located next to the Opryland Resort and Convention Center, about 20 minutes from the JW Marriot. You can take a backstage tour of the Opry's 18 themed dressing rooms, learn behind-the-scenes secrets, and just maybe step foot in "The Circle", the center of the Grand Ole Opry and the most sacred space in country music.

HISTORY & ART MUSEUMS

ANDREW JACKSON'S HERMITAGE

The Hermitage, Home of President Andrew Jackson, is one of the largest and most visited presidential homes in the United States, and recently named the #1 historic house in Tennessee. Today, The Hermitage is a 1,120-acre National Historic Landmark with more than 30 historic buildings, that welcomes some 200,000 annual visitors, including 30,000 school children, from all 50 states and many foreign countries. Visit Andrew Jackson's Hermitage to witness "The Duel: The Art of the Southern Gentleman." This 30-minute visitor experience will answer questions about dueling followed by an ACTUAL demonstration by onsite historic re-enactors. "The Duel" takes place every Thursday through Sunday throughout the day, free with paid admission. The Hermitage is about 15 miles from the JW Marriot.

BELLE MEADE PLANTATION

Belle Meade Plantation is a non-profit historic site located in Nashville. Established in 1807, Belle Meade was revered as the greatest thoroughbred stud farm in the United States. It was home to Iroquois, the first American bred horse to win the Epsom Derby and the great foundation sire, Bonnie Scotland, whose descendants include Secretariat, Seattle Slew, Native Dancer, Big Brown and California Chrome. Belle Meade was owned and operated by the Harding-Jackson family for nearly a century from 1807 until 1906. Today, their home is restored to its turn-of-the-century appearance, along with several original outbuildings. The Plantation is about 20 minutes from the JW Marriot by car.

BELMONT MANSION

Belmont Mansion is the largest house museum in Tennessee and one of a few whose history revolves around the life of a woman: Adelicia Acklen. We host visitors seven days a week for tours and are open as a rental venue for weddings and events. Tours may be purchased online or at the door. Belmont Mansion, and Belmont campus, are about 10 minutes from the JW Marriot by car.

THE FRIST

If you're in the mood to view some art in a gorgeous setting, look no further than the Frist Art Museum. Situated in a classic art deco building, the museum houses a rotating schedule of exhibitions from local, regional, national, and international sources. The Frist is a family-friendly environment, with the Martin ArtQuest Gallery providing more than 30 interactive art-making stations and free admission for youth 18 and under.

PARKS

CENTENNIAL PARK (MIDTOWN)

Smack dab in the middle of the hustle and bustle of offices, restaurants, and streets, Centennial is the perfect place to take a short walk on a lunch break. It is also home to the Parthenon replica, giving people educational benefits, along with their dose of vitamin D. Its convenient location also makes it a prime spot for events and activities. It is home to several festivals, fairs, and music series. On any given weekend, the park is full of music, food, and fun. With all of this activity, it is not the best for wildlife or nature viewing, and finding a secluded, quiet spot free of frisbees or college kids may be a challenge.

BICENTENNIAL CAPITOL MALL STATE PARK (DOWNTOWN)

This 19-acre park in the heart of Nashville serves as a monument to the bicentennial celebration of the State of Tennessee. This park offers plenty of opportunities to learn about the long history of Tennessee, while having a great experience in a beautiful green space. More information on this park and many more can be found at the Tennessee State Parks website. While in the area, consider checking out the Nashville Farmer's Market, located next door to the park. It's a great spot to grab lunch or a coffee to go!

CUMBERLAND PARK (EAST NASHVILLE)

Located on the East side of the river, Cumberland Park was completely renovated to become a go-to attraction for Nashvillians. There are a lot of features in this park, including an outdoor amphitheater for events, a rock climbing wall, water features, green space for kids to play and much more. The park is a short 1.5 mile walk from the JW Marriot.

SEVIER PARK (12 SOUTH)

A small park in a growing neighborhood, Sevier Park has the best of both worlds. It's fun and lively, but hardly ever crowded! With Las Palatas and Burger Up nearby, hunger won't be an issue either. There is a small playground area and a creek down below in the shadow of the pre-Civil War Sunnyside Mansion. The mansion invites the new growth of the 12 South neighborhood in its historic front yard. This park is about 4 miles from the JW Marriot.

LOCAL CUISINE AND RESTAURANTS
(BY NEIGHBORHOOD)

Nashville's dining scene is exploding thanks to a combination of chef-driven restaurants and classic dining spots offering up Nashville Hot Chicken, barbecue, and Meat & Three fare. Below are some top local picks, organized by neighborhood. For more information, and to find your favorite flavor, check out **https://www.visitmusiccity.com/things-to-do/food-drink**. For a more in-depth look at the different neighborhoods in Nashville, visit **www.nashvilleguru.com/neighborhoods.**

DOWNTOWN

Home to honky tonks, live music, and more boot shops than you could ever need, Downtown Nashville is the heart of "Nashvegas" energy. With more than 60 bars and restaurants, it's hard to go wrong here if you're looking for a fun evening. For a great view, consider Acme Feed & Seed, with multiple floors of dining options and a great rooftop bar overlooking the river. If you're looking for name recognition, consider Jason Aldean's Kitchen + Rooftop Bar, or Tootsies Orchid Lounge for a slice of honky tonk history.

SOBRO ("SOUTH OF BROADWAY")

Just steps from Broadway lies the SoBro neighborhood, a quieter area with a lot of new developments. Check out Bajo Sexto Taco for a quick and tasty lunch, Martin's Bar-B-Que Joint for a classic plate of Nashville BBQ, or Tennessee Brew works for live music and delicious burgers. If you're looking for some activities to go with your meal, check out Pinewood Social, featuring vintage bowling lanes and an extensive craft cocktail selection to accompany your meal.

THE GULCH

Located just a short walk from the JW Marriot, The Gulch is a small, upscale neighborhood full of great food and shops. In particular, they're known for Arnold's Country Kitchen (the quintessential Meat & Three experience, open Monday – Friday). For an upscale dinner, consider Chauhan Ale & Masala House (Indian cuisine with a Southern flair), Sambuca (American food, live music), or Sunda (Southeast Asian cuisine, with ample seating for larger groups).

GERMANTOWN

A quiet neighborhood just north of Downtown, Germantown features brick sidewalks, ample greenery, and easy street parking in a family-friendly atmosphere. There are more than 15 dining options to be found, including casual options like Red Bicycle Coffee and Crepes and Von Elrod's Beer Hall and Kitchen (with delicious soft pretzels and an extensive German and Belgian beer list), as well as more upscale dinner options like 5th & Taylor (American cuisine, reservation required) and Geist (set in an old blacksmith shop, with upscale Southern cuisine and a champagne garden).

MIDTOWN

For an area that's a little more relaxed and less touristy, consider Midtown, located near Vanderbilt University and popular among locals and students. Here you'll find Hattie B's, one of the most popular Nashville Hot Chicken joints. If the line is long, consider other Hattie B's in the area. Consider using their order-ahead option to avoid the line! In this neighborhood you'll also find several delicious restaurants with a Southern flair, like The Row or Tavern. For a nicer meal, check out Union Common for duck fat French fries and steak, or Patterson House for a speakeasy cocktail experience.

HILLSBORO VILLAGE

The neighborhood of Hillsboro Village, conveniently located just south of Vanderbilt University campus, is a great spot to do some shopping and grab a coffee or a bite to eat. The Pancake Pantry is a favorite among visitors, but be sure to get there early to beat the line! Fido is a great spot for coffee, and serves a delicious all-day breakfast menu that's popular among Vanderbilt and Belmont students alike.

12 SOUTH

If you're in the area to check out the Belmont Mansion, walk on down to the 12 South neighborhood! The street features some great shopping (including Reese Witherspoon's boutique, Draper James) as well as excellent options for eating. Consider Frothy Monkey for coffee and breakfast favorites, Five Daughters Bakery for delicious cronuts (donuts made from layered croissant dough), or Christie Cookie Co. for a fresh-baked cookie and milk. For dinner, look no further than Edley's Bar-B-Que, with incredibly tasty pulled pork, fried pickles, and jalapeño cornbread. In the mood for lighter fare? Check out Epice Lebanese Bistro for stuffed grape leaves, lentil soup, and traditional grilled chicken and lamb skewers.

PROGRAM SUMMARY

Sunday, March 22

7:30 a.m.—6:30 p.m.	**Conference Registration**
8:00 a.m.—5:00 p.m.	**Short Courses** **SC1:** Implementing Bayesian Adaptive Designs: From Theory to Practice **SC2:** Practical solutions for working with electronic health records data **SC3:** Design and Analysis of Sequential, Multiple Assignment, Randomized Trials for small and large samples
8:00 a.m.—12:00 p.m.	**Short Courses** **SC4:** Programming with hierarchical statistical models: Using the BUGS-compatible NIMBLE system for MCMC and more **SC6:** Statistical Network Analysis with Applications to Biology
10:30 a.m.—6:30 p.m.	**Fostering Diversity in Biostatistics Workshop**
1:00 p.m.—5:00 p.m.	**Short Courses** **SC5:** Multivariate meta-analysis methods **SC7:** Trial Design and Analysis Using Multisource Exchangeability Models
3:00 p.m.—6:00 p.m.	**Exhibits Open**
4:00 p.m.—6:30 p.m.	**Career Placement Services**
4:30 p.m.—7:00 p.m.	**ENAR Executive Committee Meeting**
7:30 p.m.—8:00 p.m.	**New Member Reception**
8:00 p.m.—11:00 p.m.	**Opening Mixer and Poster Session** 1. Posters: Imaging Data Analysis 2. Posters: Survival Analysis/Competing Risks 3. Posters: Machine Learning and High-Dimensional Data 4. Posters: Personalized Medicine and Biomarkers 5. Posters: Cancer Applications 6. Posters: Clinical Trials 7. Posters: Diagnostics/Prediction/Agreement 8. Posters: Adaptive Design/Experimental Design 9. Posters: Bayesian Methods 10. Posters: Causal Inference and Clinical Trials 11. Posters: Genomics/Proteomics 12. Posters: Functional Data/High Dimensional 13. Posters: Bayesian, Clustered Data, Hypothesis Testing 14. Posters: High-Dimensional Data, Missing Data and More 15. Posters: Consulting, Education, Policy, Epidemiology 16. Posters: Genetics, Computation 17. Posters: Meta-Analysis, Missing Data and More

Monday, March 23

7:30 a.m.—5:00 p.m.	**Conference Registration**
7:30 a.m.—5:00 p.m.	**Speaker Ready Room**
8:30 a.m.—5:30 p.m.	**Exhibits Open**
8:30 a.m.—10:15 a.m.	**Tutorial** **T1: Statistical methods for geometric functional data**

Monday, March 23 (continued)

8:30 a.m.—10:15 a.m.	**Scientific Program** 18. Modern Functional Data Analysis 19. Distributed and Privacy-Preserving Methods for Electronic Health Records Data 20. Innovative Statistical Methods in Environmental Mixture Analysis 21. Mentoring Throughout a Lifetime: Considerations for Mentors and Mentees at all Career Stages 22. Innovative Statistical Approaches for High-Dimensional Omic and Microbiomic Data 23. Bayesian Nonparametrics for Causal Inference and Missing Data 24. Contributed Papers: Variable Selection: How to Choose? 25. Contributed Papers: Functional Data Analysis 26. Contributed Papers: Penalized and Other Regression Models with Applications 27. Contributed Papers: Methods for Neuroimaging Data: Get the Picture? 28. Contributed Papers: Causal Effect Estimation
9:30 a.m. —4:30 p.m.	**Career Placement Services**
10:15 a.m.—10:30 a.m.	**Refreshment Break with Our Exhibitors**
10:30 a.m.–-12:15 p.m.	**Tutorial** **T2: Disease Risk Modeling and Visualization using R**
10:30 a.m.—12:15 p.m.	**Scientific Program** 29. New Perspectives on Data Integration in Genome-Wide Association Studies 30. Advances in Causal Inference and Joint Modeling with Survival and Complex Longitudinal Data 31. Opportunities and Challenges in the Analysis and Integration of Large-Scale Biobank Data 32. Compositional Nature of Microbiome Data: Challenges and New Methods 33. Statistical Modeling in Alzheimer's Disease 34. Recent Advances in Bayesian Methods for Spatial-Temporal Processes 35. Speed Posters: EHR Data, Epidemiology, Personalized Medicine, Clinical Trials 36. Contributed Papers: Adaptive Designs for Clinical Trials 37. Contributed Papers: Bayesian Semiparametric and Nonparametric Methods 38. Contributed Papers: Statistical Methods in Cancer Research 39. Contributed Papers: Network Analysis: Connecting the Dots 40. Contributed Papers: Policies and Politics: Statistical Analyses of Health Outcomes in the Real World 41. Contributed Papers: Statistical Considerations for Optimal Treatment
12:15 p.m.—1:30 p.m.	**Roundtable Luncheons**
12:30 p.m.—4:30 p.m.	**Regional Advisory Board (RAB) Luncheon Meeting (by Invitation Only)**
1:45 p.m.—3:30 p.m.	**Tutorial** **T3: Integration of Genetics and Imaging Data in Scientific Studies**
1:45 p.m.—3:30 p.m.	**Scientific Program** 42. Causal Inference with Genetic Data 43. Recent Advances in Statistical Methods for Single-Cell Omics Analysis 44. Recent Advances in Microbiome Data Analysis 45. Novel Methods to Evaluate Surrogate Endpoints 46. Recent Advances in the Uncertainty Estimation and Properties of Bayesian Additive Regression Trees 47. Current Developments in Analyzing EHR and Biobank Data 48. Speed Posters: Causal Inference/Longitudinal Data/High-Dimensional Data/Massive Data 49. Contributed Papers: Statistical Methods for Omics Data Analysis 50. Contributed Papers: Observational and Historical Data Analysis: The Rest is History 51. Contributed Papers: Immunotherapy Clincial Trial Design and Analysis 52. Contributed Papers: Machine Learning and Statistical Relational Learning 53. Contributed Papers: Time Series and and Recurrent Event Data 54. Contributed Papers: Massive Data: A Giant Problem?
3:30 p.m.—3:45 p.m.	**Refreshment Break with Our Exhibitors**
3:45 p.m.—5:30 p.m.	**Tutorial** **T4: Causal Inference Using the R TWANG Package for Mediation and Continuous Exposures**

Monday, March 23 (continued)

3:45 p.m.—5:30 p.m.	**Scientific Program** 55. Human Microbiome Studies: Novel Methods and New Studies 56. Bayesian Approaches for Complex Innovative Clinical Trial Design 57. Achieving Real-World Evidence from Real-World Data: Recent Developments and Challenges 58. Novel Spatial Modeling Approaches for Air Pollution Exposure Assessment 59. Innovations in Two Phase Sampling Designs with Applications to EHR Data 60. Recent Approaches to Multivariate Data Analysis in the Health Sciences 61. Speed Posters: Imaging Data/Survival Analysis/Spatio-Temporal 62. Contributed Papers: Imaging and Streaming Data Analysis 63. Contributed Papers: Causal Inference and Propensity Score Methods 64. Contributed Papers: Longitudinal Data and Joint Models of Longitudinal and Survival Data 65. Contributed Papers: Personalized Medicine and Biomarkers 66. Contributed Papers: Statistical Genetics: Single-Cell Sequencing Data 67. Contributed Papers: Semiparametric and Nonparametric Methods and Applications
5:30 p.m.—6:30 p.m.	**CENS Networking Mixer**
6:30 p.m.—7:30 p.m.	**President's Reception (by Invitation Only)**

Tuesday, March 24

7:30 a.m.—5:00 p.m.	**Conference Registration**
7:30 a.m.—5:00 p.m.	**Speaker Ready Room**
8:30 a.m.—5:30 p.m.	**Exhibits Open**
8:30 a.m.—10:15 a.m.	**Scientific Program** 68. Challenges and Opportunities in Methods for Precision Medicine 69. Recent Developments in Risk Estimation and Biomarker Modeling with a Focus in Alzheimer's Disease 70. Clinical Trial Designs in a New Era of Immunotherapy: Challenges and Opportunities 71. The Three M's: Meetings, Memberships, and Money! 72. Recent Advances in Joint Modeling of Longitudinal and Survival Data 73. Recent Advances in Network Meta-Analysis with Flexible Bayesian Approaches 74. Contributed Papers: Electronic Health Records Data Analysis 75. Contributed Papers: Rebel Without a Cause: Sessions on Causal Inference 76. Contributed Papers: Hypothesis Testing: Knowledge is Power 77. Contributed Papers: Missing (Data) in Action 78. Contributed Papers: Back to the Future: Prediction and Prognostic Modeling 79. Contributed Papers: M&M: Measurement Error and Modeling
9:30 a.m.—3:30 p.m.	**Career Placement Services**
10:15 a.m.—10:30 a.m.	**Refreshment Break with Our Exhibitors**
10:30 a.m.—12:15 p.m.	80. Presidential Invited Address
12:30 p.m.—4:30 p.m.	**Regional Committee Luncheon Meeting (by Invitation Only)**
1:45 p.m.—3:30 p.m.	**Tutorial** **T5: Fundamentals of difference-in-differences studies**
1:45 p.m.—3:30 p.m.	**Scientific Program** 81. Statistical Analysis of Biological Shapes 82. Improving the Development and Validation of Screening Tests for Rare Diseases 83. Causal Inference and Harmful Exposures 84. Statistical Methods for Emerging Data in Environmental Health Research 85. Bayesian Analysis in Functional Brain Imaging 86. Human Data Interaction: Gaining an Understanding of the Data Science Pipeline 87. Contributed Papers: Spatial and Spatial-Temporal Data Analysis 88. Contributed Papers: Early Phase Clinical Trials and Biomarkers 89. Contributed Papers: Electronic Health Records Data Analysis and Meta-Analysis 90. Contributed Papers: Small Things that Make a Big Difference: Microbiome Analysis 91. Contributed Papers: Statistical Genetics: Sequencing Data Analysis 92. Contributed Papers: Robust Modeling and Inference

Tuesday, March 24 (continued)

3:30 p.m.—3:45 p.m.	**Refreshment Break with Our Exhibitors**
3:45 p.m.—5:30 p.m.	**Tutorial** **T6: R package development**
3:45 p.m.—5:30 p.m.	**Scientific Program** 93. High Dimensional Methods for Mechanistic Integration of Multi-Type Omics 94. New Weighting Methods for Causal Inference 95. Using Machine Learning to Analyze Randomized Trials: Valid Estimates and Confidence Intervals Without Model Assumptions 96. Recent Developments in Semiparametric Transformation Models 97. Innovations in Statistical Neuroscience 98. Artificial Intelligence for Prediction of Health Outcomes 99. Contributed Papers: Latent Variables and Processes 100. Contributed Papers: Time-to-Event Data Analysis: Survival of the Fittest 101. Contributed Papers: Risky Business: Diagnostics, ROC, and Prediction 102. Contributed Papers: Interval-Censored and Multivariate Survival Data 103. Contributed Papers: Graphical Models and Applications 104. Contributed Papers: Support Vector Machines, Neural Networks and Deep Learning
5:30 p.m.—7 p.m.	**ENAR Business Meeting and Sponsor/Exhibitor Mixer** – Open to all ENAR Members

Wednesday, March 25

7:30 a.m. —12:00 p.m.	**Speaker Ready Room**
7:30 a.m.—9:00 a.m.	**Planning Committee (by Invitation Only)**
8:00 a.m.—12:30 p.m.	**Conference Registration**
8:00 a.m.—12:00 p.m.	**Exhibits Open**
8:30 a.m.—10:15 a.m.	**Scientific Program** 105. Advances in Statistical Modeling for Multi-omics Data Integration 106. Causal Inference and Network Dependence: From Peer Effects to the Replication Crisis in Epidemiology 107. Flexible Spatio-Temporal Models for Environmental and Ecological Processes 108. Recent Advances in Neuroimaging Analytics 109. Novel Tensor Methods for Complex Biomedical Data 110. Integrative Analysis of Clinical Trials and Real-World Evidence Studies 111. Contributed Papers: Clustered Data Methods 112. Contributed Papers: Subgroup Analysis 113. Contributed Papers: Functional Data Analysis: Below the Surface 114. Contributed Papers: HIV, Infectious Disease and More 115. Contributed Papers: Clinical Trial Design and Analysis 116. Contributed Papers: Multivariate and High-Dimensional Data Analysis
10:15 a.m.—10:30 a.m.	**Refreshment Break with Our Exhibitors**
10:30 a.m.—12:15 p.m.	**Scientific Program** 117. Asymmetrical Statistical Learning for Binary Classification 118. Recent Advances and Opportunities in Large Scale & Multi-Omic Single-Cell Data Analysis 119. Novel Statistical Methods for Complex Interval-Censored Survival Data 120. Modern Graphical Modeling of Complex Biomedical Systems 121. Highly Efficient Designs and Valid Analyses for Resource Constrained Studies 122. Statistical Analysis of Tracking Data from Personal Wearable Devices 123. Contributed Papers: Meta-Analysis Methods 124. Contributed Papers: Longitudinal Data Analysis 125. Contributed Papers: High Dimensional Data Analysis: The BIG Picture 126. Contributed Papers: Clinical “Trials and Tribulations” 127. Contributed Papers: Count Data: The Thought that Counts

SCIENTIFIC PROGRAM

Sunday, March 22

POSTER PRESENTATIONS

1. POSTERS: IMAGING DATA ANALYSIS

Sponsor: ENAR

1a. Time Varying Estimation of Tensor-on-Tensor Regression with Application in fMRI Data

Pratim Guha Niyogi* and Tapabrata Maiti, Michigan State University

1b. Estimation of Fiber Orientation Distribution through Blockwise Adaptive Thresholding

Seungyong Hwang*, Thomas Lee, Debashis Paul and Jie Peng, University of California, Davis

1c. Estimating Dynamic Connectivity Correlates of PTSD Resilience Using MultiModal Imaging

Jin Ming*, Suprateek Kundu and Jennifer Stevens, Emory University

1d. Towards an Automatic Detection Method of Chronic Active Lesions

Carolyn Lou*, Jordan D. Dworkin and Alessandra Valcarcel, University of Pennsylvania; Martina Absinta and Pascal Sati, National Institute of Neurological Disorders and Stroke, National Institutes of Health; Kelly Clark, University of Pennsylvania; Daniel Reich, National Institute of Neurological Disorders and Stroke, National Institutes of Health

1e. A Bayesian Mixture Model for Lesion Detection and Clustering in MS

Jordan D. Dworkin*, Melissa L. Martin, Arman Oganisian and Russell T. Shinohara, University of Pennsylvania

1f. Seeing Very Small Things: Applications of Mixture Modeling and Extreme Value Distributions in Microscopic Image Analysis

Miranda L. Lynch* and Sarah E.J. Bowman, Hauptman-Woodward Medical Research Institute

2. POSTERS: SURVIVAL ANALYSIS/COMPETING RISKS

Sponsor: ENAR

2a. Functional Additive Cox Model

Erjia Cui*, Andrew Leroux and Ciprian Crainiceanu, Johns Hopkins University

2b. Gene-Based Association Analysis of Survival Traits via Functional Regression based Mixed Effect Cox Models for Related Samples

Ruzong Fan*, Georgetown University Medical Center; Chi-yang Chiu, University of Tennessee Health Science Center; Bingsong Zhang, Shuqi Wang and Jingyi Shao, Georgetown University Medical Center; M'Hamed Lajmi Lakhal-Chaieb, Universite Laval; Richard J. Cook, University of Waterloo; Alexander F. Wilson and Joan E. Bailey-Wilson, Computational and Statistical Genomic Branch of the National Human Genome Research Institute, National Institutes of Health; Momiao Xiong, University of Texas Health Science Center at Houston

2c. Regression Model for the Lifetime Risk using Pseudo-Values

Sarah C. Conner* and Ludovic Trinquart, Boston University School of Public Health

2d. Proportional Subdistribution Hazards Model with Covariate-Adjusted Censoring Weight for Clustered Competing Risks Data

Manoj Khanal*, Soyoung Kim and Kwang Woo Ahn, Medical College of Wisconsin

2e. A Unified Power Series Class of Cure Rate Survival Models for Spatially Clustered Data

Sandra Hurtado Rua*, Cleveland State University; Dipak Dey, University of Connecticut

2f. Optimizing Incremental Cost-Effective Ratios for Censored Survival Time and Cost

Xinyuan Dong*, University of Washington

2g. An EM Algorithm in Fitting the Generalized Odds-Rate Model to Right Censored Data

Ennan Gu*, University of South Carolina

3. POSTERS: MACHINE LEARNING AND HIGH-DIMENSIONAL DATA

Sponsor: ENAR

3a. Distributed Quadratic Inference Functions for Integrating Studies with High-Dimensional Repeated Measures

Emily C. Hector* and Peter X.K. Song, University of Michigan

3b. Statistical Inference for the Word2vec Natural Language Processing Algorithm Applied to Electronic Health Records

Brian L. Egleston*, Stan Taylor, Michael Lutz and Richard J. Bleicher, Fox Chase Cancer Center; Slobodan Vucetic, Temple University

3c. Neural Network Survival Model for Cardiovascular Disease Prediction

Yu Deng*, Northwestern University; Lei Liu, Washington University, St. Louis; HongMei Jiang, Kho Abel, Yishu Wei, Norrina Allen, John Wilkins, Kiang Liu, Donald Lloyd-Jones and Lihui Zhao, Northwestern University

SCIENTIFIC PROGRAM
(CONTINUED)

3d. Applying Statistical Learning Algorithms on the Prediction of Response to Immune Checkpoint Blockade Therapy

Tiantian Zeng* and Chi Wang, University of Kentucky

3e. Integrative Biclustering for Characterization of Biomarker and Phenotype Associations

Weijie Zhang*, University of Minnesota

3f. Testing Presence-Absence Association in the Microbiome Using LDM and PERMANOVA

Andrea N. Lane*, Emory University; Glen Satten, Centers for Disease Control and Prevention; Yijuan Hu, Emory University

3g. Feature Selection for Support Vector Regression Using a Genetic Algorithm

Shannon B. McKearnan*, David M. Vock and Julian Wolfson, University of Minnesota

3h. Statistical Inferences for F1-scores in Multi-Class Classification Problems

Kouji Yamamoto*, Yokohama City University; Kanae Takahashi, Osaka City University; Aya Kuchiba, National Cancer Center, National Institutes of Health; Tatsuki Koyama, Vanderbilt University Medical Center

4. POSTERS: PERSONALIZED MEDICINE AND BIOMARKERS

Sponsor: ENAR

4a. Individualized Treatment Effect Estimation using Auto-Encoder and Conditional Generative Adversarial Networks

Yuanyuan Liu* and Momiao Xiong, University of Texas Health Science Center at Houston

4b. Weighted Sparse Additive Learning for ITR Estimation under Covariate Space Sparsity

Jinchun Zhang*, New York University

4c. One-Step Value Difference Test for the Existence of a Subgroup with a Beneficial Treatment Effect Using Random Forests

Dana Johnson*, Wenbin Lu and Marie Davidian, North Carolina State University

4d. Selecting Optimal Cut-Points for Early-Stage Detection in K-class Diseases Diagnosis Based on Concordance and Discordance

Jing Kersey*, Hani Samawi, Jingjing Yin, Haresh Rochani and Xinyan Zhang, Georgia Southern University

4e. Designing and Analyzing Clinical Trials for Personalized Medicine via Bayesian Models

Chuanwu Zhang*, Matthew S. Mayo, Jo A. Wick and Byron J. Gajewski, University of Kansas Medical Center

4f. Some Improved Tests for the Assessment of Bioequivalence and Biosimilarity

Rabab Elnaiem* and Thomas Mathew, University of Maryland, Baltimore County

4g. Fusing Continuous and Time-Integrated Data for Estimating Personal Air Pollution Exposures

Jenna R. Krall* and Anna Z. Pollack, George Mason University

4h. Value of Biostatistical Support in a Hospital Quality Improvement Department

Henry John Domenico*, Daniel W. Byrne and Li Wang, Vanderbilt University Medical Center

4i. Prediction of Intervention Effects in Healthcare Systems

Emily A. Scott*, Johns Hopkins Bloomberg School of Public Health; Zhenke Wu, University of Michigan; Elizabeth Colantuoni, Johns Hopkins Bloomberg School of Public Health; Sarah Kachur, Johns Hopkins HealthCare; Scott L. Zeger, Johns Hopkins Bloomberg School of Public Health

5. POSTERS: CANCER APPLICATIONS

Sponsor: ENAR

5a. Comparison of Several Bayesian Methods for Basket Trials when a Control of Subgroup-Wise Error Rate is Required

Gakuto Ogawa* and Shogo Nomura, National Cancer Center, Japan

5b. Gene Profile Modeling and Integration for EWOC Phase I Clinical Trial Design while Fully Utilizing all Toxicity Information

Feng Tian* and Zhengjia (Nelson) Chen, Rollins School of Public Health, Emory University

5c. A Pan-Cancer and Polygenic Bayesian Hierarchical Model for the Effect of Somatic Mutations on Survival

Sarah Samorodnitsky*, University of Minnesota; Katherine A. Hoadley, University of North Carolina, Chapel Hill; Eric F. Lock, University of Minnesota

5d. A Novel GEnomic NEtwork COrrelation Merging System (GENECOMS) to Investigate the Relation between Differentially Expressed Methylation Regions and Gene Modules in Bladder Cancer

Shachi Patel* and Jeffrey Thompson, University of Kansas Medical Center

SCIENTIFIC PROGRAM
(CONTINUED)

5e. Comparing the Performance of Phase I/II Oncology Trial Designs in Low-Toxicity Rate Situations

Ryo Takagi* and Isao Yokota, Hokkaido University Hospital

5f. Advantage of Using a Finite-Sample Correction when Designing Clinical Trials in Rare Diseases

Audrey Mauguen*, Memorial Sloan Kettering Cancer Center

5g. Implementation of Clusterability Testing Prior to Clustering

Naomi Brownstein*, Moffitt Cancer Center

5h. A Probabilistic Model for Leveraging Intratumor Heterogeneity Information to Enhance Estimation of the Temporal Order of Pathway Mutations during Tumorigenesis

Menghan Wang*, Chunming Liu, Arnold Stromberg and Chi Wang, University of Kentucky

5i. Functional Clustering via Weighted Dirichlet Process Modeling with Breast Cancer Genomics Data

Wenyu Gao* and Inyoung Kim, Virginia Tech

6. POSTERS: CLINICAL TRIALS

Sponsor: ENAR

6a. Sample Size Determination Method that Accounts for Selection Probability of the Maximum Tolerated Dose in Phase I Oncology Trials

Yuta Kawatsu*, Jun Tsuchida, Shuji Ando and Takashi Sozu, Tokyo University of Science; Akihiro Hirakawa, The University of Tokyo

6b. The Scale Transformed Power Prior with Applications to Studies with Different Endpoints

Brady Nifong*, Matthew A. Psioda and Joseph G. Ibrahim, University of North Carolina, Chapel Hill

6c. Design and Analysis for Three-Arm Clinical Trials with Intra-Individual Right-Left Data

Ryunosuke Machida*, National Cancer Center, Japan; Kentaro Sakamaki, Yokohama City University, Japan; Aya Kuchiba, National Cancer Center, Japan

6d. An Estimation of Efficacy of Potential Drug in Multiple Diseases with Discriminating Heterogeneity in Treatment Effects in Basket Trials

Shun Hirai*, Jun Tsuchida, Shuji Ando and Takashi Sozu, Tokyo University of Science; Akihiro Hirakawa, The University of Tokyo

6e. Longitudinal Study of Opioid Treatment on Hamilton Depression Rating Scale Using a Negative Binomial Mixed Model

Kesheng Wang*, West Virginia University; Wei Fang, West Virginia Clinical and Translational Science Institute; Toni DiChiacchio, West Virginia University; Chun Xu, University of Texas Rio Grande Valley; Ubolrat Piamjariyakul, West Virginia University

6f. Group Sequential Analysis for Sequential Multiple Assignment Randomized Trials

Liwen Wu*, Junyao Wang and Abdus S. Wahed, University of Pittsburgh

6g. Incorporating Truncation Information from Phase I Clinical Studies into Phase II Designs

Li-Ching Huang* and Fei Ye, Vanderbilt University Medical Center; Yi-Hsuan Tu, Independent Scholar; Chia-Min Chen, Nanhua University, Taiwan; Yu Shyr, Vanderbilt University Medical Center

6h. Replicability of Treatment Effects in Meta-Analyses

Kirsten R. Voorhies*, Brown University; Iman Jaljuli and Ruth Heller, Tel-Aviv University; Orestis A. Panagiotou, Brown University

7. POSTERS: DIAGNOSTICS/PREDICTION/ AGREEMENT

Sponsor: ENAR

7a. A Resampling Perspective on Evaluation of Diagnosis Accuracy: An Appendicitis Example

Calvin S. Elder*, St. Jude Children's Research Hospital; Yousef El-Gohary, Center of Colorectal and Pelvic Reconstruction; Hui Zhang, Northwestern University; Li Tang, St. Jude Children's Research Hospital

7b. Improving the Performance of Polygenic Risk Score with Rare Genetic Variants

Hongyan Xu* and Varghese George, Augusta University

7c. A Domain Level Index to Enhance the Prediction Accuracy of Pathogenic Variants

Hua-Chang Chen* and Qi Liu, Vanderbilt University Medical Center

7d. The Cornelius Project - Randomizing Real-Time Predictive Models Embedded in the Electronic Health Record to Assess Impact on Health Outcomes

Daniel W. Byrne*, Henry J. Domenico and Li Wang, Vanderbilt University

7e. Privacy-Preserving Outcome Prediction

Lamin Juwara*, McGill University

SCIENTIFIC PROGRAM
(CONTINUED)

7f. Interpretable Clustering of Hierarchical Dependent Binary Data: A Doubly-Multi-Resolution Approach

Zhenke Wu*, Yuqi Gu, Mengbing Li and Gongjun Xu, University of Michigan

7g. Estimation and Construction of Confidence Intervals for the Cutoff-Points of Continuous Biomarkers Under the Euclidean Distance in Trichotomous Settings

Brian Mosier* and Leonidas Bantis, University of Kansas Medical Center

7h. Confidence Interval of the Mean and Upper Tolerance Limit for Zero-Inflated Gamma Data

Yixuan Zou* and Derek S. Young, University of Kentucky

7i. Predictive Performance of Physical Activity Measures for 1-year up to 5-year All-Cause Mortality in NHANES 2003-2006

Lucia Tabacu*, Old Dominion University; Mark Ledbetter, Lynchburg University; Andrew Leroux and Ciprian Crainiceanu, Johns Hopkins University

8. POSTERS: ADAPTIVE DESIGN/EXPERIMENTAL DESIGN

Sponsor: ENAR

8a. An Empirical Bayesian Basket Trial Design Accounting for Uncertainties of Homogeneity and Heterogeneity of Treatment Effect among Subpopulations

Junichi Asano*, Pharmaceuticals and Medical Devices Agency; Akihiro Hirakawa, The University of Tokyo

8b. Lessons Learned in Developing an Interdisciplinary Collaboration Between Biostatistics and Forensic Nursing

Yesser Sebeh*, Georgia State University; Katherine Scafide, George Mason University; Matthew J. Hayat, Georgia State University

8c. Response-Adaptive Randomization in a Two-Stage Sequential Multiple Assignment Randomized Trial

Junyao Wang*, University of Pittsburgh

8d. Integrated Multiple Adaptive Clinical Trial Design Involving Sample Size Re-Estimation and Response-Adaptive Randomization for Continuous Outcomes

Christine M. Orndahl* and Robert A. Perera, Virginia Commonwealth University

8e. Design of a Calibrated Experiment

Blaza Toman* and Michael A. Nelson, National Institute of Standards and Technology (NIST)

8f. Modified Q-learning with Generalized Estimating Equations for Optimizing Dynamic Treatment Regimes with Repeated-Measures Outcomes

Yuan Zhang*, David Vock and Thomas Murray, University of Minnesota

8g. Development of a Spatial Composite Neighborhood SES Measure

Shanika A. De Silva*, Melissa Meeker, Yasemin Algur and Victoria Ryan, Drexel University; Leann Long, University of Alabama at Birmingham; Nyesha Black, Noire Analytics; Leslie A. McClure, Drexel University

8h. Estimating Disease Prevalence with Potentially Misclassified Dorfman Group Testing Data

Xichen Mou*, University of Memphis; Joshua M. Tebbs and Dewei Wang, University of South Carolina

9. POSTERS: BAYESIAN METHODS

Sponsor: ENAR

9a. Bayesian Spatial Analysis of County-Level Drug Mortality Rates in Virginia

Jong Hyung Lee* and Derek A. Chapman, Virginia Commonwealth University

9b. Robust Partial Reference-Free Cell Composition Estimation in Tissue Expression Profiles

Ziyi Li* and Zhenxing Guo, Emory University; Ying Cheng, Yunnan University; Peng Jin and Hao Wu, Emory University

9c. Multivariate Space-Time Disease Mapping via Quantification of Disease Risk Dependency

Daniel R. Baer* and Andrew B. Lawson, Medical University of South Carolina

9d. Bayesian Envelope in Logistic Regression

Minji Lee* and Zhihua Su, University of Florida

9e. Bayesian Kinetic Modeling for Tracer-Based Metabolomic Data

Xu Zhang*, Ya Su, Andrew N. Lane, Arnold Stromberg, Teresa W-M. Fan and Chi Wang, University of Kentucky

SCIENTIFIC PROGRAM
(CONTINUED)

9f. Forecasting Glaucoma Progression using Bayesian Structural Time Series Analysis

Manoj Pathak*, Murray State University

9g. A Three-Groups Bayesian Approach to GWAS Data with Application to Parkinson's Disease

Vivian Cheng* and Daisy Philtron, The Pennsylvania State University; Ben Shaby, Colorado State University

9h. Improving Estimation of Gene Expression Differences via Integrative Modeling of Transcriptomic and Genetic Data

Xue Zou*, William H. Majoros and Andrew S. Allen, Duke University

9i. Reliable Rates and the Effect of Prior Information with an Application to the County Health Rankings & Roadmaps Program

Guangzi Song*, Harrison Quick and Loni Philip Tabb, Drexel University

10. POSTERS: CAUSAL INFERENCE AND CLINICAL TRIALS

Sponsor: ENAR

10a. The Importance of Propensity Score Estimation to Achieve Balance in Covariates

Hulya Kocyigit*, University of Georgia

10b. Performance of Instrumental Variable and Mendelian Randomization Estimators for Count Data

Phillip Allman*, Hemant Tiwari, Inmaculada Aban and Dustin Long, University of Alabama at Birmingham; Todd MacKenzie, Dartmouth College; Gary Cutter, University of Alabama at Birmingham

10c. Improve Power Analysis in Clinical Trials with Multiple Primary Endpoints: An Application of Parametric Graphical Approaches to Multiple Comparison

Zhe Chen* and Ih Chang, Biogen

10d. Two-Stage Randomized Trial for Testing Treatment Effect for Time to Event Data

Rouba A. Chahine*, Inmaculada Aban and Dustin Long, University of Alabama at Birmingham

10e. Estimating Power for Clinical Trials with Patient Reported Outcomes Endpoints using Item Response Theory

Jinxiang Hu* and Yu Wang, University of Kansas Medical Center

10f. Bayesian Multi-Regional Clinical Trials Using Model Averaging

Nathan W. Bean*, Matthew A. Psioda and Joseph G. Ibrahim, University of North Carolina, Chapel Hill

10g. Constructing Causal Methylation Network by Additive Noise Model

Shudi Li*, Rong Jiao and Momiao Xiong, University of Texas Health Science Center at Houston

10h. Detecting Intervention Effects in a Randomized Trial within a Social Network

Shaina J. Alexandria*, Michael G. Hudgens and Allison E. Aiello, University of North Carolina, Chapel Hill

11. POSTERS: GENOMICS/PROTEOMICS

Sponsor: ENAR

11a. Kernel-Based Genetic Association Analysis for Microbiome Phenotypes

Hongjiao Liu*, University of Washington; Michael C. Wu, Fred Hutchinson Cancer Research Center

11b. True Source of Inflated Zeros in Single Cell Transcriptomics

Tae Kim* and Mengjie Chen, University of Chicago

11c. Estimating Cell Type Composition Using Isoform-Level Gene Expression Data

Hillary M. Heiling* and Douglas R. Wilson, University of North Carolina, Chapel Hill; Wei Sun, Fred Hutchinson Cancer Research Center; Naim Rashid and Joseph G. Ibrahim, University of North Carolina, Chapel Hill

11d. EWAS of Kidney Function Identifies Shared and Ethnic-Specific Loci

Anna Batorsky*, University of North Carolina, Chapel Hill; Mi Kyeong Lee and Stephanie J. London, National Institute of Environmental Health Sciences, National Institutes of Health; Josyf C. Mychaleckyj, University of Virginia; Andrew Marron, Eric A. Whitsel and Nora Franceschini, University of North Carolina, Chapel Hill; Charles E. Breeze, Altius Institute for Biomedical Sciences & University College London

11e. Deconvolutional Mixture Modeling to Account for Cell Type Composition in Tissue Samples

Zachary P. Brehm*, University of Rochester; Marc K. Halushka, Johns Hopkins University; Matthew N. McCall, University of Rochester

11f. Developing a Computational Framework for Precise TAD Boundary Prediction using Genomic Elements

Spiro C. Stilianoudakis* and Shumei Sun, Virginia Commonwealth University

SCIENTIFIC PROGRAM
(CONTINUED)

11g. Parsing Latent Factors in High-Dimensional Classification on Genomic Data

Yujia Pan* and Johann Gagnon-Bartsch, University of Michigan

11h. Estimation of Metabolomic Networks with Gaussian Graphical Models

Katherine Hoff Shutta* and Subhajit Naskar, University of Massachusetts, Amherst; Kathryn M. Rexrode, Harvard Medical School; Denise M. Scholtens, Northwestern University; Raji Balasubramanian, University of Massachusetts, Amherst

11i. Weighted Kernel Method for Integrative Metabolomic and Metagenomic Pathway Analysis

Angela Zhang*, University of Washington; Michael C. Wu, Fred Hutchinson Cancer Research Center

12. POSTERS: FUNCTIONAL DATA/HIGH DIMENSIONAL

Sponsor: ENAR

12a. Dimension Reduction Methods for Multilevel Neural Firing Rate Data

Angel Garcia de la Garza* and Jeff Goldsmith, Columbia University

12b. Amplitude-Phase Separation of Trace-Variogram and its Applications in Spatial Functional Data Analysis

Xiaohan Guo* and Sebastian Kurtek, The Ohio State University; Karthik Bharath, University of Nottingham

12c. Free-Living Walking Strides Segmentation in Wrist-Worn Accelerometry Data

Marta Karas*, Johns Hopkins Bloomberg School of Public Health; Ryan T. Roemmich, Johns Hopkins School of Medicine; Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health; Jacek K. Urbanek, Johns Hopkins School of Medicine

12d. Variable Selections for High-Dimensional Unsupervised Learning with Applications in Genomics and Regulatory Pathway Analysis

Zhipeng Wang* and David Scott, Rice University

12e. Integrative Analysis and Prediction Method for Identifying Subgroup-Specific Omics Biomarkers

Jessica Butts* and Sandra Safo, University of Minnesota

12f. A Novel FWER Controlling Procedure for Data with Generalized Reduced Rank Correlation Structure

Jiatong Sui* and Xing Qiu, University of Rochester

12g. Analyzing Accelerometer Data with Probability Magnitude Graphs

Margaret Banker* and Peter X.K. Song, University of Michigan

12h. Normalization of Minute-Level Activity Counts from Chest- and Wrist-Worn Accelerometers: An Example of Actiheart, Actiwatch, and Actigraph

Vadim Zipunnikov*, Johns Hopkins University; Jiawei Bai, Johns Hopkins Bloomberg School of Public Health

13. POSTERS: BAYESIAN, CLUSTERED DATA, HYPOTHESIS TESTING

Sponsor: ENAR

13a. Bayesian Mechanism for Categorical Data with Data Augmentation Strategy

Arinjita Bhattacharyya*, Subhadip Pal, Riten Mitra and Shesh N. Rai, University of Louisville

13b. False Coverage Rate-Adjusted Smoothed Bootstrap Simultaneous Confidence Intervals for Selected Parameters

Jing Sun*, Santu Ghosh and Varghese George, Augusta University

13c. A State-Space Approach in Handling Challenges Associated with Longitudinal Continuous Neuropsychological Outcomes

Alicia S. Chua*, Boston University School of Public Health; Yorghos Tripodis, Boston University School of Public Health & Boston University School of Medicine

13d. Combining Dependent P-values with a Quantile-Based Approach

Yu Gu*, Michael P. McDermott and Xing Qiu, University of Rochester

13e. Bayesian Estimation for Parameters of Nonlinear Multilevel Models under Burr Distributions

Mohan D. Pant*, Ismail E. Moudden and Jiangtao Luo, Eastern Virginia Medical School

13f. A Flexible and Nearly Optimal Sequential Testing Approach to Randomized Testing: QUICK-STOP

Julian Erik Hecker*, Brigham and Women's Hospital and Harvard Medical School; Ingo Ruczinski, Johns Hopkins Bloomberg School of Public Health; Michael M. Cho and Edwin Silverman, Brigham and Women's Hospital and Harvard Medical School; Brent Coull and Christoph Lange, Harvard T.H. Chan School of Public Health

13g. A Weighted Jackknife Approach Using Linear Model-Based Estimates for Clustered Data

Yejin Choi*, University of New Mexico; Ruofei Du, University of New Mexico Comprehensive Cancer Center

SCIENTIFIC PROGRAM
(CONTINUED)

14. POSTERS: HIGH-DIMENSIONAL DATA, MISSING DATA AND MORE

Sponsor: ENAR

14a. Predicting Latent Contacts from Self-Reported Social Network Data via Outcome Misclassification Adjustment

Qiong Wu*, Tianzhou Ma and Shuo Chen, University of Maryland

14b. Validate Surrogate Endpoints with Continuous and Survival Setup

Idris Demirsoy*, Florida State University; Helen Li, Regeneron Pharmaceutical

14c. New Two-Step Test for Mediation Analysis with Sets of Biomarkers

Andriy Derkach*, Memorial Sloan Kettering Cancer Center; Joshua Sampson, National Cancer Institute, National Institutes of Health; Simina Boca, Georgetown University Medical Center

14d. Meta-Analysis of Binary or Continuous Outcomes Combining Individual Patient Data and Aggregate Data

Neha Agarwala* and Anindya Roy, University of Maryland, Baltimore County

14e. A Post-Processing Algorithm for Building Longitudinal Medication Dose Data from Extracted Medication Information Using Natural Language Processing from Electronic Health Records

Elizabeth McNeer*, Cole Beck, Hannah L. Weeks, Michael L. Williams, Nathan T. James and Leena Choi, Vanderbilt University Medical Center

14f. Power and Sample Size Analysis using Various Statistical Methods in a Tumor Xenograft Study

Sheau-Chiann Chen* and Gregory D. Ayers, Vanderbilt University Medical Center; Rebecca L. Shattuck-Brandt and Ann Richmond, Vanderbilt University, Department of Veterans Affairs and Tennessee Valley Healthcare System; Yu Shyr, Vanderbilt University Medical Center

14g. Estimation and Outliers for Overdispersed Multinomial Data

Barry William McDonald*, Massey University

14h. Partial Least Squares Regression-Based Framework for Incomplete Observations in Environmental Mixture Data Analysis

Ruofei Du*, University of New Mexico Comprehensive Cancer Center

14i. Marginalized Zero-Inflated Negative Binomial Regression Model with Random Effects: Estimating Overall Treatment Effect on Lesion Counts among Multiple Sclerosis Patients (CombiRx Trial)

Steve B. Ampah*, Lloyd J. Edwards, Leann D. Long, Byron C. Jaeger and Nengjun Yi, University of Alabama at Birmingham

15. POSTERS: CONSULTING, EDUCATION, POLICY, EPIDEMIOLOGY

Sponsor: ENAR

15a. Semiparametric Shape Restricted Mixed Effect Regression Spline with Application on US Urban Birth Cohort Study Data and State-Wide Prenatal Screening Program Data

Qing Yin*, University of Pittsburgh

15b. Development and Validation of Models to Predict Foodborne Pathogen Presence and Fecal Indicator Bacteria Levels in Agricultural Water using GIS-Based, Data-Driven Approaches

Daniel L. Weller* and Tanzy Love, University of Rochester; Alexandra Belias and Martin Wiedmann, Cornell University

15c. Accounting for Competing Risks in Estimating Hospital Readmission Rates

John D. Kalbfleisch* and Kevin Zhi He, University of Michigan; Douglas E. Schaubel, University of Pennsylvania; Wenbo Wu, University of Michigan

15d. A New Framework for Cost-Effectiveness Analysis with Time-Varying Treatment and Confounding

Nicholas A. Illenberger*, University of Pennsylvania; Andrew J. Spieker, Vanderbilt University Medical Center; Nandita Mitra, University of Pennsylvania

15e. Rethinking the Introductory Biostatistics Curriculum for Non-Biostatisticians

Emily Slade*, University of Kentucky

15f. Establishing Successful Collaboration in a Competitive Environment: Case Studies from a Healthcare Setting

Jay Mandrekar*, Mayo Clinic

15g. Likelihood Ratios to Compare the Statistical Performance of Multiple Tests in Simulation Studies

Qiuxi Huang*, Boston University School of Public Health

15h. Impact of a Biostatistics Department on an Academic Medical Center

Li Wang*, Henry Domenico and Daniel W. Byrne, Vanderbilt University Medical Center

SCIENTIFIC PROGRAM

(CONTINUED)

16. POSTERS: GENETICS, COMPUTATION

Sponsor: ENAR

16a. Heterogeneity-Aware and Communication-Efficient Distributed Statistical Analysis

Rui Duan*, University of Pennsylvania; Yang Ning, Cornell University; Yong Chen, University of Pennsylvania

16b. False Discovery Rate Computation and Illustration

Megan C. Hollister* and Jeffrey D. Blume, Vanderbilt University Medical Center

16c. A Modified Genomic Control Method for Genetic Association Analysis Using a Stratified, Cluster Sample

Donald Malec*, John Pleis, Rong Wei, Bill Cai, Yulei He, Hee-Choon Shin and Guangyu Zhang, National Center for Health Statistics

16d. Semiparametric Functional Approach with Link Uncertainty

Young Ho Yun*, Virginia Tech

16e. Multi-Ethnic Phenotype Prediction via Effective Modeling of Genetic Effect Heterogeneity

Lina Yang* and Dajiang Liu, The Pennsylvania State University College of Medicine

16f. High Dimensional Sparse Regression with Auxiliary Data on the Features

Constanza Rojo* and Pixu Shi, University of Wisconsin, Madison; Ming Yuan, Columbia University; Sunduz Keles, University of Wisconsin, Madison

16g. A Unified Linear Mixed Model for Simultaneous Assessment of Familiar Relatedness and Population Structure

Tao Wang*, Medical College of Wisconsin; Paul Livermore Auer and Regina Manansala, University of Wisconsin, Milwaukee; Andrea Rau, GABI, INRA, AgroParisTech and Université Paris-Saclay, France; Nick Devogel, Medical College of Wisconsin

16h. Cubic Kernel Method for Implicit T Central Subspace

Weihang Ren* and Xiangrong Yin, University of Kentucky

16i. ODAH: A One-Shot Distributed Algorithm for Estimating Semi-Continuous Outcomes using EHR Data in Multiple Sites

Mackenzie J. Edmondson*, Chongliang Luo and Rui Duan, University of Pennsylvania; Mitchell Maltenfort and Christopher Forrest, Children's Hospital of Philadelphia; Yong Chen, University of Pennsylvania

17. POSTERS: META-ANALYSIS, MISSING DATA AND MORE

Sponsor: ENAR

17a. Multiple Imputation of Missing Covariates in Meta-Regression using Multivariate Imputation by Chained Equations

Amit K. Chowdhry* and Michael P. McDermott, University of Rochester Medical Center

17b. Test-Inversion Confidence Intervals for Estimands in Contingency Tables Subject to Equality Constraints

Qiansheng Zhu* and Joseph B. Lang, University of Iowa

17c. Bayesian Cumulative Probability Models for Continuous and Mixed Outcomes

Nathan T. James*, Bryan E. Shepherd, Leena Choi, Yuqi Tian and Frank E. Harrell, Jr., Vanderbilt University

17d. R-Squared and Goodness of Fit in the Linear Mixed Model: A Cautionary Tale

Boyi Guo* and Byron C. Jaeger, University of Alabama at Birmingham

17e. On the Optimality of Group Testing Estimation

Sarah Church* and Md S. Warasi, Radford University

17f. Bayesian Wavelet-Packet Historical Functional Linear Models

Mark J. Meyer*, Georgetown University; Elizabeth J. Malloy, American University; Brent A. Coull, Harvard T. H. Chan School of Public Health

17g. EMBVS: An EM-Bayesian Approach for Analyzing High-Dimensional Clustered Mixed Outcomes

Yunusa Olufadi* and E. Olusegun George, University of Memphis

17h. Generalized Additive Dynamic Effect Change Models: An Interpretable Extension of GAM

Yuan Yang*, Jian Kang and Yi Li, University of Michigan

17i. A Functional Generalized Linear Mixed Model for Estimating Dose Response in Longitudinal Studies

Madeleine E. St. Ville*, Clemson University; Andrew W. Bergen, Oregon Research Institute; Carolyn M. Ervin, BioRealm; Christopher McMahan, Clemson University; James W. Baurley, BioRealm; Joe Bible, Clemson University

SCIENTIFIC PROGRAM

MONDAY, MARCH 23

8:30—10:15 a.m.

18. MODERN FUNCTIONAL DATA ANALYSIS

Sponsor: IMS
Organizer: Meng Li, Rice University
Chair: Meng Li, Rice University

8:30	**Minimax Powerful Functional Analysis of Covariance Tests for Longitudinal Genome-Wide Association Studies** Yehua Li*, University of California, Riverside
8:55	**Bayesian Function-on-Scalars Regression for High-Dimensional Data** Daniel R. Kowal* and Daniel C. Bourgeois, Rice University
9:20	**Modern Functional Data Analysis for Biosciences** Ana-Maria Staicu* and Alex Long, North Carolina State University; Meredith King, Northrop Grumman
9:45	**Mean and Covariance Estimation for Functional Snippets** Jane-Ling Wang*, University of California, Davis; Zhenhua Lin, National University of Singapore
10:10	**Floor Discussion**

19. DISTRIBUTED AND PRIVACY-PRESERVING METHODS FOR ELECTRONIC HEALTH RECORDS DATA

Sponsors: ENAR, ASA Section on Statistics in Defense and National Security, ASA Health Policy Statistics Section
Organizer: Lu Tang, University of Pittsburgh
Chair: Joyce Chang, University of Pittsburgh

8:30	**Communication Efficient Federated Learning from Multiple EHRs Databases** Changgee Chang*, Zhiqi Bu and Qi Long, University of Pennsylvania
8:55	**Adaptive Noise Augmentation for Privacy-Preserving Empirical Risk Minimization** Fang Liu* and Yinan Li, University of Notre Dame
9:20	**Generating Poisson-Distributed Differentially Private Synthetic Data** Harrison Quick*, Drexel University
9:45	**dblink: Distributed End-to-End Bayesian Entity Resolution** Rebecca Steorts*, Duke University; Neil Marchant and Ben Rubinstein, University of Melbourne; Andee Kaplan, Colorado State University; Daniel Elazar, Australian Bureau of Statistics
10:10	**Floor Discussion**

20. INNOVATIVE STATISTICAL METHODS IN ENVIRONMENTAL MIXTURE ANALYSIS

Sponsors: ENAR, ASA Biometrics Section, ASA Section on Statistics and the Environment, ASA Section on Statistical Learning and Data Science
Organizer: Shanshan Zhao, National Institute of Environmental Health Sciences, National Institutes of Health
Chair: Ling-Wan Chen, National Institute of Environmental Health Sciences, National Institutes of Health

8:30	**Group Inverse-Gamma Gamma Shrinkage for Estimation and Selection in Multipollutant Models** Jonathan Boss*, University of Michigan; Jyotishka Datta, University of Arkansas; Sehee Kim and Bhramar Mukherjee, University of Michigan
8:55	**Bayesian Copula Regression for Inference on Dose-Response Curves** Federico H. Ferrari*, Duke University; Stephanie M. Engel, University of North Carolina, Chapel Hill; David B. Dunson and Amy H. Herring, Duke University
9:20	**Do Males Matter? A Couple-Based Statistical Model for Association Between Environmental Exposures to Pollutants and Infertility** Zhen Chen*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
9:45	**Accommodating Assay Limit-of-Detection in Environmental Mixture Analysis** Jason P. Fine*, University of North Carolina, Chapel Hill; Ling-Wan Chen and Shanshan Zhao, National Institute of Environmental Health Sciences, National Institutes of Health
10:10	**Floor Discussion**

21. MENTORING THROUGHOUT A LIFETIME: CONSIDERATIONS FOR MENTORS AND MENTEES AT ALL CAREER STAGES

Sponsors: ENAR, ENAR Regional Advisory Board (RAB)
Organizer: Naomi Brownstein, Moffitt Cancer Center
Chair: Emily Butler, GlaxoSmithKline

8:30	**Panel Discussion:** Leslie McClure, Drexel University Brian Millen, Eli Lilly and Company Dionne Price, U.S. Food and Drug Administration Manisha Desai, Stanford University
10:10	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

22. INNOVATIVE STATISTICAL APPROACHES FOR HIGH-DIMENSIONAL OMIC AND MICROBIOMIC DATA

Sponsors: ENAR, ASA Biometrics Section, ASA Statistics in Genomics and Genetics Section
Organizer: Subharup Guha, University of Florida
Chair: Zhigang Li, University of Florida

8:30	**Advances and Challenges in Single Cell RNA-Seq Analysis** Susmita Datta*, University of Florida; Michael Sekula and Jeremy Gaskins, University of Louisville
8:55	**Predicting DNA Methylation from Genetic Data Lacking Racial Diversity Using Shared Classified Random Effects** J. Sunil Rao* and Hang Zhang, University of Miami; Melinda Aldrich, Vanderbilt University Medical Center
9:20	**Sparse Generalized Dirichlet Distributions for Microbiome Compositional Data** Jyotishka Datta*, University of Arkansas; David B. Dunson, Duke University
9:45	**Bayesian Nonparametric Differential Analysis for Dependent Multigroup Data with Application to DNA Methylation Analyses** Subharup Guha*, University of Florida; Chiyu Gu, Monsanto Company; Veerabhadran Baladandayuthapani, University of Michigan
10:10	**Floor Discussion**

23. BAYESIAN NONPARAMETRICS FOR CAUSAL INFERENCE AND MISSING DATA

Sponsors: ENAR, ASA Bayesian Statistical Science Section, ASA Biometrics Section
Organizer: Antonio Linero, Florida State University
Chair: Yinpu Li, Florida State University

8:30	**Bayesian Nonparametric Models to Address Positivity Assumption Violations in Causal Inference** Jason Roy*, Rutgers University
8:55	**Sensitivity Analysis using Bayesian Additive Regression Trees** Nicole Bohme Carnegie*, Montana State University; Vincent Dorie, Columbia University; Masataka Harada, Fukuoka University; Jennifer Hill, New York University
9:20	**Variable Selection in Bayesian Nonparametric Models for High-Dimensional Confounding** Michael J. Daniels* and Kumaresh Dhara, University of Florida; Jason Roy, Rutgers University
9:45	**Accelerated Bayesian G-Computation Algorithms** Antonio R. Linero*, University of Texas, Austin
10:10	**Floor Discussion**

24. CONTRIBUTED PAPERS: VARIABLE SELECTION: HOW TO CHOOSE?

Sponsor: ENAR
Chair: Nicole B. Carnegie, Montana State University

8:30	**Sparse Nonparametric Regression with Regularized Tensor Product Kernel** Hang Yu*, University of North Carolina, Chapel Hill; Yuanjia Wang, Columbia University; Donglin Zeng, University of North Carolina, Chapel Hill
8:45	**Pursuing Sources of Heterogeneity in Mixture Regression** Yan Li*, University of Connecticut; Chun Yu, Jiangxi University of Finance and Economics; Yize Zhao, Yale University; Weixin Yao, University of California, Riverside; Robert H. Aseltine and Kun Chen, University of Connecticut
9:00	**An Investigation of Fully Relaxed Lasso and Second-Generation P-Values for High-Dimensional Feature Selection** Yi Zuo* and Jeffrey D. Blume, Vanderbilt University School of Medicine
9:15	**Adaptive Lasso for the Cox Regression with Interval Censored and Possibly Left Truncated Data** Chenxi Li*, Michigan State University; Daewoo Pak, University of Texas MD Anderson Cancer Center; David Todem, Michigan State University
9:30	**Variable Selection for Model-Based Clustering of Functional Data** Tanzy Love*, University of Rochester; Kyra Singh, Google; Eric Hernady, Jacob Finkelstein and Jacqueline Williams, University of Rochester
9:45	**Inconsistency in Multiple Regression Model Specifications** Changyong Feng*, Bokai Wang and Hongyue Wang, University of Rochester; Xin M. Tu, University of California, San Diego
10:00	**C2pLasso: The Categorical-Continuous Pliable Lasso to Identify Brain Regions Affecting Motor Impairment in Huntington Disease** Rakheon Kim*, Texas A&M University; Samuel Mueller, University of Sidney; Tanya Pamela Garcia, Texas A&M University

SCIENTIFIC PROGRAM
(CONTINUED)

25. CONTRIBUTED PAPERS: FUNCTIONAL DATA ANALYSIS

Sponsor: ENAR
Chair: Owais Gilani, Bucknell University

8:30	**Covariate-Adjusted Hybrid Principal Components Analysis for EEG Data** Aaron Wolfe Scheffler*, University of California, San Francisco; Abigail Dickinson, Shafali Jeste and Damla Senturk, University of California, Los Angeles
8:45	**Evidence-Based Second-Generation P-values on Functional Magnetic Resonance Imaging Data** Ya-Chen Lin* and Valerie F. Welty, Vanderbilt University; Jeffrey D. Blume, Kimberly M. Albert, Brian D. Boyd, Warren D. Taylor and Hakmook Kang, Vanderbilt University Medical Center
9:00	**Modeling Non-Linear Time Varying Dependence with Application to fMRI Data** Ivor Cribben*, Alberta School of Business
9:15	**Average Treatment Effect Estimation with Functional Confounders** Xiaoke Zhang* and Rui Miao, The George Washington University
9:30	**Model-based Statistical Depth with Applications to Functional Data** Weilong Zhao* and Zishen Xu, Florida State University; Yun Yang, University of Illinois at Urbana-Champaign; Wei Wu, Florida State University
9:45	**Bayesian Inference for Brain Activity from Multi-Resolution Functional Magnetic Resonance Imaging** Andrew Whiteman*, Jian Kang and Timothy Johnson, University of Michigan
10:00	**Floor Discussion**

26. CONTRIBUTED PAPERS: PENALIZED AND OTHER REGRESSION MODELS WITH APPLICATIONS

Sponsor: ENAR
Chair: Saryet Kucukemiroglu, U.S. Food and Drug Administration

8:30	**On More Efficient Logistic Regression Analysis via Extreme Ranking** Hani Samawi*, Georgia Southern University
8:45	**Penalized Models for Analysis of Multiple Mediators** Daniel J. Schaid* and Jason P. Sinnwell, Mayo Clinic
9:00	**Fitting Equality-Constrained, L1-Penalized Models with Inexact ADMM to Find Gene Pairs** Lam Tran*, Lan Luo and Hui Jiang, University of Michigan
9:15	**A Comparative Analysis of Penalized Linear Mixed Models in Structured Genetic Data** Anna Reisetter* and Patrick Breheny, University of Iowa
9:30	**A Two-Stage Kernel Machine Regression Model for Integrative Analysis of Alpha Diversity** Runzhe Li* and Ni Zhao, Johns Hopkins Bloomberg School of Public Health
9:45	**Penalized Semiparametric Additive Modeling for Group Testing Data** Karl B. Gregory*, Dewei Wang, University of South Carolina; Chris S. McMahan, Clemson University
10:00	**Penalized Likelihood Logistic Regression with Rare Events-An Application to the Regeneration Dynamics of Pine Species in Oak-Pine Forest Types** Dilli Bhatta*, University of South Carolina Upstate

27. CONTRIBUTED PAPERS: METHODS FOR NEUROIMAGING DATA: GET THE PICTURE?

Sponsor: ENAR
Chair: Hao Wang, Johns Hopkins University School of Medicine

8:30	**Letting the LaxKAT Out of the Bag: Packaging, Simulation, and Neuroimaging Data Analysis for a Powerful Kernel Test** Jeremy S. Rubin*, University of Maryland, Baltimore County; Simon Vandekar, Vanderbilt University; Lior Rennert, Clemson University; Mackenzie Edmonson, and Russell T. Shinohara, University of Pennsylvania
8:45	**Comparison of Two Ways of Incorporating the External Information via Linear Mixed Model Design with Application in Brain Imaging** Maria Paleczny*, Institute of Mathematics of the Jagiellonian University
9:00	**Interpretable Classification Methods for Brain-Computer Interface P300 Speller** Tianwen Ma*, Jane E. Huggins and Jian Kang, University of Michigan
9:15	**Copula Random Field with Application to Massive Neuroimaging Data Analysis** Jie He*, Jian Kang and Peter X.-K Song, University of Michigan
9:30	**Neural Networks Guided Independent Component Analysis with Application to Neuroimaging** Daiwei Zhang*, University of Michigan; Ying Guo, Emory University; Jian Kang, University of Michigan

SCIENTIFIC PROGRAM
(CONTINUED)

9:45	**Removal of Scanner Effects in Covariance of Neuroimaging Measures** Andrew Chen*, Haochang Shou and Russell T. Shinohara, University of Pennsylvania
10:00	**Classifying Brain Edema with Low-Resolution MRI** Danni Tu* and Dylan Small, University of Pennsylvania; Manu S. Goyal, Washington University School of Medicine, St. Louis; Theodore Satterthwaite, Kelly Clark and Russell T. Shinohara, University of Pennsylvania

28. CONTRIBUTED PAPERS: CAUSAL EFFECT ESTIMATION

Sponsor: ENAR
Chair: Jan De Neve, Ghent University

8:30	**Assessing Exposure Effects on Gene Expression** Sarah A. Reifeis*, Michael G. Hudgens, Karen L. Mohlke and Michael I. Love, University of North Carolina, Chapel Hill
8:45	**Sensitivity of Clinical Trial Estimands under Imperfect Compliance** Heng Chen*, Southern Methodist University; Daniel F. Heitjan, Southern Methodist University and University of Texas, Southwestern
9:00	**Borrowing from Supplemental Sources to Estimate Causal Effects from a Primary Data Source** Jeffrey A. Boatman*, David M. Vock and Joseph S. Koopmeiners, University of Minnesota
9:15	**Estimating Causal Treatment Effects: A Bayesian Inference Approach Adopting Principal Stratification with Strata Predictive Covariates** Duncan C. Rotich*, University of Kansas Medical Center; Bin Dong, Janssen Research & Development; Jeffrey A. Thompson, University of Kansas Medical Center
9:30	**Estimating Causal Effects in the Presence of Positivity Violations** Yaqian Zhu* and Nandita Mitra, University of Pennsylvania; Jason Roy, Rutgers University
9:45	**Estimating Causal Effect of Multiple Treatments with Censored Data in Observational Studies** Youfei Yu*, Min Zhang and Bhramar Mukherjee, University of Michigan
10:00	**Floor Discussion**

MONDAY, MARCH 23
10:15 a.m.—10:30 a.m.

REFRESHMENT BREAK WITH OUR EXHIBITORS

MONDAY, MARCH 23
10:30 a.m.—12:15 p.m.

29. NEW PERSPECTIVES ON DATA INTEGRATION IN GENOME-WIDE ASSOCIATION STUDIES

Sponsor: IMS
Organizer: Qiongshi Lu, University of Wisconsin, Madison
Chair: Hyunseng Kang, University of Wisconsin, Madison

10:30	**TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits** Jingjing Yang*, Emory University School of Medicine; Sini Nagpal, Georgia Institute of Technology; Xiaoran Meng and Shizhen Tang, Emory University School of Public Health; Gregory C. Gibson, Georgia Institute of Technology; David A. Bennett, Rush University Medical Center; Philip L. De Jager, Columbia University; Aliza P. Wingo, Atlanta VA Medical Center; Thomas S. Wingo and Michael P. Epstein, Emory University School of Medicine
10:55	**Integrating Gene Expression Regulatory Variation Across Populations and Tissues to Understand Complex Traits** Heather E. Wheeler*, Loyola University Chicago
11:20	**Transcriptome-Wide Transmission Disequilibrium Analysis Identifies Novel Risk Genes for Autism Spectrum Disorder** Qiongshi Lu*, Kunling Huang and Yuchang Wu, University of Wisconsin, Madison
11:45	**Model Checking and More Powerful Inference in Transcriptome-Wide Association Studies** Wei Pan*, University of Minnesota
12:10	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

30. ADVANCES IN CAUSAL INFERENCE AND JOINT MODELING WITH SURVIVAL AND COMPLEX LONGITUDINAL DATA

Sponsors: ENAR, ASA Biometrics Section, ASA Statistics in Genomics and Genetics Section
Organizer: Zhigang Li, University of Florida
Chair: James O'Malley, Dartmouth College

10:30	**Causal Proportional Hazards Estimation with a Binary Instrumental Variable** Limin Peng* and Behzad Kianian, Emory University; Jung In Kim and Jason Fine, University of North Carolina, Chapel Hill
11:00	**Joint Modeling of Zero-Inflated Longitudinal Microbiome and Time-to-Event Data** Huilin Li*, Jiyuan Hu and Chan Wang, New York University; Martin Blaser, Rutgers University
11:30	**Causal Comparative Effectiveness Analysis of Dynamic Continuous-Time Treatment Initiation Rules with Sparsely Measured Outcomes and Death** Liangyuan Hu*, Icahn School of Medicine at Mount Sinai; Joseph W. Hogan, Brown University
12:00	**Discussant:** Joseph Hogan, Brown University

31. OPPORTUNITIES AND CHALLENGES IN THE ANALYSIS AND INTEGRATION OF LARGE-SCALE BIOBANK DATA

Sponsors: ENAR, ASA Biometrics Section, ASA Section in Statistics in Epidemiology, ASA Statistics in Genomics and Genetics Section
Organizer: Ryan Sun, University of Texas MD Anderson Cancer Center
Chair: Ryan Sun, University of Texas MD Anderson Cancer Center

10:30	**Empowering GWAS Analysis with Missing Data Using Surrogate Phenotypes in Biobanks** Xihong Lin*, Harvard University; Zachary McCaw, Google
10:55	**Fast and Efficient Generalized Estimating Equations for Fitting Non-Linear Model to Biobank Scale Data** Nilanjan Chatterjee* and Diptavo Dutta, Johns Hopkins University
11:20	**Modeling Functional Enrichment Improves Polygenic Prediction Accuracy in UK Biobank and 23andMe Data Sets** Carla Marquez-Luna*, Icahn School of Medicine at Mount Sinai; Steven Gazal, Harvard T.H. Chan School of Public Health; Po-Ru Loh, Brigham and Women's Hospital and Harvard Medical School; Samuel S. Kim, Massachusetts Institute of Technology; Nicholas Furlotte and Adam Auton, 23andMe Inc; Alkes L. Price, Harvard T.H. Chan School of Public Health
11:45	**Handling Sampling and Selection Bias in Association Studies Embedded in Electronic Health Records** Bhramar Mukherjee* and Lauren J. Beesley, University of Michigan
12:10	**Floor Discussion**

32. COMPOSITIONAL NATURE OF MICROBIOME DATA: CHALLENGES AND NEW METHODS

Sponsors: ENAR, ASA Biometrics Section, ASA Section in Statistics in Epidemiology, ASA Statistics in Genomics and Genetics Section
Organizer: Michael Sohn, University of Rochester
Chair: Michael Sohn, University of Rochester

10:30	**Association Testing for Longitudinal Multiomics Data** Anna M. Plantinga*, Williams College
11:00	**Scalable Inference for Count Compositional Microbiome Data** Justin D. Silverman*, Duke University
11:30	**Robust and Powerful Differential Composition Tests on Clustered Microbiome Data** Zhengzheng Tang* and Guanhua Chen, University of Wisconsin, Madison
12:00	**Discussant:** Hongzhe Li, University of Pennsylvania

33. STATISTICAL MODELING IN ALZHEIMER'S DISEASE

Sponsors: ENAR, ASA Health Policy Statistics Section
Organizer: Guoqiao Wang, Washington University in St. Louis
Chair: Chengjie Xiong, Washington University in St. Louis

10:30	**Bent Lines and Quantiles in Longitudinal Modeling of Alzheimer's Progression** Rick Chappell*, University of Wisconsin, Madison
10:55	**Partly Conditional Modeling for Ordinal Outcomes with Application to Alzheimer's Disease Progression** Dandan Liu* and Jacquelyn Neal, Vanderbilt University
11:20	**Leveraging Disease Progression Modeling to Improve Clinical Trial Design in Alzheimer's Disease** Barbara Wendelberger*, Melanie Quintana and Scott Berry, Berry Consultants
11:45	**Integrative Modeling and Dynamic Prediction of Alzheimer's Disease** Sheng Luo*, Duke University; Kan Li, Merck & Co., Inc.
12:00	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

34. RECENT ADVANCES IN BAYESIAN METHODS FOR SPATIAL-TEMPORAL PROCESSES

Sponsors: ENAR, ASA Bayesian Statistical Science Section, ASA Section in Statistics in Epidemiology
Organizer: Zehang Li, Yale School of Public Health
Chair: Howard H. Chung, Emory University

10:30	**Multivariate Disease Mapping using Directed Acyclic Graph Autoregressive Models** Abhi Datta*, Johns Hopkins University
10:55	**Modeling Heroin-Related EMS Calls in Space and Time** Zehang Richard Li*, Forrest Crawford and Gregg Gonsalves, Yale School of Public Health
11:20	**Bayesian Spatial Prediction of Collective Efficacy Across an Urban Environment** Catherine Calder*, University of Texas, Austin
11:45	**Estimating Subnational Variation in Health Indicators in a Low- and Medium-Income Countries Setting** Jon Wakefield*, University of Washington
12:00	**Floor Discussion**

35. SPEED POSTERS: EHR DATA, EPIDEMIOLOGY, PERSONALIZED MEDICINE, CLINICAL TRIALS

Sponsor: ENAR
Chair: Chenguang Wang, Johns Hopkins University

35a. INVITED POSTER: Extending Difference-in-Difference Methods to Test the Impact of State-Level Marijuana Laws on Substance Use Using Published Prevalence Estimates

Christine M. Mauro* and Melanie M. Wall, Columbia University Mailman School of Public Health

35b. INVITED POSTER: Methods of Analysis when an Outcome Variable is a Prediction with Berkson Error

Pamela A. Shaw*, University of Pennsylvania; Paul Gustafson, University of British Columbia; Daniela Sotres-Alvarez, University of North Carolina, Chapel Hill; Victor Kipnis, National Cancer Institute, National Institutes of Health; Laurence Freedman, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center

35c. Confidence Intervals for the Youden Index and Its Optimal Cut-Off Point in the Presence of Covariates

Xinjie Hu*, Gengsheng Qin and Chenxue Li, Georgia State University; Jinyuan Chen, Lanzhou University

35d. Critical Window Variable Selection for Pollution Mixtures

Joshua L. Warren*, Yale University

WITHDRAWN

35e. Learning Individualized Treatment Rules for Multiple-Domain Latent Outcomes

Yuan Chen*, Columbia University; Donglin Zeng, University of North Carolina, Chapel Hill; Yuanjia Wang, Columbia University

35f. Semi-Parametric Efficient Prediction of Binary Outcomes when Some Predictors are Incomplete via Post-Stratification

Yaqi Cao*, University of Pennsylvania; Sebastien Haneuse, Harvard T.H. Chan School of Public Health; Yingye Zheng, Fred Hutchinson Cancer Research Center; Jinbo Chen, University of Pennsylvania

35g. Optimal Sampling Plans for Functional Linear Regression Models

Hyungmin Rha*, Ming-Hung Kao and Rong Pan, Arizona State University

35h. Optimal Experimental Design for Big Data: Applications in Brain Imaging

Eric W. Bridgeford*, Shangsi Wang, Zeyi Wang, Brian Caffo and Joshua Vogelstein, Johns Hopkins University

35i. New Statistical Learning for Evaluating Nested Dynamic Treatment Regimes with Test-and-Treat Observational Data

Ming Tang*, Lu Wang and Jeremy M.G. Taylor, University of Michigan

35j. A Sequential Strategy for Determining Confidence in Individual Treatment Decisions in Personalized Medicine

Nina Orwitz*, Eva Petkova and Thaddeus Tarpey, New York University

35k. Hidden Analyses: A Systematic Framework of Data Analyses Prior to Statistical Modeling and Recommendations for More Transparent Reporting

Marianne Huebner*, Michigan State University; Werner Vach, University Hospital Basel, Switzerland; Saskia le Cessie, Leiden University Medical Center, Netherlands; Carsten Schmidt, University Medicine of Greifswald, Germany; Lara Lusa, University of Primorksa, Slovenia

35l. A Bayesian Adaptive Design for Early Phase Biomarker Discovery Study

Yi Yao*, Ying Yuan and Liang Li, University of Texas MD Anderson Cancer Center

35m. Association Between Tooth Loss and Cancer Mortality: NHANES 1999-2015

Xiaobin Zhou*, Agnes Scott College; Kelli O'Connell and Mengmeng Du, Memorial Sloan Kettering Cancer Center

SCIENTIFIC PROGRAM
(CONTINUED)

36. CONTRIBUTED PAPERS: ADAPTIVE DESIGNS FOR CLINICAL TRIALS

Sponsor: ENAR
Chair: Jingshu O. Wang, The University of Chicago

10:30	**Keyboard Design for Phase I Drug-Combination Trials** Haitao Pan*, St. Jude Children's Research Hospital; Ruitao Lin and Ying Yuan, University of Texas MD Anderson Cancer Center
10:45	**Interim Adaptive Decision-Making for Small n Sequential Multiple Assignment Randomized Trial** Yan-Cheng Chao* and Thomas M. Braun, University of Michigan; Roy N. Tamura, University of South Florida; Kelley M. Kidwell, University of Michigan
11:00	**Bayesian Adaptive Enrichment Trial Design for Continuous Predictive Biomarkers with Possibly Non-Linear or Non-Monotone Effects** Yusha Liu*, Rice University; Lindsay Ann Renfro, University of Southern California
11:15	**Robust Blocked Response-Adaptive Randomization Designs** Thevaa Chandereng* and Rick Chappell, University of Wisconsin, Madison
11:30	**Streamlined Hyperparameter Tuning in Mobile Health** Marianne Menictas*, Harvard University
11:45	**A Two-Stage Sequential Design for Selecting the t Best Treatments** Mingyue Wang* and Pinyuen Chen, Syracuse University
12:00	**Adaptive Monitoring: Optimal Burn-in to Control False Discoveries Allowing Unlimited Monitoring** Jonathan J. Chipman*, Huntsman Cancer Institute, University of Utah; Jeffrey D. Blume and Robert A. Greevy, Jr., Vanderbilt University

37. CONTRIBUTED PAPERS: BAYESIAN SEMIPARAMETRIC AND NONPARAMETRIC METHODS

Sponsor: ENAR
Chair: Ana-Maria Staicu, North Carolina State University

10:30	**Heterogeneity Pursuit for Spatial Point Process with Applications: A Bayesian Semiparametric Recourse** Jieying Jiao*, Guanyu Hu and Jun Yan, University of Connecticut
10:45	**A Bayesian Finite Mixture Model-Based Clustering Method with Variable Selection for Identifying Disease Phenotypes** Shu Wang*, University of Florida
11:00	**A Bayesian Nonparametric Model for Zero-Inflated Outcomes: Prediction, Clustering, and Causal Estimation** Arman Oganisian* and Nandita Mitra, University of Pennsylvania; Jason A. Roy, Rutgers University
11:15	**Longitudinal Structural Topic Models for Estimating Latent Health Trajectories using Administrative Claims Data** Mengbing Li* and Zhenke Wu, University of Michigan
11:30	**Novel Semiparametric Bayesian Methods for the Competing Risks Data with Length-Biased Sampling** Tong Wang*, Texas A&M University
11:45	**A Bayesian Nonparametric Approach for Estimating Causal Effects for Longitudinal Data** Kumaresh Dhara* and Michael J. Daniels, University of Florida
12:00	**Floor Discussion**

38. CONTRIBUTED PAPERS: STATISTICAL METHODS IN CANCER RESEARCH

Sponsor: ENAR
Chair: Ivor Cribben, Alberta School of Business

10:30	**Identifying Gene-Environment Interactions Using Integrative Multidimensional Omics Data for Cancer Outcomes** Yaqing Xu*, Yale University; Mengyun Wu, Shanghai University of Finance and Economics; Shuangge Ma, Yale University
10:45	**Bayesian Modeling of Metagenomic Sequencing Data for Discovering Microbial Biomarkers in Colorectal Cancer Detection** Shuang Jiang*, Southern Methodist University; Qiwei Li, University of Texas, Dallas; Andrew Y. Koh, Guanghua Xiao and Xiaowei Zhan, University of Texas Southwestern Medical Center
11:00	**Propensity Score Methods in the Presence of Missing Covariates** Kay See Tan*, Memorial Sloan Kettering Cancer Center
11:15	**Pathway-Structured Predictive Modeling for Multi-Level Drug Response in Multiple Myeloma** Xinyan Zhang*, Georgia Southern University; Bingzong Li and Wenzhuo Zhuang, Soochow University; Nengjun Yi, University of Alabama at Birmingham
11:30	**Integrative Network Based Analysis of Metabolomic and Transcriptomic Data for Understanding Biological Mechanism of Lung Cancer** Christopher M. Wilson*, Brooke L. Fridley and Doug W. Cress, Moffitt Cancer Center; Farnoosh Abbas Aghabazadeh, Princess Margaret Cancer Centre

SCIENTIFIC PROGRAM
(CONTINUED)

11:45	**A General Framework for Multi-Gene, Multi-Cancer Mendelian Risk Prediction Models** Jane W Liang*, Harvard T.H. Chan School of Public Health; Gregory Idos, Christine Hong and Stephen B. Gruber, University of Southern California Norris Comprehensive Cancer Center; Giovanni Parmigiani and Danielle Braun, Dana-Farber Cancer Institute
12:00	**The Impact of Design Misspecification in Oncology Trials with Survival Endpoint** Tyler Zemla* and Jennifer Le-Rademacher, Mayo Clinic

39. CONTRIBUTED PAPERS: NETWORK ANALYSIS: CONNECTING THE DOTS

Sponsor: ENAR
Chair: Maiying Kong, University of Louisville

10:30	**Bayesian Assessment of Homogeneity and Consistency for Network Meta-Analysis** Cheng Zhang*, Hao Li and Ming-Hui Chen, University of Connecticut; Joseph G. Ibrahim, University of North Carolina, Chapel Hill; Arvind K. Shah and Jianxin Lin, Merck & Co., Inc.
10:45	**Bayesian Community Detection for Multiple Networks** Luoying Yang* and Zhengwu Zhang, University of Rochester Medical Center
11:00	**Semi-Parametric Bayes Regression with Network Valued Covariates** Xin Ma*, Suprateek Kundu and Jennifer Stevens, Emory University
11:15	**Scalable Network Estimation with L0 Penalty** Junghi Kim*, U.S. Food and Drug Administration; Hongtu Zhu, Univ[illegible] Chapel Hill; Xiao Wang, Purdue University; Kim-Anh Do, University of Texas MD Anderson Cancer Center **WITHDRAWN**
11:30	**Disease Prediction by Integrating Marginally Weak Signals and Local Predictive Gene/Brain Networks** Yanming Li*, University of Michigan
11:45	**Scalar-on-Network Regression Via Gradient Boosting** Emily Morris* and Jian Kang, University of Michigan
12:00	**Floor Discussion**

40. CONTRIBUTED PAPERS: POLICIES AND POLITICS: STATISTICAL ANALYSES OF HEALTH OUTCOMES IN THE REAL WORLD

Sponsor: ENAR
Chair: Ciprian M. Crainiceanu, Johns Hopkins University

10:30	**The Challenges of Electronic Health Record Use to Estimate Individualized Type 2 Diabetes Treatment Strategies** Erica EM Moodie* and Gabrielle Simoneau, McGill University
10:45	**Incorporating Statistical Methods to Address Spatial Confounding in Large EHR Data Studies** Jennifer Bobb* and Andrea Cook, Kaiser Permanente Washington
11:00	**A Spatial Causal Analysis of Wildland Fire-Contributed PM2.5 Using Numerical Model Output** Alexandra E. Larsen*, Duke University School of Medicine; Shu Yang and Brian J. Reich, North Carolina State University; Ana Rappold, U.S. Environmental Protection Agency
11:15	**Propensity Score Matching with Time-Varying Covariates: An Application in the Prevention of Recurrent Preterm Birth** Erinn M. Hade*, Giovanni Nattino, Heather A. Frey and Bo Lu, The Ohio State University
11:30	**A Bayesian Spatio-Temporal Abundance Model for Surveillance of the Opioid Epidemic** David M. Kline*, The Ohio State University; Lance A. Waller, Emory University; Staci A. Hepler, Wake Forest University
11:45	**Health Co-Benefits of the Implementation of Global Climate Mitigation Commitments** Gavin Shaddick*, University of Exeter
12:00	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

41. CONTRIBUTED PAPERS: STATISTICAL CONSIDERATIONS FOR OPTIMAL TREATMENT

Sponsor: ENAR
Chair: Andrada E. Ivanescu, Montclair State University

10:30	**Optimal Treatment Regime Estimation using Pseudo Observation with Censored Data** Taehwa Choi* and Sangbum Choi, Korea University
10:45	**Boosting Algorithms for Estimating Optimal Individualized Treatment Rules** Duzhe Wang*, University of Wisconsin, Madison; Haoda Fu, Eli Lilly and Company; Po-Ling Loh, University of Wisconsin, Madison
11:00	**Capturing Heterogeneity in Repeated Measures Data by Fusion Penalty** Lili Liu*, Shandong University and Washington University in St. Louis; Lei Liu, Washington University in St. Louis
11:15	**Optimal Individualized Decision Rules Using Instrumental Variable Methods** Hongxiang Qiu* and Marco Carone, University of Washington; Ekaterina Sadikova, Maria Petukhova and Ronald C. Kessler, Harvard Medical School; Alex Luedtke, University of Washington
11:30	**Sample Size and Timepoint Tradeoffs for Comparing Dynamic Treatment Regimens in a Longitudinal SMART** Nicholas J. Seewald* and Daniel Almirall, University of Michigan
11:45	**Floor Discussion**

MONDAY, MARCH 23
12:15 p.m. — 1:30 p.m.

ROUNDTABLE LUNCHEONS

MONDAY, MARCH 23
1:45 p.m. — 3:30 p.m.

42. CAUSAL INFERENCE WITH GENETIC DATA

Sponsor: IMS
Organizer: Qingyuan Zhao, University of Cambridge
Chair: Richard Charnigo, University of Kentucky

1:45	**Estimating Causal Relationship for Complex Traits with Weak and Heterogeneous Genetic Effects** Jingshu Wang*, The University of Chicago; Qingyuan Zhao, University of Cambridge; Jack Bowden, Gibran Hemani and George Davey Smith, University of Bristol; Nancy R. Zhang and Dylan Small, University of Pennsylvania
2:15	**Distinguishing Genetic Correlation from Causation in GWAS** Luke J. O'Connor*, Broad Institute; Alkes L. Price, Harvard T.H. Chan School of Public Health
2:45	**Robust Methods with Two-Sample Summary Data Mendelian Randomization** Hyunseung Kang*, University of Wisconsin, Madison
3:15	**Discussant:** Qingyuan Zhao, University of Cambridge

43. RECENT ADVANCES IN STATISTICAL METHODS FOR SINGLE-CELL OMICS ANALYSIS

Sponsor: IMS
Organizer: Yuchao Jiang, University of North Carolina, Chapel Hill
Chair: Rhonda Bacher, University of Florida

1:45	**Fast and Accurate Alignment of Single-Cell RNA-seq Samples Using Kernel Density Matching** Mengjie Chen*, Yang Li and Qi Zhan, The University of Chicago
2:10	**Novel Computational Methods for Analyzing Single Cell Multi-Omics Data** Wei Chen*, University of Pittsburgh
2:35	**DNA Copy Number Profiling: From Bulk to Single-Cell Sequencing** Yuchao Jiang*, University of North Carolina, Chapel Hill
3:00	**Statistical Analysis of Spatial Expression Pattern for Spatially Resolved Transcriptomic Studies** Xiang Zhou*, Shiquan Sun and Jiaqiang Zhu, University of Michigan
3:25	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

44. RECENT ADVANCES IN MICROBIOME DATA ANALYSIS

Sponsor: IMS
Organizer: Anru Zhang, University of Wisconsin, Madison
Chair: Chi Zhang, Indiana University

1:45	**Incorporating Auxiliary Information to Improve Microbiome-Based Prediction Models** Michael C. Wu*, Fred Hutchinson Cancer Research Center
2:10	**Estimation and inference with non-random missing data and latent factors** Christopher McKennan*, University of Pittsburgh
2:35	**Statistical Methods for Tree Structured Microbiome Data** Hongyu Zhao*, Yale University; Tao Wang and Yaru Song, Shanghai Jiao Tong University; Can Yang, Hong Kong University of Science and Technology
3:00	**High-Dimensional Log-Error-in-Variable Regression with Applications to Microbial Compositional Data Analysis** Anru Zhang*, Pixu Shi and Yuchen Zhou, University of Wisconsin, Madison
3:25	**Floor Discussion**

45. NOVEL METHODS TO EVALUATE SURROGATE ENDPOINTS

Sponsors: ENAR, ASA Biometrics Section, ASA Biopharmaceutical Section
Organizer: Ludovic Trinquart, Boston University School of Public Health
Chair: Michael LaValley, Boston University School of Public Health

1:45	**Using a Surrogate Marker for Early Testing of a Treatment Effect** Layla Parast*, RAND; Tianxi Cai, Harvard University; Lu Tian, Stanford University
2:10	**Mediation Analysis with Illness-Death Model for Right-Censored Surrogate and Clinical Outcomes** Isabelle Weir*, Harvard T.H. Chan School of Public Health; Jennifer Rider and Ludovic Trinquart, Boston University
2:35	**Incorporating Patient Subgroups During Surrogate Endpoint Validation** Emily Roberts*, Michael Elliott and Jeremy MG Taylor, University of Michigan
3:00	**Assessing a Surrogate Predictive Value: A Causal Inference Approach** Ariel Alonso Abad*, University of Leuven; Wim Van der Elst, Janssen Pharmaceutica; Geert Molenberghs University of Leuven
3:25	**Floor Discussion**

46. RECENT ADVANCES IN THE UNCERTAINTY ESTIMATION AND PROPERTIES OF BAYESIAN ADDITIVE REGRESSION TREES

Sponsors: ENAR, IMS, ASA Bayesian Statistical Science Section, ASA Section on Statistical Learning and Data Science
Organizer: Yaoyuan Vincent Tan, Rutgers School of Public Health
Chair: Chanmin Kim, Boston University School of Public Health

1:45	**Heteroscedastic BART via Multiplicative Regression Trees** Matthew T. Pratola*, The Ohio State University; Hugh A. Chipman, Acadia University; Edward I. George, University of Pennsylvania; Robert E. McCulloch, Arizona State University
2:10	**Bayesian Nonparametric Modeling with Tree Ensembles for Predicting Patient Outcomes** Robert E. McCulloch*, Arizona State University; Rodney Sparapani, Purushottam Laud and Brent Logan, Medical College of Wisconsin
2:35	**Bayesian Decision Tree Ensembles in Fully Nonparametric Problems** Yinpu Li*, Florida State University; Antonio Linero, University of Texas, Austin; Junliang Du, Florida State University
3:00	**On Theory for BART** Veronika Rockova* and Enakshi Saha, The University of Chicago
3:25	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

47. CURRENT DEVELOPMENTS IN ANALYZING EHR AND BIOBANK DATA

Sponsor: ENAR
Organizer: Xue Zhong, Vanderbilt University
Chair: Xue Zhong, Vanderbilt University

1:45	**Adventures with Large Biomedical Datasets: Diseases, Medical Records, Environment and Genetics** Andrey Rzhetsky*, The University of Chicago
2:10	**Association Analysis of Biobank Scale Data Using Minimal Sufficient Statistics** Dajiang Liu*, Penn State College of Medicine
2:35	**Use of Electronic Health Records and a Biobank for Pharmacogenomic Studies: Promises and Challenges** Leena Choi*, Vanderbilt University Medical Center
3:00	**Assessing the Progress of Alzheimer's Disease Via Electronic Medical Records** Zhijun Yin*, Vanderbilt University Medical Center
3:25	**Floor Discussion**

48. SPEED POSTERS: CAUSAL INFERENCE/ LONGITUDINAL DATA/HIGH-DIMENSIONAL DATA/MASSIVE DATA

Sponsor: ENAR

Chair: Yong Lin, Rutgers University

48a. INVITED POSTER: Bipartite Causal Inference with Interference for Evaluating Air Pollution Regulations

Corwin M. Zigler*, University of Texas, Austin and Dell Medical School

48b. Doubly Robust Estimation of Causal Effects with Covariate-Balancing Propensity Score and Machine-Learning-Based Outcome Prediction

Byeong Yeob Choi*, University of Texas Health Science Center at San Antonio

48c. Percentile-Based Residuals for Model Assessment

Sophie Berube*, Abhirup Datta, Chenguang Wang, Qingfeng Li and Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health

48d. Change-Point Detection in Multivariate Time Series

Tong Shen*, University of California, Irvine; Xu Gao, Google; Hernando Ombao, King Abdullah University of Science and Technology; Zhaoxia Yu, University of California, Irvine

48e. Approaches for Modeling Spatially Varying Associations Between Multi-Modal Images

Alessandra M. Valcarcel*, University of Pennsylvania; Simon N. Vandekar, Vanderbilt University; Tinashe Tapera, Azeez Adebimpe and David Roalf, University of Pennsylvania; Armin Raznahan, National Institute of Mental Health, National Institutes of Health; Theodore Satterthwaite, Russell T. Shinohara and Kristin Linn, University of Pennsylvania

48f. Generalizing Trial Findings using Nested Trial Designs with Sub-Sampling of Non-Randomized Individuals

Sarah E. Robertson* and Issa J. Dahabreh, Brown University; Miguel A. Hernan, Harvard University; Ashley L. Buchanan, University of Rhode Island; Jon A. Steingrimsson, Brown University

48g. Causal Inference with Multiple Mediators in a Survival Context

Hui Zeng* and Vernon Michael Chinchilli, The Pennsylvania State University

48h. Adjusting for Compliance in SMART Designs

William Jeremy Artman*, Ashkan Ertefaie and Brent Johnson, University of Rochester

48i. Statistical Inference for Cox Proportional Hazards Model with a Diverging Number of Covariates

Lu Xia*, University of Michigan; Bin Nan, University of California, Irvine; Yi Li, University of Michigan

48j. Bayesian Focal-Area Detection for Multi-Class Dynamic Model with Application to Gas Chromatography

Byung-Jun Kim*, Virginia Tech

48k. The Survivor Separable Effects

Mats Julius Stensrud* and Miguel Hernan, Harvard T. H. Chan School of Public Health; Jessica Julius Young, Harvard Medical School

48l. Adjusted Cox Scores for GWAS and PheWAS Screening in R

Elizabeth A. Sigworth*, Ran Tao, Frank Harrell and Qingxia Chen, Vanderbilt University

48m. Microbiome Quantile Regression

Myung Hee Lee*, Weill Cornell Medical College

SCIENTIFIC PROGRAM

(CONTINUED)

49. CONTRIBUTED PAPERS: STATISTICAL METHODS FOR OMICS DATA ANALYSIS

Sponsor: ENAR
Chair: Yehua Li, University of California, Riverside

1:45	**Mean-Correlation Relationship Biases Co-Expression Analysis** Yi Wang* and Stephanie C. Hicks, Johns Hopkins Bloomberg School of Public Health; Kasper D. Hansen, Johns Hopkins Bloomberg School of Public Health and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine
2:00	**Efficient Detection and Classification of Epigenomic Changes Under Multiple Conditions** Pedro L. Baldoni*, Naim U. Rashid and Joseph G. Ibrahim, University of North Carolina, Chapel Hill
2:15	**BREM-SC: A Bayesian Random Effects Mixture Model for Clustering Single Cell Multi-Omics Data** Xinjun Wang*, Zhe Sun, Yanfu Zhang, Heng Huang, Kong Chen, Ying Ding and Wei Chen, University of Pittsburgh
2:30	**Co-Localization Between Sequence Constraint and Epigenomic Information Improves Interpretation of Whole Genome Sequencing Data** Danqing Xu*, Chen Wang and Krzysztof Kiryluk, Columbia University; Joseph D. Buxbaum, Icahn School of Medicine at Mount Sinai; Iuliana Ionita-Laza, Columbia University
2:45	**Covariate Adaptive False Discovery Rate Control with Applications to Epigenome-Wide Association Studies** Jun Chen*, Mayo Clinic; Xianyang Zhang, Texas A&M University
3:00	**Estimation of Cell-Type Proportions in Complex Tissue** Gregory J. Hunt*, William & Mary; Johann A. Gagnon-Bartsch, University of Michigan
3:15	**Floor Discussion**

50. CONTRIBUTED PAPERS: OBSERVATIONAL AND HISTORICAL DATA ANALYSIS: THE REST IS HISTORY

Sponsor: ENAR
Chair: Paul J. Rathouz, University of Texas, Austin

1:45	**Identifying the Optimal Timing of Surgery from Observational Data** Xiaofei Chen* and Daniel F. Heitjan, Southern Methodist University and University of Texas Southwestern Medical Center; Gerald Greil and Haekyung Jeon-Slaughter, University of Texas Southwestern Medical Center
2:00	**Historical Control Borrowing in Adaptive Designs "To Borrow or Not to Borrow?"** Nusrat Harun*, Cincinnati Children's Hospital Medical Center; Mi-Ok Kim, University of California, San Francisco; Maurizio Macaluso, Cincinnati Children's Hospital Medical Center
2:15	**Weighted F Test and Weighted Chi-Square Test for Multiple Group Comparisons in Observational Studies** Maiying Kong*, Xiaofang Yan and Qi Zheng, University of Louisville
2:30	**Bayesian Probability of Success of Clinical Trials for the Generalized Linear Model Using Historical Data** Ethan M. Alt*, Matthew A. Psioda and Joseph G. Ibrahim, University of North Carolina, Chapel Hill
2:45	**Borrowing Strength from Auxiliary Variables and Historical Data for Counties with very Small Sample Sizes or No Data** Hui Xie*, Deborah B. Rolka and Lawrence Barker, Centers for Disease Control and Prevention
3:00	**Adaptive Combination of Conditional Treatment Effect Estimators Based on Randomized and Observational Data** David Cheng*, VA Boston Healthcare System; Ross Prentice, University of Washington School of Public Health and Community Medicine; Tianxi Cai, Harvard T.H. Chan School of Public Health
3:15	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

51. CONTRIBUTED PAPERS: IMMUNOTHERAPY CLINICAL TRIAL DESIGN AND ANALYSIS

Sponsor: ENAR
Chair: Ming-Hui Chen, University of Connecticut

1:45	**Time-to-Event Model-Assisted Designs to Accelerate and Optimize Early-Phase Immunotherapy Trials** Ruitao Lin*, University of Texas MD Anderson Cancer Center
2:00	**Designing Cancer Immunotherapy Trials with Delayed Treatment Effect Using Maximin Efficiency Robust Statistics** Xue Ding* and Jianrong Wu, University of Kentucky
2:15	**Cancer Immunotherapy Trial Design with Cure Rate and Delayed Treatment Effect** Jing Wei* and Jianrong Wu, University of Kentucky
2:30	**Cancer Immunotherapy Trial Design with Long-Term Survivors** Jianrong Wu* and Xue Ding, University of Kentucky
2:45	**Evaluate the Properties of Cure Model in the Context of Immuno-oncology Trials** Quyen Duong* and Jennifer Le-Rademacher, Mayo Clinic
3:00	**Phase I/II Dose-Finding Interval Design for Immunotherapy** Yeonhee Park*, Medical University of South Carolina
3:15	**Floor Discussion**

52. CONTRIBUTED PAPERS: MACHINE LEARNING AND STATISTICAL RELATIONAL LEARNING

Sponsor: ENAR
Chair: Mohan D. Pant, Eastern Virginia Medical School

1:45	**Merging versus Ensembling in Multi-Study Machine Learning: Theoretical Insight from Random Effects** Zoe Guan* and Giovanni Parmigiani, Harvard T.H. Chan School of Public Health, Dana-Farber Cancer Institute; Prasad Patil, Boston University
2:00	**Informative Dynamic ODE-based-Network Learning (IDOL) from Steady Data** Chixiang Chen*, Ming Wang and Rongling Wu, The Pennsylvania State University
2:15	**Examining the Regulatory Use of Machine Learning for Drug Safety Studies** Jae Joon Song*, Hana Lee and Tae Hyun Jung, U.S. Food and Drug Administration
2:30	**Mixture Proportion Estimation in Positive-Unlabeled Learning** James Patrick Long*, University of Texas MD Anderson Cancer Center; Zhenfeng Lin, Microsoft
2:45	**Unsupervised Learning of Disease Heterogeneity and Patient Subgroups using Diagnosis Codes in Electronic Medical Records** Yaomin Xu*, Vanderbilt University Medical Center
3:00	**Deep Learning for Cell Painting Image Analysis** Yuting Xu*, Andy Liaw and Shubing Wang, Merck & Co., Inc.
3:15	**Model Building Methods in Machine Learning for Clinical Outcome Prediction** Jarcy Zee* and Qian Liu, Arbor Research Collaborative for Health; Laura H. Mariani, University of Michigan; Abigail R. Smith, Arbor Research Collaborative for Health

53. CONTRIBUTED PAPERS: TIME SERIES AND RECURRENT EVENT DATA

Sponsor: ENAR
Chair: S. Yaser Samadi, Southern Illinois University

1:45	**Integer-Valued Autoregressive Process with Flexible Marginal and Innovation Distributions** Matheus Bartolo Guerrero*, King Abdullah University of Science and Technology; Wagner Barreto-Souza, Universidade Federal de Minas Gerais; Hernando Ombao, King Abdullah University of Science and Technology
2:00	**Analysis of N-of-1 Trials Using Bayesian Distributed Lag Model with AR(p) Error** Ziwei Liao*, Ying Kuen Cheung and Ian Kronish, Columbia University; Karina Davidson, Feinstein Institute for Medical Research
2:15	**An Estimating Equation Approach for Recurrent Event Models with Non-Parametric Frailties** Lili Wang*, University of Michigan; Douglas E. Schaubel, University of Pennslyvania
2:30	**Shape-Preserving Prediction for Stationary Functional Time Series** Shuhao Jiao* and Hernando Ombao, King Abdullah University of Science and Technology

SCIENTIFIC PROGRAM
(CONTINUED)

2:45	**A Class of Dynamic Additive-Multiplicative Models for Recurrent Event Data** Russell S. Stocker*, Indiana University of Pennsylvania
3:00	**Causal Dependence between Multivariate Time Series** Yuan Wang*, Washington State University; Louis Scharf, Colorado State University
3:15	**Floor Discussion**

54. CONTRIBUTED PAPERS: MASSIVE DATA: A GIANT PROBLEM?

Sponsor: ENAR
Chair: Sharon M. Lutz, Harvard Medical School and Harvard Pilgrim Health Care Institute

1:45	**Irreproducibility in Large-Scale Drug Sensitivity Data** Zoe L. Rehnberg* and Johann A. Gagnon-Bartsch, University of Michigan
2:00	**A New Integrated Marked Point Process Approach to Analyze Highly Multiplexed Cellular Imaging Data** Coleman R. Harris*, Qi Liu, Eliot McKinley, Joseph Roland, Ken Lau, Robert Coffey and Simon Vandekar, Vanderbilt University Medical Center
2:15	**Comparison of Methods to Analyze Clustered Time-to-Event Data with Competing Risks** Yuxuan Wang*, Guanqun Meng, Wenhan Lu, Zehua Pan, Can Meng, Erich Greene, Peter Peduzzi and Denise Esserman, Yale Center for Analytical Sciences
2:30	**False Discovery Rates for Second-Generation p-Values in Large-Scale Inference** Valerie Welty* and Jeffrey Blume, Vanderbilt University
2:45	**Drives of Inpatient Readmissions: Insights from Analysis of National Inpatient Database** Haileab Hilafu* and Bogdan Bichescu, University of Tennessee
3:00	**Large Scale Hypothesis Testing with Reduced Variance of the False Discovery Proportion** Olivier Thas*, I-BioStat, Data Science Institute, Hasselt University, Belgium, Ghent University, Belgium and University of Wollongong, Australia; Stijn Hawinkel, Ghent University, Belgium; Luc Bijnens, Janssen Pharmaceuticals
3:15	**Floor Discussion**

MONDAY, MARCH 23
3:30 p.m. — 3:45 p.m.

REFRESHMENT BREAK WITH OUR EXHIBITORS

MONDAY, MARCH 23
3:45 p.m. — 5:30 p.m.

55. HUMAN MICROBIOME STUDIES: NOVEL METHODS AND NEW STUDIES

Sponsors: ENAR, IMS, ASA Biometrics Section, ASA Statistics in Genomics and Genetics Section
Organizer: Ni Zhao, Johns Hopkins Bloomberg School of Public Health
Chair: Ni Zhao, Johns Hopkins Bloomberg School of Public Health

3:45	**A Novel Method for Compositional Analysis of the Microbiome Data** Yijuan Hu*, Emory University
4:10	**Estimating the Overall Contribution of Human Oral Microbiome to the Risk of Developing Cancers Based on Prospective Studies** Jianxin Shi*, National Cancer Institute, National Institutes of Health
4:35	**Multi-Group Analysis of Compositions of Microbiomes with Bias Correction (MANCOM-BC)** Shyamal D. Peddada* and Huang Lin, University of Pittsburgh
5:00	**A Powerful Microbial Group Association Test Based on the Higher Criticism Analysis for Sparse Microbial Association Signals** Ni Zhao* and Hyunwook Koh, Johns Hopkins University
5:25	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

56. BAYESIAN APPROACHES FOR COMPLEX INNOVATIVE CLINICAL TRIAL DESIGN

Sponsors: ENAR, ASA Bayesian Statistical Science Section, ASA Biometrics Section, ASA Statistical Consulting Section, ASA Biopharmaceutical Section
Organizer: Joseph Ibrahim, University of North Carolina, Chapel Hill
Chair: Brady Nifong, University of North Carolina, Chapel Hill

3:45	**Bayesian Clinical Trial Design using Historical Data that Inform the Treatment Effect** Joseph G. Ibrahim* and Matthew A. Psioda, University of North Carolina, Chapel Hill
4:10	**Advanced Hierarchical Modeling in Clinical Trials** Kert Viele*, Berry Consultants
4:35	**Bayesian Sequential Monitoring of Clinical Trials** Matthew Austin Psioda* and Evan Kwiatkowski, University of North Carolina, Chapel Hill; Mat Soukup and Eugenio Andraca-Carrera, U.S. Food and Drug Administration
5:00	**Bayesian Clinical Trial Designs using SAS** Fang Chen*, SAS Institute Inc.; Guanghan Frank Liu, Merck & Co. Inc.
5:25	**Floor Discussion**

57. ACHIEVING REAL-WORLD EVIDENCE FROM REAL-WORLD DATA: RECENT DEVELOPMENTS AND CHALLENGES

Sponsors: ENAR, ASA Biometrics Section
Organizer: Haiwen Shi, U.S. Food and Drug Administration
Chair: Haiwen Shi, U.S. Food and Drug Administration

3:45	**Real-World Data and Analytics for Regulatory Decision-Making: FDA/CDRH Experience** Lilly Yue*, U.S. Food and Drug Administration
4:15	**RWD, EHRs, PROs; Using Data to Inform the Patient Trajectory and Experience** Warren A. Kibbe*, Duke University
4:45	**Addressing Confounding in Real-World Evidence using Propensity Scores** John D. Seeger*, Optum
5:15	**Discussant:** Lisa LaVange, University of North Carolina, Chapel Hill

58. NOVEL SPATIAL MODELING APPROACHES FOR AIR POLLUTION EXPOSURE ASSESSMENT

Sponsors: ENAR, ASA Section on Statistics and the Environment
Organizer: Yawen Guan, University of Nebraska, Lincoln
Chair: Kate Calder, The Ohio State University

3:45	**Spatiotemporal Data Fusion Model for Air Pollutants in the Near-Road Environment using Mobile Measurements and Dispersion Model Output** Owais Gilani*, Bucknell University; Veronica J. Berrocal, University of California, Irvine; Stuart A. Batterman, University of Michigan
4:10	**Multi-Resolution Data Fusion of Air Quality Model Outputs for Improved Air Pollution Exposure Assessment: An Application to PM2.5** Veronica J. Berrocal*, University of California, Irvine
4:35	**Multivariate Spectral Downscaling for PM2.5 Species** Yawen Guan*, University of Nebraska, Lincoln; Brian Reich, North Carolina State University; James Mulholland, Georgia Institute of Technology; Howard Chang, Emory University
5:00	**Functional Regression for Predicting Air Pollution Concentrations from Spatially Misaligned Data** Meredith Franklin* and Khang Chau, University of Southern California
5:25	**Floor Discussion**

59. INNOVATIONS IN TWO PHASE SAMPLING DESIGNS WITH APPLICATIONS TO EHR DATA

Sponsors: ENAR, ASA Biometrics Section
Organizer: Pamela Shaw, University of Pennsylvania
Chair: Bryan Shepherd, Vanderbilt University

3:45	**Optimal and Nearly-Optimal Designs for Studies with Measurement Errors** Gustavo G. C. Amorim*, Bryan E. Shepherd, Ran Tao and Sarah C. Lotspeich, Vanderbilt University Medical Center; Pamela A. Shaw, University of Pennsylvania; Thomas Lumley, University of Auckland
4:15	**The Mean Score and Efficient Two-Phase Sampling for Discrete-Time Survival Models with Error Prone Exposures** Kyunghee Han*, University of Pennsylvania; Thomas Lumley, University of Auckland; Bryan E. Shepherd, Vanderbilt University Medical Center; Pamela A. Shaw, University of Pennsylvania

SCIENTIFIC PROGRAM
(CONTINUED)

4:45	**Two-Phase Designs Involving Incomplete Life History Processes** Richard J. Cook*, University of Waterloo
5:15	**Discussant:** Jianwen Cai, University of North Carolina, Chapel Hill

60. RECENT APPROACHES TO MULTIVARIATE DATA ANALYSIS IN THE HEALTH SCIENCES

Sponsors: ENAR, ASA Bayesian Statistical Science Section, ASA Biometrics Section, ASA Section on Statistics in Epidemiology, ASA Statistics in Genomics and Genetics Section, ASA Health Policy Statistics Section
Organizer: Brian Neelon, Medical University of South Carolina
Chair: Christopher Schmid, Brown University

3:45	**A Multivariate Discrete Failure Time Model for the Analysis of Infant Motor Development** Brian Neelon*, Medical University of South Carolina
4:10	**Incorporating a Bivariate Neighborhood Effect of a Single Neighborhood Identifier in a Hierarchical Model** James O'Malley*, Dartmouth College; Peter James, Harvard T.H. Chan School of Public Health; Todd A. MacKenzie and Jinyoung Byun, Dartmouth College; SV Subramanian, Harvard T.H. Chan School of Public Health; Jason B. Block, Harvard Pilgrim Health Care
4:35	**hubViz: A Bayesian Model for Hub-Centric Visualization of Multivariate Binary Data** Dongjun Chung*, The Ohio State University; Jin Hyun Nam, Medical University of South Carolina; Ick Hoon Jin, Yonsei University
5:00	**On Nonparametric Estimation of Causal Networks with Additive Faithfulness** Kuang-Yao Lee*, Temple University; Tianqi Liu, Google; Bing Li, The Pennsylvania State University; Hongyu Zhao, Yale University
5:25	**Floor Discussion**

61. SPEED POSTERS: IMAGING DATA/SURVIVAL ANALYSIS/SPATIO-TEMPORAL

Sponsor: ENAR

Chair: Layla Parast, RAND

61a. INVITED POSTER: A Geometric Approach Towards Evaluating fMRI Preprocessing Pipelines

Martin Lindquist*, Johns Hopkins Bloomberg School of Public Health

61b. Non-Parametric Estimation of Spearman's Rank Correlation with Bivariate Survival Data

WITHDRAWN

Svetlana K. Eden*, Vanderbilt University; Chun Li, Case Western Reserve University; Bryan Shepherd, Vanderbilt University

61c. Nonparametric Tests for Semi-Competing Risks Data under Markov Illness-Death Model

Jing Li* and Giorgos Bakoyannis, Indiana University; Ying Zhang, University of Nebraska Medical Center; Sujuan Gao, Indiana University

61d. Parsimonious Covariate Selection for Interval Censored Data

Yi Cui*, State University of New York at Albany; Xiaoxue Gu, North Dakota State University; Bo Ye, State University of New York at Albany

61e. Identifying Amenity Typologies in the Built Environment: A Bayesian Non-Parametric Approach

Adam T. Peterson*, University of Michigan; Veronica Berrocal, University of California, Irvine; Brisa Sánchez, Drexel University

61f. Estimation of a Buffering Window in Functional Linear Cox Regression Models for Spatially-Defined Environmental Exposure

Jooyoung Lee*, Harvard T.H. Chan School of Public Health; Donna Spiegelman, Yale School of Public Health; Molin Wang, Harvard T.H. Chan School of Public Health

61g. An Alternative Sensitivity Analysis for Informative Censoring

Patrick O'Connor*, Chiu-Hsieh Hsu, Denise Roe and Chengcheng Hu, University of Arizona; Jeremy M.G. Taylor, University of Michigan

61h. Displaying Survival of Patient Groups Defined by Covariate Paths: Extensions of the Kaplan-Meier Estimator

Melissa Jay*, University of Iowa; Rebecca Betensky, New York University

SCIENTIFIC PROGRAM
(CONTINUED)

61i. Semiparametric Transformation Model for Clustered Competing Risks Data

Yizeng He* and Soyoung Kim, Medical College of Wisconsin; Lu Mao, University of Wisconsin, Madison; Kwang Woo Ahn, Medical College of Wisconsin

61j. Partial Linear Single Index Mean Residual Life Models

Peng Jin* and Mengling Liu, New York University School of Medicine

61k. Evaluating the Diagnostic Accuracy of a New Biomarker for Prostate Cancer: Challenges in Small Samples

Joshua I. Banks*, Jungreem Woo, Sandra Santasusagna, Benjamin Leiby and Josep Domingo-Domenech, Thomas Jefferson University

61l. Identifying Spatio-Temporal Variation in Breast Cancer Incidence Among Different Age Cohorts Using Bayesian Hierarchical Modeling

Amy E. Hahn*, Jacob Oleson and Paul Romitti, University of Iowa

61m. One-to-One Feature Matching with Application to Multi-Level Modeling

David Degras*, University of Massachusetts, Boston

62. CONTRIBUTED PAPERS: IMAGING AND STREAMING DATA ANALYSIS

Sponsor: ENAR
Chair: Daniel Kowal, Rice University

3:45	**Generalizable Two-Stage PCA for Confounding Adjustment** Sarah M. Weinstein*, Kristin A. Linn*, and Russell T. Shinohara*, University of Pennsylvania
4:00	**Permutation-Based Inference for Spatially Localized Signals in Longitudinal MRI Data** Jun Young Park* and Mark Fiecas, University of Minnesota
4:15	**Geostatistical Modeling of Positive Definite Matrices: An Application to Diffusion Tensor Imaging** Zhou Lan*, The Pennsylvania State University; Brian Reich, North Carolina State University; Joseph Guinness, Cornell University; Dipankar Bandyopadhyay, Virginia Commonwealth University
4:30	**Length Penalized Probabilistic Principal Curve with Application to Pharmacologic Colon Imaging Study** Huan Chen*, Johns Hopkins Bloomberg School of Public Health
4:45	**Image-on-Scalar Regression Via Interpretable Regularized Reduced Rank Regression** Tianyu Ding*, Dana Tudorascu, Annie Cohen and Robert Krafty, University of Pittsburgh
5:00	**Automatic Transformation and Integration to Improve Visualization and Discovery of Latent Effects in Imaging Data** Johann A. Gagnon-Bartsch*, University of Michigan; Gregory J. Hunt, William & Mary
5:15	**Floor Discussion**

63. CONTRIBUTED PAPERS: CAUSAL INFERENCE AND PROPENSITY SCORE METHODS

Sponsor: ENAR
Chair: Wei Chen, University of Pittsburgh

3:45	**Generalizing Randomized Trial Findings to a Target Population using Complex Survey Population Data** Benjamin Ackerman*, Catherine R. Lesko and Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health
4:00	**Weak-Instrument Robust Tests in Two-Sample Summary-Data Mendelian Randomization** Sheng Wang* and Hyunseung Kang, University of Wisconsin, Madison
4:15	**Propensity Score Matching Methods in Unbalanced Studies with Optimal Caliper Choice** Ziliang Zhu*, University of North Carolina, Chapel Hill; Toshio Kimura, Regeneron Pharmaceuticals, Inc.; Xinyi He, Agios Pharmaceuticals, Inc.; Zhen Chen, Regeneron Pharmaceuticals, Inc.
4:30	**Mendelian Randomization with Statistical Warranty of All Core Assumptions** Zhiguang Huo*, University of Florida
4:45	**Improved Propensity Score for Matching** Ernesto Ulloa*, Marco Carone and Alex Luedtke, University of Washington
5:00	**A Likelihood Ratio Test for Multi-Dimensional Mediation Effects** Wei Hao* and Peter X.K. Song, University of Michigan
5:15	**Floor Discussion**

SCIENTIFIC PROGRAM

(CONTINUED)

64. CONTRIBUTED PAPERS: LONGITUDINAL DATA AND JOINT MODELS OF LONGITUDINAL AND SURVIVAL DATA

Sponsor: ENAR
Chair: Liang Li, University of Texas MD Anderson Cancer Center

3:45	**Estimation of the Joint Distribution of Survival Time and Mark Variable in the Presence of Dependent Censoring** Busola O. Sanusi*, Michael G. Hudgens and Jianwen Cai, University of North Carolina, Chapel Hill
4:00	**A Multilevel Mixed Effects Varying Coefficient Model with Multilevel Predictors and Random Effects for Modeling Hospitalization Risk in Patients on Dialysis** Yihao Li*, University of California, Los Angeles; Danh V. Nguyen, University of California, Irvine; Esra Kurum, University of California, Riverside; Connie M. Rhee, University of California, Irvine; Yanjun Chen, University of California, Irvine Institute of Clinical and Translational Science; Kamyar Kalantar-Zadeh, University of California, Irvine; Damla Senturk, University of California, Los Angeles
4:15	**Structural Joint Modeling of Longitudinal and Survival Data** Bryan Blette*, University of North Carolina, Chapel Hill; Peter Gilbert, Fred Hutchinson Cancer Research Center; Michael Hudgens, University of North Carolina, Chapel Hill
4:30	**Bayesian Models for Joint Longitudinal and Competing Risks Data** Allison KC Furgal*, Ananda Sen and Jeremy M.G. Taylor, University of Michigan
4:45	**Joint Model for Survival and Multivariate Sparse Functional Data with Application to a Study of Alzheimer's Disease** Cai Li*, Yale University; Luo Xiao, North Carolina State University
5:00	**Bayesian Semiparametric Joint Models to Study Growth and Islet Autoimmunity in Subjects at High Risk for Type 1 Diabetes** Xiang Liu*, Roy Tamura, Kendra Vehik and Jeffrey Krischer, University of South Florida
5:15	**Marginal Inference in Transition Models with Generalized Estimating Equations: What is Being Estimated?** Danping Liu*, National Cancer Institute, National Institutes of Health; Joe Bible, Clemson University; Paul S. Albert, National Cancer Institute, National Institutes of Health; Bruce G. Simons-Morton, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

65. CONTRIBUTED PAPERS: PERSONALIZED MEDICINE AND BIOMARKERS

Sponsor: ENAR
Chair: Dong-Yun Kim, National Heart, Lung, and Blood Institute, National Institutes of Health

3:45	**Synergistic Self-Learning of Individualized Dietary Supplement Rules from Multiple Health Benefit Outcomes** Yiwang Zhou* and Peter X.K. Song, University of Michigan
4:00	**Integrative Network Learning for Multi-Modality Biomarker Data** Shanghong Xie*, Columbia University; Donglin Zeng, University of North Carolina, Chapel Hill; Yuanjia Wang, Columbia University
4:15	**An Optimal Design of Experiments Approach to Closed-Loop Target-Controlled Induction of Anesthesia for Robustness to Interpatient PK/PD Variability: A Simulation Study** Ryan T. Jarrett* and Matthew S. Shotwell, Vanderbilt University
4:30	**Utilization of Residual Lifetime Quantiles to Optimize Personalized Biomarker Screening Intervals** Fang-Shu Ou* and Phillip J. Schulte, Mayo Clinic; Martin Heller, Private Practitioner
4:45	**Precision Medicine Using MixedBART for Repeated Measures** Charles K. Spanbauer* and Rodney Sparapani, Medical College of Wisconsin
5:00	**A Statistical Method to Estimate Sleep Duration from Actigraphy Data** Jonggyu Baek*, University of Massachusetts Medical School; Margaret Banker, Erica C. Jansen and Karen E. Peterson, University of Michigan; E. Andrew Pitchford, Iowa State University; Peter X. K. Song, University of Michigan
5:15	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

66. CONTRIBUTED PAPERS: STATISTICAL GENETICS: SINGLE-CELL SEQUENCING DATA

Sponsor: ENAR
Chair: Qing Pan, The George Washington University

3:45	**SMNN: Batch Effect Correction for Single-Cell RNA-seq Data via Supervised Mutual Nearest Neighbor Detection** Gang Li*, Yuchen Yang, Huijun Qian and Kirk C. Wilhelmsen, University of North Carolina, Chapel Hill; Yin Shen, University of California, San Francisco; Yun Li, University of North Carolina, Chapel Hill
4:00	**Multiple Phenotype-Multiple Genotype Testing with Principal Components** Andy Shi*, Harvard University; Ryan Sun, University of Texas MD Anderson Cancer Center; Xihong Lin, Harvard University
4:15	**Single-Cell ATAC-seq Signal Extraction and Enhancement with SCATE** Zhicheng Ji*, Weiqiang Zhou and Hongkai Ji, Johns Hopkins University
4:30	**A Neural Network Based Dropout Correction for Single-Cell RNA-Seq Data with High Sparsity** Lingling An*, Xiang Zhang and Siyang Cao, University of Arizona
4:45	**A Novel Surrogate Variable Analysis Framework in Large-Scale Single-Cell RNA-seq Data Integration** Chao Huang*, Yue Julia Wang and Madison Layfield, Florida State University
5:00	**Robust Normalization of Single-Cell RNA-seq Data using Local Smoothing and Median Ratio** Chih-Yuan Hsu*, Qi Liu and Yu Shyr, Vanderbilt University Medical Center
5:15	**Subpopulation Identification for Single-Cell RNA-Sequencing Data Using Functional Data Analysis** Kyungmin Ahn* and Hironobu Fujiwara, RIKEN Center for Biosystems Dynamics Research, Japan

67. CONTRIBUTED PAPERS: SEMIPARAMETRIC AND NONPARAMETRIC METHODS AND APPLICATIONS

Sponsor: ENAR
Chair: Aaron W. Scheffler, University of California, San Francisco

3:45	**A Semiparametric Alternative Method to Conditional Logistic Regression for Combining Biomarkers under Matched Case-Control Studies** Wen Li* and Ruosha Li, University of Texas Health Science Center at Houston; Ziding Feng, Fred Hutchinson Cancer Research Center; Jing Ning, University of Texas MD Anderson Cancer Center
4:00	**Exponential and Super-Exponential Convergence of Misclassification Probabilities in Nonparametric Modeling** Richard Charnigo* and Cidambi Srinivasan, University of Kentucky
4:15	**Zero-Inflated Quantile Rank-Score Based Test (ZIQRank) with Application to scRNA-seq Differential Gene Expression Analysis** Wodan Ling*, Fred Hutchinson Cancer Research Center; Ying Wei, Columbia University; Wenfei Zhang, Sanofi
4:30	**A Nonparametric MC-SIMEX Method** Lili Yu*, Congjian Liu, Jingjing Yin and Jun Liu, Georgia Southern University
4:45	**Nonparametric Regression for Error-Prone Homogeneous Pooled Data** Dewei Wang*, University of South Carolina
5:00	**Testing for Uniform Stochastic Ordering among k Populations** Chuan-Fa Tang*, University of Texas, Dallas; Dewei Wang, University of South Carolina
5:15	**k-Tuple Partially Rank-Ordered Set Sampling** Kaushik Ghosh* and Marvin C. Javier, University of Nevada, Las Vegas

SCIENTIFIC PROGRAM
(CONTINUED)

TUESDAY, MARCH 24
8:30 a.m. — 10:15 a.m.

68. CHALLENGES AND OPPORTUNITIES IN METHODS FOR PRECISION MEDICINE

Sponsor: IMS
Organizer Yingqi Zhao, Fred Hutchinson Cancer Research Center
Chair: Xinyuan Dong, University of Washington

8:30	**Subgroup-Effects Models (SGEM) for Analysis of Personal Treatment Effects**	Peter X.K. Song*, Ling Zhou and Shiquan Sun, University of Michigan; Haoda Fu, Eli Lilly and Company
8:55	**Kernel Optimal Orthogonality Weighting for Estimating Effects of Continuous Treatments**	Michele Santacatterina*, Cornell University
9:20	**Inference on Individualized Treatment Rules from Observational Studies with High-Dimensional Covariates**	Yingqi Zhao* and Muxuan Liang, Fred Hutchinson Cancer Research Center; Young-Geun Choi, Sookmyung University; Yang Ning, Cornell University; Maureen Smith, University of Wisconsin, Madison
9:45	**Integrative Analysis of Electronic Health Records for Precision Medicine**	Yuanjia Wang*, Columbia University; Jitong Luo and Donglin Zeng, University of North Carolina, Chapel Hill
10:10	**Floor Discussion**	

69. RECENT DEVELOPMENTS IN RISK ESTIMATION AND BIOMARKER MODELING WITH A FOCUS IN ALZHEIMER'S DISEASE

Sponsors: ENAR, ASA Biometrics Section, ASA Health Policy Statistics Section, ASA Mental Health Statistics Section
Organizer: Zheyu Wang, Johns Hopkins University
Chair: Danping Liu, National Cancer Institute, National Institutes of Health

8:30	**Analyzing Semi-Competing Risks Data as a Longitudinal Bivariate Process**	Sebastien Haneuse*, Harvard T.H. Chan School of Public Health; Daniel Nevo, University of Tel Aviv
8:55	**Biomarker Models for Early Alzheimer's Disease Risk Prediction Before Symptoms Appear**	Zheyu Wang*, Johns Hopkins University
9:20	**A Statistical Test on the Ordering of Changes in Biomarkers for Preclinical Alzheimer's Disease**	Chengjie Xiong*, Washington University in St. Louis
9:45	**Changepoint Estimation for Biomarkers of Alzheimer's Disease**	Laurent Younes*, Johns Hopkins University
10:10	**Floor Discussion**	

70. CLINICAL TRIAL DESIGNS IN A NEW ERA OF IMMUNOTHERAPY: CHALLENGES AND OPPORTUNITIES

Sponsors: ENAR, ASA Biometrics Section, ASA Biopharmaceutical Section
Organizer: Yeonhee Park, Medical University of South Carolina
Chair: Yeonhee Park, Medical University of South Carolina

8:30	**Immune-Oncology Agents: Endpoints and Designs**	Hao Wang* and Gary Rosner, Johns Hopkins University School of Medicine
8:55	**Adaptive Dose Finding Based on Safety and Feasibility in Early-Phase Clinical Trials of Adoptive Cell Immunotherapy**	Nolan A. Wages* and Camilo E. Fadul, University of Virginia
9:20	**Novel Bayesian Phase I/II Designs for Identifying Safe and Efficacious Treatments for Immunotherapy**	J. Jack Lee*, University of Texas MD Anderson Cancer Center
9:45	**Impact of Design Misspecification in Immuno-Oncology Trials**	Jennifer Le-Rademacher*, Quyen Duong, Tyler Zemla and Sumithra J. Mandrekar, Mayo Clinic
10:10	**Floor Discussion**	

SCIENTIFIC PROGRAM
(CONTINUED)

71. THE THREE M'S: MEETINGS, MEMBERSHIPS, AND MONEY!

Sponsors: CENS, ENAR
Organizers: Jing Li, Richard M. Fairbanks School of Public Health, Indiana University and Hannah Weeks, Vanderbilt University
Chair: Will A. Eagan, Purdue University

8:30	**Panel Discussion:** Jeff Goldsmith, Columbia University Donna LaLonde, American Statistical Association Nandita Mitra, University of Pennsylvania Perelman School of Medicine Sarah Ratcliffe, University of Virginia
10:00	**Floor Discussion**

72. RECENT ADVANCES IN JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA

Sponsors: ENAR, ASA Biometrics Section, ASA Section on Statistics in Defense and National Security, ASA Section on Statistics and the Environment, ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section, ASA Mental Health Statistics Section
Organizer: Abdus Sattar, Case Western Reserve University
Chair: Jeffrey Albert, Case Western Reserve University

8:30	**Assessing Importance of Biomarkers: A Bayesian Joint Modeling Approach of Longitudinal and Survival Data with Semicompeting Risks** Ming-Hui Chen* and Fan Zhang, University of Connecticut; Xiuyu Julie Cong, Boehringer Ingelheim (China) Investment Co., Ltd.; Qingxia Chen, Vanderbilt University
8:55	**Inference with Joint Models Under Misspecified Random Effects Distributions** Sanjoy Sinha*, Carleton University; Abdus Sattar, Case Western Reserve University
9:20	**Personalized Decision Making for Biopsies in Prostate Cancer Active Surveillance Programs** Dimitris Rizopoulos*, Erasmus University Medical Center
9:45	**Quantifying Direct and Indirect Effect for Longitudinal Mediator and Survival Outcome Using Joint Modeling Approach** Cheng Zheng*, University of Wisconsin, Milwaukee; Lei Liu, Washington University in St. Louis
10:10	**Floor Discussion**

73. RECENT ADVANCES IN NETWORK META-ANALYSIS WITH FLEXIBLE BAYESIAN APPROACHES

Sponsors: ENAR, ASA Bayesian Statistical Science Section, ASA Biometrics Section, ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section
Organizer: Hwanhee Hong, Duke University School of Medicine
Chair: Roland Matsouaka, Duke University School of Medicine

8:30	**Data-Adaptive Synthesis of Historical Information through Network-Meta-Analytic-Predictive Priors** Jing Zhang*, University of Maryland; Hwanhee Hong, Duke University School of Medicine; Yong Chen, University of Pennsylvania; Cher Dallal, University of Maryland
9:00	**Bayesian Flexible Hierarchical Skew Heavy-Tailed Multivariate Meta Regression Models for Individual Patient Data with Applications** Sung Duk Kim*, National Cancer Institute, National Institutes of Health; Ming-Hui Chen, University of Connecticut; Joseph G. Ibrahim, University of North Carolina, Chapel Hill; Arvind K. Shah and Jianxin Lin, Merck Research Laboratories
9:30	**Bayesian Network Meta-Analysis for Estimating Population Treatment Effects** Hwanhee Hong*, Duke University School of Medicine
10:00	**Discussant:** Christopher Schmid, Brown University School of Public Health

74. CONTRIBUTED PAPERS: ELECTRONIC HEALTH RECORDS DATA ANALYSIS

Sponsor: ENAR
Chair: Adam Ciarleglio, George Washington University

8:30	**Estimating Individualized Treatment Rules for Multicategory Type 2 Diabetes Treatments Using Electronic Health Records** Jitong Lou*, University of North Carolina, Chapel Hill; Yuanjia Wang, Columbia University; Lang Li, The Ohio State University; Donglin Zeng, University of North Carolina, Chapel Hill
8:45	**Modeling Heterogeneity and Missing Data in Electronic Health Records** Rebecca Anthopolos*, New York University; Qixuan Chen and Ying Wei, Columbia University Mailman School of Public Health
9:00	**Modeling Valid Drug Dosage in the Presence of Conflicting Information Extracted from Electronic Health Records** Michael L. Williams*, Hannah L. Weeks, Cole Beck, Elizabeth McNeer and Leena Choi, Vanderbilt University Medical Center

SCIENTIFIC PROGRAM
(CONTINUED)

9:15	**Case Contamination in Electronic Health Records-Based Case-Control Studies** Jill Schnall*, Lu Wang, Scott Damrauer, Michael Levin and Jinbo Chen, University of Pennsylvania
9:30	**Quantile Rank Test for Dynamic Heterogeneous Genetic Effect in Longitudinal Electronic Health Record Analysis** Tianying Wang*, Ying Wei, Iuliana Ionita-Laza, Zixu Wang and Chunhua Weng, Columbia University
9:45	**Leveraging Electronic Health Data for Embedded Pragmatic Clinical Trials within Health Care Systems: Lessons Learned from the NIH Collaboratory** Andrea J. Cook*, Kaiser Permanente Washington Health Research Institute
10:00	**Floor Discussion**

75. CONTRIBUTED PAPERS: REBEL WITHOUT A CAUSE: SESSIONS ON CAUSAL INFERENCE

Sponsor: ENAR
Chair: Kesheng Wang, West Virginia University

8:30	**A New Method for Estimating a Principal Stratum Causal Effect Conditioning on a Post-Treatment Intermediate Response** Xiaoqing Tan*, University of Pittsburgh; Judah Abberbock, GlaxoSmithKline; Priya Rastogi and Gong Tang, University of Pittsburgh
8:45	**Detecting Heterogeneous Treatment Effect with Instrumental Variables** Michael W. Johnson*, University of Wisconsin, Madison; Jiongyi Cao, The University of Chicago; Hyunseung Kang, University of Wisconsin, Madison
9:00	**A Groupwise Approach for Inferring Heterogeneous Treatment Effects in Causal Inference** Chan Park* and Hyunseung Kang, University of Wisconsin, Madison
9:15	**Estimating Complier Quantile Causal Treatment Effects with Randomly Censored Data and A Binary Instrumental Variable** Bo Wei* and Limin Peng, Emory University; Mei-jie Zhang, Medical College of Wisconsin; Jason Fine, University of North Carolina, Chapel Hill
9:30	**Causal Effects in Twin Studies: The Role of Interference** Bonnie Smith* and Elizabeth Ogburn, Johns Hopkins Bloomberg School of Public Health; Saonli Basu and Matthew McGue, University of Minnesota; Daniel Scharfstein, Johns Hopkins Bloomberg School of Public Health
9:45	**Causal Inference from Self-Controlled Case Series Studies Using Targeted Maximum Likelihood Estimation** Yaru Shi*, Fang Liu and Jie Chen, Merck & Co., Inc.
10:00	**Caution Against Examining the Role of Reverse Causality in Mendelian Randomization** Sharon M. Lutz* and Ann C. Wu, Harvard Medical School and Harvard Pilgrim Health Care Institute; Christoph Lange, Harvard T.H. Chan School of Public Health

76. CONTRIBUTED PAPERS: HYPOTHESIS TESTING: KNOWLEDGE IS POWER

Sponsor: ENAR
Chair: Daniel J. Schaid, Mayo Clinic

8:30	**A Score Based Test for Functional Linear Concurrent Regression** Rahul Ghosal* and Arnab Maity, North Carolina State University
8:45	**Differential Expression Analysis in Single-Cell RNA Sequencing with G-modeling-based Two-Sample Test** Jingyi Zhai* and Hui Jiang, University of Michigan
9:00	**Detect with BERET** Duyeol Lee*, Kai Zhang and Michael R. Kosorok, University of North Carolina, Chapel Hill
9:15	**Resampling-Based Stepwise Multiple Testing Procedures with Applications to Clinical Trial Data** Jiwei He* and Feng Li, U.S. Food and Drug Administration; Yan Gao, The University of Illinois at Chicago; Mark Rothmann, U.S. Food and Drug Administration

SCIENTIFIC PROGRAM
(CONTINUED)

9:30	**Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models** Rong Ma*, T. Tony Cai and Hongzhe Li, University of Pennsylvania
9:45	**Hypothesis Testing to Determine if Two Penalties Are Better Than One: Should Second Order Terms have the Same Penalty as Main Effects?** Todd A. MacKenzie*, Iben Ricket, Jiang Gui and Kimon Bekelis, Dartmouth College
10:00	**Floor Discussion**

77. CONTRIBUTED PAPERS: MISSING (DATA) IN ACTION

Sponsor: ENAR
Chair: Jason Roy, Rutgers University

8:30	**Identifying Treatment Effects using Trimmed Means when Data are Missing Not at Random** Alex J. Ocampo*, Harvard University
8:45	**A Bayesian Multivariate Skew-Normal Mixture Model for Longitudinal Data with Intermittent Missing Observations: An Application to Infant Motor Development** Carter Allen* and Brian Neelon, Medical University of South Carolina; Sara E. Benjamin-Neelon, Johns Hopkins Bloomberg School of Public Health
9:00	**Estimation, Variable Selection and Statistical Inference in a Linear Regression Model under an Arbitrary Missingness Mechanism** Chi Chen* and Jiwei Zhao, State University of New York at Buffalo
9:15	**Influence Function Based Inference in Randomized Trials with Non Monotone Missing Binary Outcomes** Lamar Hunt* and Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health
9:30	**Multiple Imputation Variance Estimation in Studies with Missing or Misclassified Inclusion Criteria** Mark J. Giganti*, Center for Biostatistics in AIDS Research; Bryan E. Shepherd, Vanderbilt University
9:45	**Missing Data in Deep Learning** David K. Lim*, Naim U. Rashid and Joseph G. Ibrahim, University of North Carolina, Chapel Hill
10:00	**An Approximated Expectation-Maximization Algorithm for Analysis of Data with Missing Values** Gong Tang*, University of Pittsburgh

78. CONTRIBUTED PAPERS: BACK TO THE FUTURE: PREDICTION AND PROGNOSTIC MODELING

Sponsor: ENAR
Chair: Dongjun Chung, Medical University of South Carolina

8:30	**High Dimensional Classified Mixed Model Prediction** Mengying Li* and J. Sunil Rao, University of Miami
8:45	**Connecting Population-Level AUC and Latent Scale-Invariant R-square via Semiparametric Gaussian Copula and Rank Correlations** Debangan Dey* and Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health
9:00	**Artificial Intelligence and Agent-Based Modeling - Prediction and Simulation Issue** Nicolas J. Savy* and Philippe Saint-Pierre, Toulouse Institute of Mathematics
9:15	**Improving Survival Prediction Using a Novel Feature Selection and Feature Reduction Framework Based on the Integration of Clinical and Molecular Data** Lisa Neums*, Richard Meier, Devin C. Koestler and Jeffrey A. Thompson, University of Kansas Medical Center and University of Kansas Cancer Center
9:30	**Quantile Regression for Prediction of High-Cost Patients** Scott S. Coggeshall*, VA Puget Sound
9:45	**Joint Prediction of Variable Importance Rank from Binary and Survival Data via Adaptively Weighted Random Forest** Jihwan Oh* and John Kang, Merck & Co., Inc.
10:00	**External Validation Study of SMART Vascular Event Prediction Model Using UK Primary Care Data Between 2000-2017** Laura H. Gunn*, University of North Carolina, Charlotte & Imperial College London; Ailsa McKay, Azeem Majeed and Kosh Ray, Imperial College London

SCIENTIFIC PROGRAM

(CONTINUED)

79. CONTRIBUTED PAPERS: M&M: MEASUREMENT ERROR AND MODELING

Sponsor: ENAR
Chair: Mark Meyer, Georgetown University

8:30	**Statistical Analysis of Data Reproducibility Measures** Zeyi Wang* and Eric Bridgeford, Johns Hopkins Bloomberg School of Public Health; Joshua T. Vogelstein, Johns Hopkins University; Brian Caffo, Johns Hopkins Bloomberg School of Public Health
8:45	**An Approximate Quasi-Likelihood Approach to Analyzing Error-Prone Failure Time Outcomes and Exposures** Lillian A. Boe* and Pamela A. Shaw, University of Pennsylvania
9:00	**Improving the Efficiency of Generalized Raking Estimators to Address Correlated Covariate and Failure-Time Outcome Error** Eric J. Oh*, University of Pennsylvania; Thomas Lumley, University of Auckland; Bryan E. Shepherd, Vanderbilt University; Pamela A. Shaw, University of Pennsylvania
9:15	**Surrogate-Assisted Subsampling in Logistic Regression with Outcome Misclassification** Chongliang Luo*, Arielle Marks-Anglin and Yong Chen, University of Pennsylvania
9:30	**Impact of Design Considerations in Sensitivity to Time Recording Errors in Pharmacokinetic Modeling** Hannah L. Weeks* and Matthew S. Shotwell, Vanderbilt University
9:45	**Floor Discussion**

TUESDAY, MARCH 24

10:15 a.m. — 10:30 a.m.

REFRESHMENT BREAK WITH OUR EXHIBITORS

TUESDAY, MARCH 24

10:30 a.m. — 12:15 p.m.

80. PRESIDENTIAL INVITED ADDRESS

Sponsor: ENAR
Organizer/Chair: Michael J. Daniels, University of Florida

10:30	**Introduction**
10:35	**Distinguished Student Paper Awards**
10:45	**Medical Product, Healthcare Delivery, and Road Safety Policies: Seemingly Unrelated Regulatory Questions** **Sharon-Lise Normand,** Ph.D., S. James Adelstein Professor of Health Care Policy (Biostatistics), Department of Health Care Policy, Harvard Medical School, Department of Biostatistics, Harvard T.H. Chan School of Public Health

SCIENTIFIC PROGRAM
(CONTINUED)

TUESDAY, MARCH 24
1:45 p.m. — 3:30 p.m.

81. STATISTICAL ANALYSIS OF BIOLOGICAL SHAPES

Sponsors: ENAR, ASA Biometrics Section, ASA Section on Statistics in Imaging
Organizer: Anuj Srivastava, Florida State University
Chair: Anuj Srivastava, Florida State University

1:45	**Manifold-Valued Data Analysis of Brain Networks** Ian L. Dryden*, Simon P. Preston and Katie E. Severn, University of Nottingham
2:10	**Shape Analysis for Mitochondria Data** Todd Ogden*, Columbia University; Ruiyi Zhang, Florida State University; Martin Picard, Columbia University; Anuj Srivastava, Florida State University
2:35	**Geometric Methods for Image-Based Statistical Analysis of Shape and Texture of Glioblastoma Multiforme Tumors** Sebastian Kurtek*, The Ohio State University; Karthik Bharath, University of Nottingham; Veera Baladandayuthapani and Arvind Rao, University of Michigan
3:00	**Fiber Bundles in Probabilistic Models** Lorin Crawford*, Brown University; Bruce Wang, Princeton University; Timothy Sudijono, Brown University; Henry Kirveslahti, Duke University; Tingran Gao, The University of Chicago; Doug M. Boyer and Sayan Mukherjee, Duke University
3:25	**Floor Discussion**

82. IMPROVING THE DEVELOPMENT AND VALIDATION OF SCREENING TESTS FOR RARE DISEASES

Sponsor: ENAR, ASA Section on Medical Devices and Diagnostics
Organizer: Gene Pennello, U.S. Food and Drug Administration
Chair: Norberto Pantoja-Galicia, U.S. Food and Drug Administration

1:45	**From Prediction to Policy: Risk Stratification to Improve the Efficiency of Early Detection for Cancer** Ruth Etzioni*, Fred Hutchinson Cancer Research Center
2:10	**A Simple Framework to Identify Optimal Cost-Effective Risk Thresholds for a Single Screen: Comparison to Decision Curve Analysis** Hormuzd Katki*, National Cancer Institute, National Institutes of Health; Ionut Bebu, The George Washington University
2:35	**Sample Weighted Semiparametric Estimation of Cause-Specific Cumulative Risk and Incidence Using Left or Interval-censored Data from Electronic Health Records** Noorie Hyun*, Medical College of Wisconsin; Hormuzd A. Katki and Barry I. Graubard, National Cancer Institute, National Institutes of Health
3:00	**A Statistical Review: Why Average Weighted Accuracy, not Accuracy or AUC?** Qing Pan*, Yunyun Jiang and Scott Evans, The George Washington University
3:25	**Floor Discussion**

83. CAUSAL INFERENCE AND HARMFUL EXPOSURES

Sponsors: ENAR, ASA Biometrics Section, ASA Section on Statistics and the Environment, ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section
Organizer: Maria Cuellar, University of Pennsylvania
Chair: Daniel Malinsky, Johns Hopkins University

1:45	**Envisioning Hypothetical Interventions on Occupational Exposures to Protect Worker Health: Applications of the Parametric G-formula** Andreas M. Neophytou*, Colorado State University
2:10	**A Causal Inference Framework for Cancer Cluster Investigations Using Publicly Available Data** Rachel C. Nethery* and Yue Yang, Harvard T.H. Chan School of Public Health; Anna J. Brown, The University of Chicago; Francesca Dominici, Harvard T.H. Chan School of Public Health
2:35	**Estimating the Effects of Precinct Level Policing Policies Through Causal Inference with Interference** Joseph Antonelli* and Brenden Beck, University of Florida
3:00	**Exploring Evidence of Residual Confounding in Tropical Cyclone Epidemiology Using a Negative Exposure Control Analysis** Brooke Anderson*, Colorado State University; Meilin Yan, Peking University
3:25	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

84. STATISTICAL METHODS FOR EMERGING DATA IN ENVIRONMENTAL HEALTH RESEARCH

Sponsors: ENAR, ASA Section on Statistics and the Environment, ASA Section on Statistics in Epidemiology
Organizer: Jenna Krall, George Mason University
Chair: Jenna Krall, George Mason University

1:45	**Bayesian Joint Modeling of Chemical Structure and Dose Response Curves** Kelly R. Moran*, David Dunson and Amy H. Herring, Duke University
2:10	**Source-Specific Exposure Assessment by using Bayesian Spatial Multivariate Receptor Modeling** Eun Sug Park*, Texas A&M Transportation Institute
2:35	**The Impact of Complex Social and Environmental Mixtures on Educational Outcomes in Young Children** Kathy B. Ensor*, Rice University; Mercedes Bravo, Research Triangle Institute and Rice University; Daniel Kowal, Henry Leong and Marie Lynn Miranda, Rice University
3:00	**Accounting for Mixtures in Risk Assessment** Chris Gennings*, Icahn School of Medicine at Mount Sinai
3:25	**Floor Discussion**

85. BAYESIAN ANALYSIS IN FUNCTIONAL BRAIN IMAGING

Sponsors: ENAR, IMS, ASA Bayesian Statistical Science Section, ASA Biometrics Section, ASA Section on Statistics in Imaging, ASA Mental Health Statistics Section
Organizer: Donatello Telesca, University of California, Los Angeles
Chair: Donatello Telesca, University of California, Los Angeles

1:45	**Functional Regression Methods for Functional Neuroimaging** Jeffrey Scott Morris*, University of Pennsylvania; Hongxiao Zhu, Virginia Tech University; Michelle Miranda, University of Victoria; Neel Desai, Rice University; Veera Baladandayuthapani, University of Michigan; Philip Rausch, Humboldt University
2:10	**A Grouped Beta Process Model for Multivariate Resting-State EEG Microstate Analysis on Twins** Mark Fiecas*, University of Minnesota; Brian Hart, UnitedHealthGroup; Stephen Malone, University of Minnesota
2:35	**Bayesian Analysis of Multidimensional Functional Data** John Shamshoian*, Donatello Telesca and Damla Senturk, University of California, Los Angeles
3:00	**Encompassing Semiparametric Bayesian Inference for Stationary Points in Gaussian Process Regression Models with Applications to Event-Related Potential Analysis** Meng Li*, Cheng-Han Yu and Marina Vannucci, Rice University
3:25	**Floor Discussion**

86. HUMAN DATA INTERACTION: GAINING AN UNDERSTANDING OF THE DATA SCIENCE PIPELINE

Sponsors: ENAR, ASA Section on Statistical Learning and Data Science
Organizer: Jeff Leek, Johns Hopkins Bloomberg School of Public Health
Chair: Jeff Leek, Johns Hopkins Bloomberg School of Public Health

1:45	**Tools for Analyzing R Code the Tidy Way** Lucy D'Agostino McGowan*, Wake Forest University
2:10	**Domain Specific Languages for Data Science** Hadley Wickham*, RStudio
2:35	**The Challenges of Analytic Workflows: Perspectives from Data Science Educators** Sean Kross*, University of California, San Diego
3:00	**Building A Software Package in Tandem with Machine Learning Methods Research Can Result in Both More Rigorous Code and More Rigorous Research** Nick Stayer*, Vanderbilt University
3:25	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

87. CONTRIBUTED PAPERS: SPATIAL AND SPATIAL-TEMPORAL DATA ANALYSIS

Sponsor: ENAR
Chair: Fridtjof Thomas, University of Tennessee Health Science Center

1:45	**Bayesian Spatial-Temporal Accelerated Failure Time Models for Survival Data from Cancer Registries** Ming Wang*, The Pennsylvania State University; Zheng Li, Novartis; Lijun Zhang, The Pennsylvania State University; Yimei Li, University of Pennsylvania; Vern M. Chinchilli, The Pennsylvania State University
2:00	**Where did All the Good Fish Go? Spatio- Temporal Modelling of Research Vessel Data with R** Ethan Lawler* and Joanna Mills Flemming, Dalhousie University
2:15	**Assessing Meteorological Drivers of Air Pollution in the Eastern United States via a Bayesian Quantile Regression Model with Spatially Varying Coefficients** Stella Coker Watson Self*, University of South Carolina; Christopher S. McMahan, Brook Russell and Derek Andrew Brown, Clemson University
2:30	**Spatio-Temporal Mixed Effects Single Index Models** Hamdy F. F. Mahmoud*, Virginia Tech and Assiut University, Egypt; Inyoung Kim, Virginia Tech
2:45	**Bayesian Spatial Blind Source Separation via Thresholded Gaussian Processes** Ben Wu*, University of Michigan; Ying Guo, Emory University; Jian Kang, University of Michigan
3:00	**Incorporating Spatial Structure into Bayesian Spike-and-Slab Lasso GLMs** Justin M. Leach*, Inmaculada Aban and Nengjun Yi, University of Alabama at Birmingham
3:15	**Floor Discussion**

88. CONTRIBUTED PAPERS: EARLY PHASE CLINICAL TRIALS AND BIOMARKERS

Sponsor: ENAR
Chair: Lingling An, University of Arizona

1:45	**Building an Allostatic Load Scale using Item Response Theory** Shelley H. Liu*, Kristen Dams-O'Connor and Julie Spicer, Icahn School of Medicine at Mount Sinai
2:00	**Subgroup-Specific Dose Finding in Phase I Clinical Trials Based on Time to Toxicity Allowing Adaptive Subgroup Combination** Andrew G. Chapple*, Louisiana State University; Peter F. Thall, University of Texas MD Anderson Cancer Center
2:15	**Evaluation of Continuous Monitoring Approach in Early Phase Oncology Trial** Suhyun Kang* and Jingyi Liu, Eli Lilly and Company
2:30	**PA-CRM: A Continuous Reassessment Method for Pediatric Phase I Trials with Concurrent Adult Trials** Yimei Li*, University of Pennsylvania; Ying Yuan, University of Texas MD Anderson Cancer Center
2:45	**Two-Stage Enrichment Clinical Trial Design with Adjustment for Misclassification in Predictive Biomarkers** Yong Lin*, Weichung Joe Shih and Shou-En Lu, Rutgers University
3:00	**Incorporating Real-World Evidence or Historical Data to Improve Phase I Clinical Trial Designs** Yanhong Zhou*, Ying Yuan and J. Jack Lee, University of Texas MD Anderson Cancer Center
3:15	**Density Estimation Based on Pooled Biomarkers using Dirichlet Process Mixtures** Zichen Ma*, University of South Carolina

SCIENTIFIC PROGRAM
(CONTINUED)

89. CONTRIBUTED PAPERS: ELECTRONIC HEALTH RECORDS DATA ANALYSIS AND META-ANALYSIS

Sponsor: ENAR
Chair: Jung-Ying Tzeng, North Carolina State University

1:45 **The Impact of Covariance Priors on Arm-Based Bayesian Network Meta-Analyses with Binary Outcomes**
Zhenxun Wang*, University of Minnesota; Lifeng Lin, Florida State University; James S. Hodges and Haitao Chu, University of Minnesota

2:00 **A Bayesian Multivariate Meta-Analysis of Prevalence Data**
Lianne Siegel* and Kyle Rudser, University of Minnesota; Siobhan Sutcliffe, Washington University School of Medicine; Alayne Markland, University of Alabama at the Birmingham VA Medical Center; Linda Brubaker and Sheila Gahagan, University of California, San Diego; Ann E. Stapleton, University of Washington; Haitao Chu, University of Minnesota

2:15 **An Augmented Estimation Procedure for EHR-based Association Studies Accounting for Differential Misclassification**
Jiayi Tong* and Jing Huang, University of Pennsylvania; Jessica Chubak, Kaiser Permanente Washington Health Research Institute; Xuan Wang, Zhejiang University; Jason H. Moore, Rebecca Hubbard and Yong Chen, University of Pennsylvania

2:30 **Testing Calibration of Risk Prediction Models Using Positive-Only EHR Data**
Lingjiao Zhang*, University of Pennsylvania; Yanyuan Ma, The Pennsylvania State University; Daniel Herman and Jinbo Chen, University of Pennsylvania

2:45 **Bias Reduction Methods for Propensity Scores Estimated from Mismeasured EHR-Derived Covariates**
Joanna Grace Harton*, Rebecca A. Hubbard and Nandita Mitra, University of Pennsylvania

3:00 **Bayesian Network Meta-Regression for Partially Collapsed Ordinal Outcomes: Latent Counts Approach**
Yeongjin Gwon*, University of Nebraska Medical Center; Ming-Hui Chen, University of Connecticut; Mo May, Xun Jiang and Amy Xia, Amgen Inc.; Joseph Ibrahim, University of North Carolina, Chapel Hill

3:15 **Efficient and Robust Methods for Causally Interpretable Meta-Analysis: Transporting Inferences From Multiple Randomized Trials to a Target Population**
Issa J. Dahabreh*, Jon A. Steingrimsson and Sarah E. Robertson, Brown University; Lucia C. Petito, Northwestern University; Miguel A. Hernán, Harvard University

90. CONTRIBUTED PAPERS: SMALL THINGS THAT MAKE A BIG DIFFERENCE: MICROBIOME ANALYSIS

Sponsor: ENAR
Chair: Olivier Thas, Ghent University, Hasselt University and University of Wollongong

1:45 **Robust Inter-Taxa Dependency Estimation for High-Dimensional Microbiome Data**
Arun A. Srinivasan*, The Pennsylvania State University; Danning Li, Jilin University; Lingzhou Xue and Xiang Zhan, The Pennsylvania State University

2:00 **Analysis of Compositions of Microbiomes with Bias Correction**
Huang Lin* and Shyamal Das Peddada, University of Pittsburgh

2:15 **Zero-Inflated Poisson Factor Model with Application to Microbiome Absolute Abundance Data**
Tianchen Xu*, Columbia University; Ryan T. Demmer, University of Minnesota; Gen Li, Columbia University

SCIENTIFIC PROGRAM
(CONTINUED)

2:30	**Zero-Inflated Topic Models for Human Microbiome Data** Rebecca A. Deek* and Hongzhe Li, University of Pennsylvania
2:45	**Bayesian Modeling of Microbiome Count Data for Network Analysis** Qiwei Li*, University of Texas, Dallas; Shuang Jiang, Southern Methodist University; Xiaowei Zhan, University of Texas Southwestern Medical Center
3:00	**Sparse Kernel RV for Identifying Genomic Features Related to Microbiome Community Composition** Nanxun Ma*, Universitya of Washington; Anna Plantinga, Williams College; Michael C. Wu, Fred Hutchinson Cancer Research Center
3:15	**A Bayesian Semiparametric Approach to Wild-Type Distribution Estimation: Accounting for Contamination and Measurement Error (BayesACME)** Will A. Eagan* and Bruce A. Craig, Purdue University

91. CONTRIBUTED PAPERS: STATISTICAL GENETICS: SEQUENCING DATA ANALYSIS

Sponsor: ENAR
Chair: Xiang Liu, University of South Florida

1:45	**lncDIFF: A Novel Quasi-Likelihood Method for Differential Expression Analysis of Non-Coding RNA** Qian Li*, University of South Florida; Xiaoqing Yu, Ritu Chaudhary, Robbert J. Slebos, Christine Chung and Xuefeng Wang, Moffitt Cancer Center
2:00	**ASEP: Gene-based Detection of Allele-Specific Expression in a Population by RNA-seq** Jiaxin Fan* and Jian Hu, University of Pennsylvania Perelman School of Medicine; Chenyi Xue, Hanrui Zhang and Muredach P. Reilly, Columbia University; Rui Xiao and Mingyao Li, University of Pennsylvania Perelman School of Medicine
2:15	**A Sparse Negative Binomial Classifier with Covariate Adjustment for RNA-seq Data** Md Tanbin Rahman*, University of Texas MD Anderson Cancer Center; Hsin-En Huang, An-Shun Tai and Wen-Ping Hsieh, National Tsing Hua University; George Tseng, University of Pittsburgh
2:30	**A Functional Regression Based Approach for Gene-Based Association Testing of Quantitative Trait in Family Studies** Chi-Yang Chiu*, University of Tennessee Health Science Center
2:45	**A Systematic Evaluation of Single-Cell RNA-seq Imputation Methods** Wenpin Hou*, Zhicheng Ji, Hongkai Ji and Stephanie C. Hicks, Johns Hopkins University
3:00	**A Comprehensive Evaluation of Preprocessing Methods for Single-Cell RNA Sequencing Data** Shih-Kai Chu*, Qi Liu and Yu Shyr, Vanderbilt University Medical Center
3:15	**Fast Clustering for Single-Cell RNA-seq Data using Mini-Batch k-Means** Stephanie C. Hicks*, Johns Hopkins Bloomberg School of Public Health; Ruoxi Liu, Johns Hopkins University; Yuwei Ni, Weill Cornell Medical College; Elizabeth Purdom, University of California, Berkeley; Davide Risso, University of Padova

92. CONTRIBUTED PAPERS: ROBUST MODELING AND INFERENCE

Sponsor: ENAR
Chair: Julia Wrobel, Colorado School of Public Health

1:45	**Robust Estimation with Outcome Misclassification and Covariate Measurement Error in Logistic Regression** Sarah C. Lotspeich*, Bryan E. Shepherd and Gustavo G.C. Amorim, Vanderbilt University Medical Center; Pamela A. Shaw, University of Pennsylvania; Ran Tao, Vanderbilt University Medical Center
2:00	**Implementing Interventions to Combat the Opioid Epidemic During a Rising Tide of Activities Aimed at Improving Patient Outcomes: Evaluating the Robustness of Parallel-Group and Stepped-Wedge Cluster Randomized Trials to Confounding from External Events** Lior Rennert*, Clemson University
2:15	**Robust Statistical Models for Impact Injury Risk Estimation** Anjishnu Banerjee* and Narayan Yoganandan, Medical College of Wisconsin
2:30	**Joint Testing of Donor/Recipient Genetic Matching Scores and Recipient Genotype has Robust Power for Finding Genes Associated with Transplant Outcomes** Victoria L. Arthur*, University of Pennsylvania Perelman School of Medicine; Sharon Browning, University of Washington; Bao-Li Chang and Brendan Keating, University of Pennsylvania; Jinbo Chen, University of Pennsylvania Perelman School of Medicine

SCIENTIFIC PROGRAM
(CONTINUED)

2:45	**A Robust Bayesian Copas Selection Model for Detecting and Correcting Publication Bias** Ray Bai*, Yong Chen and Mary Regina Boland, University of Pennsylvania
3:00	**Estimation of Knots in Linear Spline Models** Guangyu Yang*, University of Michigan, Ann Arbor; Baqun Zhang, Shanghai University of Finance and Economics; Min Zhang, University of Michigan, Ann Arbor
3:15	**Floor Discussion**

TUESDAY, MARCH 24
3:30 p.m. — 3:45 p.m.

REFRESHMENT BREAK WITH OUR EXHIBITORS

TUESDAY, MARCH 24
3:45 p.m. — 5:30 p.m.

93. HIGH DIMENSIONAL METHODS FOR MECHANISTIC INTEGRATION OF MULTI-TYPE OMICS

Sponsor: IMS
Organizer: Qi Zhang, University of Nebraska, Lincoln
Chair: Min Jin Ha, University of Texas MD Anderson Cancer Center

3:45	**Integrating Heterogeneous Longitudinal Omics Data with Personalized Dynamic Network Analysis** Xing Qiu*, University of Rochester; Leqin Wu, Jinan University; Ya-xiang Yuan, Chinese Academy of Sciences; Hulin Wu, University of Texas Health Science Center at Houston
4:10	**INFIMA Leverages Multi-Omic Model Organism Data to Identify Target Genes for Human GWAS Variants** Sunduz Keles* and Chenyang Dong, University of Wisconsin, Madison
4:35	**Nonlinear Moderated Mediation Analysis with Genetical Genomics Data** Yuehua Cui* and Bin Gao, Michigan State University; Xu Liu, Shanghai University of Finance and Economics
5:00	**High Dimensional Mediation Analysis for Causal Gene Selection** Qi Zhang*, University of Nebraska, Lincoln
5:25	**Floor Discussion**

94. NEW WEIGHTING METHODS FOR CAUSAL INFERENCE

Sponsors: ENAR, IMS, ASA Biometrics Section, ASA Section on Statistics in Epidemiology
Organizer: Roland Matsouaka, Duke University
Chair: Hwanhee Hong, Duke University

3:45	**Propensity Score Weighting for Causal Inference with Multiple Treatments** Fan Li*, Yale School of Public Health; Fan Li, Duke University
4:10	**Methods for Balancing Covariates when Estimating Heterogeneous Treatment Effects in Observational Data** Laine Thomas* and Fan Li, Duke University; Daniel Wojdyla, Duke Clinical Research Institute; Siyun Yang, Duke University
4:35	**Flexible Regression Approach to Propensity Score Analysis and its Relationship with Matching and Weighting** Liang Li*, University of Texas MD Anderson Cancer Center; Huzhang Mao, Eli Lilly and Company
5:00	**Robust Inference when Combining Probability and Non-Probability Samples with High-Dimensional Data** Shu Yang*, North Carolina State University; Jae Kwang Kim, Iowa State University; Rui Song, North Carolina State University
5:25	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

95. USING MACHINE LEARNING TO ANALYZE RANDOMIZED TRIALS: VALID ESTIMATES AND CONFIDENCE INTERVALS WITHOUT MODEL ASSUMPTIONS

Sponsors: ENAR, IMS, ASA Biometrics Section, ASA Section on Statistical Learning and Data Science
Organizer: Michael Rosenblum, Johns Hopkins Bloomberg School of Public Health
Chair: Bingkai Wang, Johns Hopkins University

3:45	**Performance Evaluation of Flexible Strategies for Estimating HIV Vaccine Efficacy** Alex Luedtke*, University of Washington
4:10	**Inference for Model-Light Machine Learning in Precision Medicine** Michael Kosorok*, University of North Carolina, Chapel Hill
4:35	**Synthetic Difference in Differences** David A. Hirshberg*, Stanford University; Dmitry Arkhangelsky, CEMFI, Madrid; Susan Athey, Guido Imbens and Stefan Wager, Stanford University
5:00	**Machine Learning Versus Standard Methods for Covariate Adjustment: Performance Comparison Across 10 Completed Randomized Trials** Michael M. Rosenblum*, Johns Hopkins Bloomberg School of Public Health
5:25	**Floor Discussion**

96. RECENT DEVELOPMENTS IN SEMIPARAMETRIC TRANSFORMATION MODELS

Sponsors: ENAR, ASA Biometrics Section
Organizer: Chun Li, Case Western Reserve University
Chair: Gustavo Amorim, Vanderbilt University Medical Center

3:45	**Semiparametric Regression Models for Indirectly Observed Outcomes** Jan De Neve* and Heidelinde Dehaene, Ghent Unversity
4:15	**Addressing Outcome Detection Limits using Semiparametric Cumulative Probability Models** Bryan E. Shepherd* and Yuqi Tian, Vanderbilt University
4:45	**Cumulative Probability Models for Big Data** Chun Li*, Case Western Reserve University
5:15	**Discussant:** Frank Harrell, Vanderbilt University

97. INNOVATIONS IN STATISTICAL NEUROSCIENCE

Sponsors: ENAR, ASA Section on Statistics in Imaging
Organizer: Jeff Goldsmith, Columbia University
Chair: Jeff Goldsmith, Columbia University

3:45	**A Study of Longitudinal Trends in Time-Frequency Transformations of EEG Data During a Learning Experiment** Damla Senturk*, Joanna Boland, Shafali Jeste and Donatello Telesca, University of California, Los Angeles
4:10	**Improved Diagnostics and Prognostics using MRI in Multiple Sclerosis** Russell Shinohara*, University of Pennsylvania
4:35	**Intensity Warping for Multisite MRI Harmonization** Julia L. Wrobel*, Colorado School of Public Health; Melissa Martin and Taki Shinohara, University of Pennsylvania; Jeff Goldsmith, Columbia University
5:00	**Bayesian Approaches for Estimating Dynamic Functional Network Connectivity in fMRI Data** Michele Guindani*, University of California, Irvine
5:25	**Floor Discussion**

98. ARTIFICIAL INTELLIGENCE FOR PREDICTION OF HEALTH OUTCOMES

Sponsors: ENAR, ASA Biometrics Section, ASA Health Policy Statistics Section, ASA Section on Statistical Learning and Data Science
Organizer: Lihui Zhao, Northwestern University
Chair: Lei Liu, Washington University in St. Louis

3:45	**Distributed Learning from Multiple EHR Databases for Predicting Medical Events** Qi Long*, University of Pennsylvania; Ziyi Li, Emory University; Kirk Roberts and Xiaoqian Jiang, University of Texas Health Science Center at Houston
4:10	**Deep Learning with Time-to-Event Outcomes** Jon Steingrimsson*, Samantha Morrison and Constantine Gatsonis, Brown University
4:35	**A Scalable Discrete-Time Survival Model for Neural Networks** Balasubramanian Narasimhan*, Stanford University
5:00	**Deep Learning for Dynamic Prediction of Cardiovascular Events** Lihui Zhao*, Northwestern University
5:25	**Floor Discussion**

SCIENTIFIC PROGRAM

(CONTINUED)

99. CONTRIBUTED PAPERS: LATENT VARIABLES AND PROCESSES

Sponsor: ENAR
Chair: Donglin Zeng, University of North Carolina, Chapel Hill

3:45	**Modeling the Effects of Multiple Exposures with Unknown Group Memberships: A Bayesian Latent Variable Approach** Alexis E. Zavez*, University of Rochester Medical Center; Emeir M. McSorley, Ulster University; Sally W. Thurston, University of Rochester Medical Center
4:00	**A Time-Dependent Structural Model Between Latent Classes and Competing Risks Outcomes** Teng Fei*, John Hanfelt and Limin Peng, Emory University
4:15	**Dirichlet Depths for Point Process** Kai Qi*, Yang Chen and Wei Wu, Florida State University
4:30	**Acknowledging the Dilution Effect in Group Testing Regression: A New Approach** Stefani C. Mokalled*, Christopher S. McMahan and Derek A. Brown, Clemson University; Joshua M. Tebbs, University of South Carolina; Christopher R. Bilder, University of Nebraska, Lincoln
4:45	**Modeling Brain Waves as a Mixture of Latent Processes** Guillermo Cuauhtemoctzin Granados Garcia* and Hernando Ombao, King Abdullah University of Science and Technology; Mark Fiecas, University of Minnesota; Babak Shahbaba, University of California, Irvine
5:00	**A Method to Flexibly Incorporate Covariates in Latent Class Analysis with Application to Mild Cognitive Impairment** Grace Kim* and John Hanfelt, Emory University Rollins School of Public Health
5:15	**Exploration of Misspecification in Latent Class Trajectory Analysis (LCTA) and Growth Mixture Modeling (GMM): Error Structure Matters** Megan L. Neely*, Jane Pendergast and Bida Gu, Duke University; Natasha Dmitreava, Duke University Medical Center; Carl Pieper, Duke University

100. CONTRIBUTED PAPERS: TIME-TO-EVENT DATA ANALYSIS: SURVIVAL OF THE FITTEST

Sponsor: ENAR
Chair: Richard Chappell, University of Wisconsin, Madison

3:45	**Survival Analysis under the Cox Proportional Hazards Model with Pooled Covariates** Paramita Saha Chaudhuri* and Lamin Juwara, McGill University
4:00	**Quantile Association Regression on Bivariate Survival Data** Ling-Wan Chen*, National Institute of Environmental Health Sciences, National Institutes of Health; Yu Cheng and Ying Ding, University of Pittsburgh; Ruosha Li, University of Texas Health Science Center at Houston
4:15	**Restricted Mean Survival Time as a Function of Restriction Time** Yingchao Zhong*, University of Michigan; Douglas E. Schaubel, University of Pennsylvania
4:30	**Quantile Regression on Cause-Specific Inactivity Time** Yichen Jia* and Jong-Hyeon Jeong, University of Pittsburgh
4:45	**Relaxing the Independence Assumption in Relative Survival Analysis: A Parametric Approach** Reuben Adatorwovor* and Jason Fine, University of North Carolina at Chapel Hill; Aurelien Latouche, Conservatoire National des Arts et Métiers and Institut Curie, St-Cloud, France
5:00	**Estimation of Effect Measures in Survival Analysis that Allow Causal Interpretation** Kjetil Røysland*, University of Oslo
5:15	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

101. CONTRIBUTED PAPERS: RISKY BUSINESS: DIAGNOSTICS, ROC, AND PREDICTION

Sponsor: ENAR
Chair: Gong Tang, University of Pittsburgh

3:45	**NMADiagT: An R package for Network Meta-Analysis of Multiple Diagnostic Tests** Boyang Lu*, University of Minnesota; Qinshu Lian, Genentech; James S. Hodges and Haitao Chu, University of Minnesota
4:00	**Informative Back-End Screening** Michael R. Stutz* and Joshua M. Tebbs, University of South Carolina
4:15	**Patient-Reported Outcome (PRO) Assessment in Diagnostic Devices: A Novel Approach** Saryet Kucukemiroglu* and Manasi Sheth, U.S. Food and Drug Administration
4:30	**A Placement-Value Based Approach to Concave ROC Curves** Soutik Ghosal* and Zhen Chen, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
4:45	**Inference in ROC Curves for Two-Phase Nested Case-Control Biomarker Studies** Leonidas E. Bantis*, University of Kansas Medical Center; Ziding Feng, Fred Hutchinson Cancer Research Center
5:00	**Diagnostic Evaluation of Quantitative Features of Functional Markers** Jeong Hoon Jang*, Indiana University; Amita K. Manatunga, Emory University
5:15	**Evaluation of Multiple Diagnostic Tests using Multi-Institutional Data with Missing Components** Jiasheng Shi*, Children's Hospital of Philadelphia; Jing Huang, University of Pennsylvania and The Children's Hospital of Philadelphia; Yong Chen, University of Pennsylvania

102. CONTRIBUTED PAPERS: INTERVAL-CENSORED AND MULTIVARIATE SURVIVAL DATA

Sponsor: ENAR
Chair: Jianxin Shi, National Cancer Institute, National Institutes of Health

3:45	**A Divide-and-Combine Approach of Multivariate Survival Analysis in Big Data** Wei Wang* and Shou-En Lu, Rutgers University; Jerry Q. Cheng, New York Institute of Technology
4:00	**Nonparametric Inference for Nonhomogeneous Multi-State Processes Based on Clustered Observations** Giorgos Bakoyannis*, Indiana University
4:15	**Flexible, Unified Approach for Analyzing Arbitrarily-Censored and/or Left-Truncated Interval-Censored Data** Prabhashi Withana Gamage*, James Madison University; Christopher McMahan, Clemson University; Lianming Wang, University of South Carolina
4:30	**Potential Intransitivity of Win-Ratio Preferences: Is it a Problem and What Do We Do About It?** David Oakes*, University of Rochester
4:45	**Bayesian Analysis of Multivariate Survival Data Based on Vine Copulas** Guanyu Hu*, University of Connecticut; Dooti Roy, Boehringer Ingelheim; Dipak Dey, University of Connecticut
5:00	**Non-parametric estimation in an illness-death model with component-wise censoring** Anne Eaton*, University of Minnesota
5:15	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

103. CONTRIBUTED PAPERS: GRAPHICAL MODELS AND APPLICATIONS

Sponsor: ENAR
Chair: Russell Stocker, Indiana University of Pennsylvania

3:45	**Inference of Large Modified Poisson-Type Graphical Models: Application to RNA-seq Data in Childhood Atopic Asthma Studies** Rong Zhang*, University of Pittsburgh; Juan C. Celedon and Wei Chen, UPMC Children's Hospital of Pittsburgh; Zhao Ren, University of Pittsburgh
4:00	**Assisted Estimation of Gene Expression Graphical Models** Huangdi Yi*, Yale School of Public Health; Yifan Sun, Renmin University of China; Qingzhao Zhang, Xiamen University; Yang Li, Renmin University of China; Shuangge Ma, Yale School of Public Health
4:15	**Directed Acyclic Graph Assisted Methods for Estimating Average Treatment Effect** Jingchao Sun*, Maiying Kong, Scott Davis Duncan and Subhadip Pal, University of Louisville
4:30	**Gene Network Analysis Based on Single Cell RNA Sequencing Data** Meichen Dong* and Fei Zou, University of North Carolina, Chapel Hill
4:45	**Selection and Estimation of Conditional Graphical Models** Stephen Salerno* and Yi Li, University of Michigan
5:00	**Joint Estimation of the Two-Level Gaussian Graphical Models across Multiple Classes** Inyoung Kim*, Virginia Tech; Liang Shan, University of Alabama at Birmingham
5:15	**Floor Discussion**

104. CONTRIBUTED PAPERS: SUPPORT VECTOR MACHINES, NEURAL NETWORKS AND DEEP LEARNING

Sponsor: ENAR
Chair: Tanzy M. Love, University of Rochester

3:45	**ForgeNet: A Graph Deep Neural Network Model Using Tree-Based Ensemble Classifiers for Feature Graph Construction** Yunchuan Kong* and Tianwei Yu, Emory University
4:00	**GWAS-Based Deep Learning for Survival Prediction** Tao Sun*, Wei Chen and Ying Ding, University of Pittsburgh
4:15	**An Inferential Framework for Individualized Minimal Clinically Importance Difference with a Linear Structure** Zehua Zhou* and Jiwei Zhao, State University of New York at Buffalo
4:30	**Deep Neural Networks for Survival Analysis Using Pseudo Values** Lili Zhao*, University of Michigan; Feng Dai, Merck & Co., Inc.
4:45	**Neural Networks for Clustered and Longitudinal Data using Mixed Effects Models** Francesca Mandel* and Ian Barnett, University of Pennsylvania
5:00	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

WEDNESDAY, MARCH 25
8:30 a.m. — 10:15 a.m.

105. ADVANCES IN STATISTICAL MODELING FOR MULTI-OMICS DATA INTEGRATION

Sponsor: IMS
Organizer: Sunyoung Shin, University of Texas, Dallas
Chair: Sunyoung Shin, University of Texas, Dallas

8:30	**Gene-Set Integrative Omics Analysis Using Tensor-Based Association Tests** Jung-Ying Tzeng*, North Carolina State University; Meng Yang, The SAS Institute; Wenbin Lu, North Carolina State University; Jeff Miecznikowski, University of Buffalo; Sheng-Mao Chang, National Cheng-Kung University
8:55	**Radiogenomic Analysis of Lower Grade Gliomas Incorporating Tumor Heterogeneity in Imaging Through Densities** Shariq Mohammed*, University of Michigan; Sebastian Kurtek, The Ohio State University; Karthik Bharath, University of Nottingham; Arvind Rao and Veerabhadran Baladandayuthapani, University of Michigan
9:20	**Bayesian Regression and Clustering Models to Incorporate Multi-Layer Overlapping Group Structure in Multi-Omics Applications** George Tseng*, University of Pittsburgh
9:45	**Graphical Models for Data Integration and Mediation Analysis** Min Jin Ha*, University of Texas MD Anderson Cancer Center; Veera Baladandayuthapani, University of Michigan
10:10	**Floor Discussion**

106. CAUSAL INFERENCE AND NETWORK DEPENDENCE: FROM PEER EFFECTS TO THE REPLICATION CRISIS IN EPIDEMIOLOGY

Sponsors: ENAR, ASA Biometrics Section, ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section
Organizer: Corwin Zigler, University of Texas, Austin and Dell Medical School
Chair: Corwin Zigler, University of Texas, Austin and Dell Medical School

8:30	**Social Network Dependence, the Replication Crisis, and (in)valid Inference** Elizabeth L. Ogburn*, Johns Hopkins University
8:55	**Nonparametric Identification of Causal Intervention Effects Under Contagion** Forrest W. Crawford* and Xiaoxuan Cai, Yale School of Public Health; Wen Wei Loh, University of Ghent
9:20	**Bayesian Auto-g-Computation of Network Causal Effects: Incarceration and Infection in a High Risk Network** Isabel R. Fulcher*, Harvard Medical School; Eric J. Tchetgen Tchetgen, University of Pennsylvania; Ilya Shpitser, Johns Hopkins University
9:45	**Heterogeneous Causal Effects under Network Interference** Laura Forastiere*, Yale University; Costanza Tortú and Falco Bargagli-Stoffi, IMT Lucca, Italy
10:10	**Floor Discussion**

107. FLEXIBLE SPATIO-TEMPORAL MODELS FOR ENVIRONMENTAL AND ECOLOGICAL PROCESSES

Sponsor: ENAR, ASA Section on Statistics and the Environment
Organizer: Alexandra Schmidt, McGill University
Chair: Alexandra Schmidt, McGill University

8:30	**Evaluating Proxy Influence in Assimilated Paleoclimate Reconstructions - Testing the Exchangeability of Two Ensembles of Spatial Processes** Bo Li* and Trevor Harris, University of Illinois at Urbana-Champaign; Nathan Steiger and Jason Smerdon, Columbia University; Naveen Narisetty, University of Illinois at Urbana-Champaign; J. Derek Tucker, Sandia National Lab
8:55	**Fusing Multiple Existing Space-Time Categorical Land Cover Datasets** Amanda S. Hering*, Baylor University; Nicolás Rodríguez-Jeangros and John E. McCray, Colorado School of Mines
9:20	**Inverse Reinforcement Learning for Animal Behavior from Environmental Cues** Toryn L.J. Schafer* and Christopher K. Wikle, University of Missouri
9:45	**High-dimensional multivariate Geostatistics: A Bayesian Matrix-Normal Approach** Lu Zhang* and Sudipto Banerjee, UCLA-Fielding School of Public Health
10:10	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

108. RECENT ADVANCES IN NEUROIMAGING ANALYTICS

Sponsors: ENAR, ASA Biometrics Section
Organizer: Zainab Albar, Case Western Reserve University School of Medicine and Quantitative Health Sciences
Chair: Abdus Sattar, Case Western Reserve University School of Medicine and Quantitative Health Sciences

8:30	**Covariance Regression in Brain Imaging** Brian S. Caffo*, Johns Hopkins University; Yi Zhao, Indiana University Purdue University Indianapolis; Bingkai Wang, Johns Hopkins University; Xi (Rossi) Luo, University of Texas Health Science Center at Houston
8:55	**Bayesian Modeling of Multiple Structural Connectivity Networks During the Progression of Alzheimer's Disease** Christine Peterson*, The University of Texas MD Anderson Cancer Center
9:20	**Modeling Lead-Lag Dynamics in High Dimensional Time Series** Hernando Ombao* and Chee-Ming Ting, King Abdullah University of Science and Technology; Marco Pinto, Oslo Metropolitan University
9:45	**Modeling Positive Definite Matrices in Diffusion Tensor Imaging** Dipankar Bandyopadhyay*, Virginia Commonwealth University; Zhou Lan, The Pennsylvania State University; Brian J. Reich, North Carolina State University
10:10	**Floor Discussion**

109. NOVEL TENSOR METHODS FOR COMPLEX BIOMEDICAL DATA

Sponsors: ENAR, ASA Biometrics Section, ASA Statistics in Genomics and Genetics Section, ASA Section on Statistical Learning and Data Science, ICSA
Organizer: Gen Li, Columbia University
Chair: Gen Li, Columbia University

8:30	**Generalized Tensor Regression with Covariates on Multiple Modes** Miaoyan Wang*, Zhuoyan Xu and Jiaxin Hu, University of Wisconsin, Madison
8:55	**Co-Manifold Learning on Tensors** Eric Chi*, North Carolina State University
9:20	**Nonparametric Regression for Brain Imaging Data Analysis** Weining Shen*, University of California, Irvine
9:45	**Brain Regions Identified as Being Associated with Verbal Reasoning through the Use of Imaging Regression via Internal Variation** Xuan Bi*, University of Minnesota; Long Feng and Heping Zhang, Yale University
10:10	**Floor Discussion**

110. INTEGRATIVE ANALYSIS OF CLINICAL TRIALS AND REAL-WORLD EVIDENCE STUDIES

Sponsors: ENAR, IMS, ASA Biometrics Section
Organizer: Shu Yang, North Carolina State University
Chair: Shu Yang, North Carolina State University

8:30	**On Using Electronic Health Records to Improve Optimal Treatment Rules in Randomized Trials** Peng Wu*, Columbia University and Visa Inc.; Donglin Zeng, University of North Carolina, Chapel Hill; Haoda Fu, Eli Lilly and Company; Yuanjia Wang, Columbia University
8:55	**Making Use of Information Contained in Existing Black-Box-Type Risk Calculators** Peisong Han*, Jeremy M.G. Taylor and Bhramar Mukherjee, University of Michigan
9:20	**Integrative Analysis of Randomized Clinical Trials with Real World Evidence Studies** Lin Dong*, Wells Fargo Bank; Shu Yang, North Carolina State University
9:45	**Risk Projection for Time-to-Event Outcome Leveraging External Summary Statistics with Source Individual-Level Data** Jiayin Zheng*, Li Hsu and Yingye Zheng, Fred Hutchinson Cancer Research Center
10:10	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

111. CONTRIBUTED PAPERS: CLUSTERED DATA METHODS

Sponsor: ENAR
Chair: Sanjoy K. Sinha, Carleton University

8:30	**Modeling Tooth-Loss using Inverse Probability Censoring Weights in Longitudinal Clustered Data with Informative Cluster Size** Aya A. Mitani*, Harvard T. H. Chan School of Public Health; Elizabeth K. Kaye, Boston University Henry M. Goldman School of Dental Medicine; Kerrie P. Nelson, Boston University School of Public Health
8:45	**Partially Pooled Propensity Score Models for Average Treatment Effect Estimation with Multilevel Data** Youjin Lee*, University of Pennsylvania; Trang Nguyen and Elizabeth Stuart, Johns Hopkins Bloomberg School of Public Health
9:00	**Outcome-Guided Disease Subtyping for High-Dimensional Omics Data** Peng Liu*, Lu Tang and George Tseng, University of Pittsburgh
9:15	**The Impact of Sample Size Re-Estimation using Baseline ICC in Cluster Randomized Trials: A Simulation Study** Kaleab Z. Abebe*, Kelley A. Jones, Taylor Paglisotti and Elizabeth Miller, University of Pittsburgh; Daniel J. Tancredi, University of California, Davis
9:30	**Hypothesis Testing for Community Detection in Network Data** Chetkar Jha*, Mingyao Li and Ian Barnett, University of Pennsylvania
9:45	**On the Interplay Between Exposure Misclassification and Informative Cluster Size** Glen McGee*, Harvard University; Marianthi-Anna Kioumourtzoglou, Columbia University; Marc G. Weisskopf, Sebastien Haneuse and Brent A. Coull, Harvard University
10:00	**An Alternative to the Logistic GLMM with Normal Random Effects for Estimating Dose Response in the Presence of Extreme Between Subject Heterogeneity** Joe Bible* and Christopher McMahan, Clemson University

112. CONTRIBUTED PAPERS: SUBGROUP ANALYSIS

Sponsor: ENAR
Chair: Sheng Luo, Duke University

8:30	**Inference on Selected Subgroups in Clinical Trials** Xinzhou Guo*, Harvard University; Xuming He, University of Michigan
8:45	**A Simultaneous Inference Procedure to Identify Subgroups in Targeted Therapy Development with Time-to-Event Outcomes** Yue Wei*, University of Pittsburgh; Jason Hsu, The Ohio State University; Ying Ding, University of Pittsburgh
9:00	**Cross-Platform Omics Prediction (CPOP) Procedure Enables Precision Medicine** Kevin Y.X. Wang*, The University of Sydney; Varsha Tembe and Gullietta Pupo, Melanoma Institute Australia and The University of Sydney; Garth Tarr and Samuel Mueller, The University of Sydney; Graham Mann, Melanoma Institute Australia and The University of Sydney; Jean Y.H. Yang, The University of Sydney
9:15	**Bayesian Subgroup Analysis in Regression using Mixture Models** Yunju Im* and Aixin Tan, University of Iowa
9:30	**Adaptive Subgroup Identification in Phase I-II Clinical Trials** Alexandra M. Curtis* and Brian J. Smith, University of Iowa; Andrew G. Chapple, Louisiana State University School of Public Health
9:45	**Identifying Effect Modifiers and Subgroups that May Benefit from Treatment when the Number of Covariates is Large** John A. Craycroft*, Maiying Kong and Subhadip Pal, University of Louisville
10:00	**Floor Discussion**

SCIENTIFIC PROGRAM

(CONTINUED)

113. CONTRIBUTED PAPERS: FUNCTIONAL DATA ANALYSIS: BELOW THE SURFACE

Sponsor: ENAR
Chair: Lihui Zhao, Northwestern University

8:30	**Imaging Genetics: Where the Statistics of fMRI and Genome-Wide Association Studies Collide** Kristen N. Knight*, University of Georgia
8:45	**Bayesian Quantile Monotone Single-Index Model for Bounded Response Using Functional and Scalar Predictors** Bradley B. Hupf*, Debajyoti Sinha, Eric Chicken and Greg Hajcak, Florida State University
9:00	**Sparse Log-Contrast Regression with Functional Compositional Predictors: Linking Gut Microbiome Trajectory in Early Postnatal Period to Neurobehavioral Development of Preterm Infants** Zhe Sun*, Wanli Xu and Xiaomei Cong, University of Connecticut; Gen Li, Columbia University; Kun Chen, University of Connecticut
9:15	**Principle ERP Reduction and Analysis** Emilie Campos*, Chad Hazlett, Patricia Tan, Holly Truong, Sandra Loo, Charlotte DiStefano, Shafali Jeste and Damla Senturk, University of California, Los Angeles
9:30	**Approaches for Extending Multiple Imputation to Handle Scalar and Functional Data** Adam Ciarleglio*, The George Washington University
9:45	**Statistical Analysis of Heart Rate Variability from Electrocardiogram Data** Andrada E. Ivanescu*, Montclair State University; Naresh Punjabi and Ciprian M. Crainiceanu, Johns Hopkins University
10:00	**Interpretable Principal Components Analysis for Multilevel Multivariate Functional Data, with Application to EEG Experiments** Jun Zhang* and Greg J. Siegle, University of Pittsburgh; Wendy D' Andrea, New School for Social Research; Robert T. Krafty, University of Pittsburgh

114. CONTRIBUTED PAPERS: HIV, INFECTIOUS DISEASE AND MORE

Sponsor: ENAR
Chair: Ming Wang, The Pennsylvania State University

8:30	**A Hybrid Compartment/Agent-Based Model for Infectious Disease Modeling** Shannon Gallagher*, National Institute of Allergy and Infectious Diseases, National Institutes of Health; William Eddy, Carnegie Mellon University
8:45	**Analysis of Two-Phase Studies using Generalized Method of Moments** Prosenjit Kundu*, Johns Hopkins Bloomberg School of Public Health; Nilanjan Chatterjee, Johns Hopkins Bloomberg School of Public Health and Johns Hopkins University School of Medicine
9:00	**Bias and Efficiency in Group Testing Estimation for Infectious Disease Surveillance** Katherine M. Bindbeutel* and Md S. Warasi, Radford University
9:15	**Mediation Effect Sizes for Latent Outcome Models using Explained Variance Decomposition** Yue Jiang*, University of North Carolina, Chapel Hill; Shanshan Zhao, National Institute of Environmental Health Sciences, National Institutes of Health; Jason Peter Fine, University of North Carolina, Chapel Hill
9:30	**Toward Evaluation of Disseminated Effects of Non-Randomized HIV Prevention Interventions Among Observed Networks of People who Inject Drugs** Ashley Buchanan*, Natallia Katenka and TingFang Lee, University of Rhode Island; M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington; Samuel Friedman, New York University; Georgios Nikolopoulos, University of Cyprus
9:45	**Joint Model of Adherence to Dapivirine-containing Vaginal Ring and HIV-1 Risk** Qi Dong*, University of Washington; Elizabeth R. Brown and Jingyang Zhang, Fred Hutchinson Cancer Research Center
10:00	**The Mechanistic Analysis of Founder Virus Data in Challenge Models** Ana Maria Ortega-Villa* and Dean A. Follmann, National Institute of Allergy and Infectious Diseases, National Institutes of Health

SCIENTIFIC PROGRAM
(CONTINUED)

115. CONTRIBUTED PAPERS: CLINICAL TRIAL DESIGN AND ANALYSIS

Sponsor: ENAR
Chair: Fang Liu, University of Notre Dame

8:30	**Bayesian Design of Clinical Trials for Joint Models of Longitudinal and Time-to-Event Data** Jiawei Xu*, Matthew A. Psioda and Joseph G. Ibrahim, University of North Carolina, Chapel Hill
8:45	**Statistical Support for Designing Non-Inferiority Trials: An Application to Rheumatoid Arthritis** Rebecca Rothwell* and Gregory Levin, U.S. Food and Drug Administration
9:00	**Determining Mental Health Condition Patterns in Veterans with a Lifetime PTSD Diagnosis** Ilaria Domenicano*, Department of Veterans Affairs Cooperative Studies Program and Yale School of Public Health; Lori L. Davis, Tuscaloosa Veterans Affairs Medical Center and University of Alabama School of Medicine; Lisa Mueller, Edith Nourse Rogers Memorial Veterans Hospital; Tassos Constantino Kyriakides, Department of Veterans Affairs Cooperative Studies Program and Yale School of Public Health
9:15	**Estimation of Ascertainment Bias and its Effect on Power in Clinical Trials with Time-to-Event Outcomes** Erich J. Greene*, Peter Peduzzi, James Dziura, Can Meng and Denise Esserman, Yale Center for Analytical Sciences
9:30	**Design and Analysis Considerations for Utilizing a Tailoring Function in a snSMART with Continuous Outcomes** Holly E. Hartman*, University of Michigan; Roy N. Tamura, University of South Florida; Matthew J. Schipper and Kelley Kidwell, University of Michigan
9:45	**Two-Part Proportional Mixed Effects Model for Clinical Trials in Alzheimer's Disease** Guoqiao Wang*, Yan Li, Chengjie Xiong, Lei Liu, Andrew Aschenbrenner, Jason Hassenstab, Eric McDade and Randall Bateman, Washington University in St. Louis
10:00	**Floor Discussion**

116. CONTRIBUTED PAPERS: MULTIVARIATE AND HIGH-DIMENSIONAL DATA ANALYSIS

Sponsor: ENAR
Chair: Qiwei Li, University of Texas, Dallas

8:30	**On Genetic Correlation Estimation with Summary Statistics from Genome-Wide Association Studies** Bingxin Zhao* and Hongtu Zhu, University of North Carolina, Chapel Hill
8:45	**Multivariate Association Analysis with Correlated Traits in Related Individuals** Souvik Seal*, University of Minnesota
9:00	**Grafted and Vanishing Random Subspaces** Matthew Corsetti* and Tanzy Love, University of Rochester
9:15	**Modeling Repeated Multivariate Data to Estimate Individuals' Trajectories with Application to Scleroderma** Ji Soo Kim*, Johns Hopkins University; Ami Shah and Laura Hummers, Johns Hopkins University School of Medicine; Scott L. Zeger, Johns Hopkins University
9:30	**Nonignorable Item Nonresponse in Multivariate Outcomes** Sijing Li* and Jun Shao, University of Wisconsin, Madison
9:45	**Multivariate Association Analysis with Somatic Mutation Data** Chad He*, Fred Hutchinson Cancer Research Center; Yang Liu, Wright State University; Ulrike Peters and Li Hsu, Fred Hutchinson Cancer Research Center
10:00	**Floor Discussion**

WEDNESDAY, MARCH 25
10:15 a.m. — 10:30 a.m.

REFRESHMENT BREAK WITH OUR EXHIBITORS

SCIENTIFIC PROGRAM
(CONTINUED)

WEDNESDAY, MARCH 25
10:30 a.m. — 12:15 p.m.

117. ASYMMETRICAL STATISTICAL LEARNING FOR BINARY CLASSIFICATION

Sponsor: IMS
Organizer: Jingyi Li, University of California, Los Angeles
Chair: Anqi Zhao, National University of Singapore

10:30	**Introduction to Neyman-Pearson Classification** Jingyi Jessica Li*, University of California, Los Angeles; Xin Tong, University of Southern California; Yang Feng, Columbia University
10:55	**A Unified View of Asymmetric Binary Classification** Wei Vivian Li*, Rutgers, The State University of New Jersey; Xin Tong, University of Southern California; Jingyi Jessica Li, University of California, Los Angeles
11:20	**Neyman-Pearson Classification: Parametrics and Sample Size Requirement** Yang Feng*, New York University
11:45	**Intentional Control of Type I Error over Unconscious Data Distortion: A Neyman-Pearson Approach to Text Classification** Xin Tong*, University of Southern California; Lucy Xia, Hong Kong University of Science and Technology; Richard Zhao, The Pennsylvania State University; Yanhui Wu, University of Southern California
12:10	**Floor Discussion**

118. RECENT ADVANCES AND OPPORTUNITIES IN LARGE SCALE & MULTI-OMIC SINGLE-CELL DATA ANALYSIS

Sponsors: ENAR, ASA Biometrics Section, ASA Statistics in Genomics and Genetics Section
Organizer: Rhonda Bacher, University of Florida
Chair: Mengjie Chen, University of Chicago

10:30	**Statistical Analysis of Coupled Single-Cell RNA-seq and Immune Profiling Data** Hongkai Ji* and Zhicheng Ji, Johns Hopkins Bloomberg School of Public Health
11:00	**Assessing Consistency of Single Cell Unsupervised Multi-Omics Methods** Michael I. Love*, University of North Carolina, Chapel Hill
11:30	**Statistical Methods for Identifying and Characterizing Cell Populations using High-Dimensional Single-Cell Data** Raphael Gottardo*, Fred Hutchinson Cancer Research Center
12:00	**Discussant:** Rhonda Bacher, University of Florida

119. NOVEL STATISTICAL METHODS FOR COMPLEX INTERVAL-CENSORED SURVIVAL DATA

Sponsors: ENAR, ASA Biometrics Section, ASA Section on Statistics in Epidemiology
Organizer: Sedigheh Mirzaei Salehabadi, St. Jude Children's Research Hospital
Chair: Sedigheh Mirzaei Salehabadi, St. Jude Children's Research Hospital

10:30	**Semiparametric Regression Analysis of Multiple Censored Events in Family Studies** Donglin Zeng*, University of North Carolina, Chapel Hill; Fei Gao, Fred Hutchinson Cancer Research Center; Yuanjia Wang, Columbia University
10:55	**AModeling Interval Censored Time to Event Outcomes with Inflation of Zeros, with Application to Pediatric HIV Studies** Raji Balasubramanian*, University of Massachusetts, Amherst
11:20	**Case-Cohort Studies with Multiple Interval-Censored Disease Outcomes** Qingning Zhou*, University of North Carolina, Charlotte; Jianwen Cai and Haibo Zhou, University of North Carolina, Chapel Hill
11:45	**Adjusting for Covariate Measurement Error in Survival Analysis under Competing Risks** Sharon Xiangwen Xie* and Carrie Caswell, University of Pennsylvania
12:10	**Floor Discussion**

SCIENTIFIC PROGRAM
(CONTINUED)

120. MODERN GRAPHICAL MODELING OF COMPLEX BIOMEDICAL SYSTEMS

Sponsor: ENAR, ASA Bayesian Statistical Science Section
Organizer: Lili Zhao, University of Michigan
Chair: Lili Zhao, University of Michigan

10:30	**A Tripartite Latent Graph for Phenotype Discovery in EHR Data** Peter Mueller*, University of Texas, Austin; Yang Ni, Texas A&M University; Yuan Ji, The University of Chicago
10:55	**The Reduced PC-Algorithm: Improved Causal Structure Learning in Large Random Networks** Ali Shojaie*, University of Washington
11:20	**Latent Network Estimation and Variable Selection for Compositional Data via Variational EM** Nathan Osborne*, Rice University; Christine B. Peterson, MD Anderson Cancer Center; Marina Vannucci, Rice University
11:45	**Personalized Integrated Network Estimation** Veera Baladandayuthapani*, University of Michigan; Min Jin Ha, University of Texas MD Anderson Cancer Center; Yang Ni, Texas A&M University; Francesco C. Stingo, University of Florence, Italy
12:10	**Floor Discussion**

121. HIGHLY EFFICIENT DESIGNS AND VALID ANALYSES FOR RESOURCE CONSTRAINED STUDIES

Sponsors: ENAR, ASA Biometrics Section, ASA Section on Statistics in Epidemiology
Organizer: Jonathan Schildcrout, Vanderbilt University Medical Center
Chair: Jonathan Schildcrout, Vanderbilt University Medical Center

10:30	**Semiparametric Generalized Linear Models for Analysis of Longitudinal Data with Biased Observation-Level Sampling** Paul Rathouz*, University of Texas, Austin
10:55	**Cluster-Based Outcome-Dependent Sampling in Resource-Limited Settings: Inference in Small-Samples** Sara M. Sauer*, Harvard T.H. Chan School of Public Health; Bethany Hedt-Gauthier, Harvard Medical School; Claudia Rivera-Rodriguez, University of Auckland; Sebastien Haneuse, Harvard T.H. Chan School of Public Health
11:20	**Optimal Designs of Two-Phase Studies** Ran Tao*, Vanderbilt University Medical Center; Donglin Zeng and Danyu Lin, University of North Carolina, Chapel Hill
11:45	**Predictive Case Control Designs for Modification Learning** Patrick James Heagerty*, University of Washington; Katherine Tan, Flatiron Health
12:10	**Floor Discussion**

122. STATISTICAL ANALYSIS OF TRACKING DATA FROM PERSONAL WEARABLE DEVICES

Sponsors: ENAR, ASA Section on Statistical Learning and Data Science
Organizer: Jonggyu Baek, University of Massachusetts Medical School
Chair: Peter X.K. Song, University of Michigan

10:30	**Smartphone-Based Estimation of Sleep** Ian J. Barnett* and Melissa Martin, University of Pennsylvania
11:00	**Quantifying Mortality Risks using Accelerometry Data Collected According to the Complex Survey Weighted Design** Ekaterina Smirnova*, Virginia Commonwealth University; Andrew Leroux, Johns Hopkins University; Lucia Tabacu, Old Dominion University; Ciprian Crainiceanu, Johns Hopkins University
11:30	**Circadian Rhythm for Physical Activity of Infants Under 1-year Old** Jiawei Bai*, Sara Benjamin-Neelon and Vadim Zipunnikov, Johns Hopkins University
12:00	**Discussant:** Ciprian Crainiceanu, Johns Hopkins University

SCIENTIFIC PROGRAM

(CONTINUED)

123. CONTRIBUTED PAPERS: META-ANALYSIS METHODS

Sponsor: ENAR
Chair: Jing Zhang, University of Maryland

10:30	**A Three-Groups Bayesian Approach for Identifying Genetic Modifiers from Disparate Data Sources, with Application to Parkinson's Disease** Daisy Philtron*, The Pennsylvania State University; Benjamin Shaby, Colorado State University; Vivian Cheng, The Pennsylvania State University
10:45	**Multi-Trait Analysis of Rare-Variant Association Summary Statistics using MTAR** Lan Luo*, University of Wisconsin, Madison; Judong Shen, Hong Zhang, Aparna Chhibber and Devan V. Mehrotra, Merck & Co., Inc.; Zheng-zheng Tang, Wisconsin Institute for Discovery at University of Wisconsin, Madison
11:00	**Empirical Bayes Approach to Integrate Multiple External Summary-Level Information into Current Study** Tian Gu*, Jeremy M.G. Taylor and Bhramar Mukherjee, University of Michigan
11:15	**Tradeoff between Fixed-Effect and Random-Effects Meta-Analyses** Yipeng Wang* and Lifeng Lin, Florida State University
11:30	**Bayesian Approach to Assessing Publication Bias with Controlled False Positive Rate in Meta-Analyses of Odds Ratios** Linyu Shi* and Lifeng Lin, Florida State University
11:45	**A Bayesian Hierarchical CACE Model Accounting for Incomplete Noncompliance Data in Meta-Analysis** Jincheng (Jeni) Zhou*, Amgen; James S. Hodges and Haitao Chu, University of Minnesota
12:00	**Meta-Analysis of Gene Set Coexpression** Haocan Song*, Vanderbilt University Medical Center; Yan Guo, University of New Mexico; Fei Ye, Vanderbilt University Medical Center

124. CONTRIBUTED PAPERS: LONGITUDINAL DATA ANALYSIS

Sponsor: ENAR
Chair: Erinn M. Hade, The Ohio State University

10:30	**Regression Analysis of Sparse Asynchronous Longitudinal Data with Informative Observation Times** Dayu Sun*, University of Missouri; Hui Zhao, Zhongnan University of Economics and Law; Jianguo Sun, University of Missouri
10:45	**Modeling Continuous Longitudinal Response Data using Ordinal Regression** Yuqi Tian* and Bryan E. Shepherd, Vanderbilt University; Chun Li, Case Western Reserve University; Jonathan S. Schildcrout, Vanderbilt University
11:00	**Novel Joint Models for Identifying Determinants of Cognitive Decline in the Presence of Informative Drop-out and Observation Times** Kendra Davis-Plourde* and Yorghos Tripodis, Boston University
11:15	**Multiple Imputation of an Expensive Covariate in Outcome Dependent Sampling Designs for Longitudinal Data** Chiara Di Gravio*, Ran Tao and Jonathan S. Schildcrout, Vanderbilt University
11:30	**Real-Time Regression Analysis of Streaming Clustered Data with Possible Abnormal Data Batches** Lan Luo* and Peter X.K. Song, University of Michigan
11:45	**Modeling Disease Progression with Time-Dependent Risk Factors and Time-Varying Effects using Longitudinal Data** Jacquelyn E. Neal* and Dandan Liu, Vanderbilt University
12:00	**Informative Visit Processes in Longitudinal Data from the Health Sciences** Fridtjof Thomas*, University of Tennessee Health Science Center; Csaba P. Kovesdy, Memphis VA Medical Center; Yunusa Olufadi, University of Memphis

SCIENTIFIC PROGRAM
(CONTINUED)

125. CONTRIBUTED PAPERS: HIGH DIMENSIONAL DATA ANALYSIS: THE BIG PICTURE

Sponsor: ENAR
Chair: Kaushik Ghosh, University of Nevada, Las Vegas

10:30	**Capturing Skewness and Sparsity in High Dimensions** Xiaoqiang Wu*, Yiyuan She and Debajyoti Sinha, Florida State University
10:45	**Efficient Greedy Search for High-Dimensional Linear Discriminant Analysis** Hannan Yang* and Quefeng Li, University of North Carolina, Chapel Hill
11:00	**Parallelized Large-Scale Estimation and Inference for High-Dimensional Clustered Data with Binary Outcomes** Wenbo Wu*, Kevin He and Jian Kang, University of Michigan School of Public Health
11:15	**A Generalized Framework for High-Dimensional Inference based on Leave-One-Covariate-Out LASSO Path** Xiangyang Cao*, Karl Gregory and Dewei Wang, University of South Carolina
11:30	**Iterative Algorithm to Select Vine Copula According to Expert Knowledge and Pairwise Correlations** Philippe Saint Pierre*, University of Toulouse; Nazih Benoumechiara, Sorbonnes University; Nicolas J. Savy, University of Toulouse
11:45	**Floor Discussion**

126. CONTRIBUTED PAPERS: CLINICAL 'TRIALS AND TRIBULATIONS'

Sponsor: ENAR
Chair: Rachel Nethery, Harvard T.H. Chan School of Public Health

10:30	**Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Adjustment for Additional Baseline Variables** Bingkai Wang*, Michael Rosenblum, Ryoko Susukida, Ramin Mojtabai and Masoumeh Aminesmaeili, Johns Hopkins University
10:45	**Dynamic Borrowing in the Presence of Treatment Effect Heterogeneity** Ales Kotalik* and David Vock, University of Minnesota; Eric Donny, Wake Forest School of Medicine; Dorothy Hatsukami and Joseph Koopmeiners, University of Minnesota
11:00	**Bayesian Methods to Compare Dose Levels to Placebo in a Small n Sequential Multiple Assignment Randomized Trial (snSMART)** Kimberly A. Hochstedler* and Fang Fang, University of Michigan; Roy N. Tamura, University of South Florida; Thomas M. Braun and Kelley M. Kidwell, University of Michigan
11:15	**Sample Size Calculation in Comparative Clinical Trials with Longitudinal Count Data: Incorporation of Misspecification of the Variance Function and Correlation Matrix** Masataka Igeta*, Hyogo College of Medicine; Shigeyuki Matsui, Nagoya University Graduate School of Medicine
11:30	**Sequential Interval Estimation of Patient Accrual Rate in Clinical Trials** Dongyun Kim*, National Heart Lung and Blood Institute, National Institutes of Health; Sung-Min Han, OSEHRA
11:45	**Statistical Analysis of Glucose Variability** Jiangtao Luo*, Ismail El Moudden and Mohan Pant, Eastern Virginia Medical School
12:00	**The Impact of Precision on Go/No-Go Decision in Proof-of-Concept Trials** Macaulay Okwuokenye*, Brio Dexteri Pharmaceutical Consultants

SCIENTIFIC PROGRAM

(CONTINUED)

127. CONTRIBUTED PAPERS: COUNT DATA: THE THOUGHT THAT COUNTS

Sponsor: ENAR
Chair: Sandra Hurtado Rua, Cleveland State University

10:30	**Probabilistic Canonical Correlation Analysis for Sparse Count Data** Lin Qiu* and Vernon M. Chinchilli, The Pennsylvania State University
10:45	**Bayesian Credible Subgroups for Count Regression and Its Application to Safety Evaluation in Clinical Studies** Duy Ngo*, Western Michigan University; Patrick Schnell, The Ohio State University; Shahrul Mt-Isa, MSD Research Laboratories; Jie Chen and Greg Ball, Merck & Co., Inc.; Dai Feng, AbbVie Inc.; Richard Baumgartner, Merck & Co., Inc.
11:00	**Analysis of Panel Count Data with Time-Dependent Coefficient and Covariate Effects** Yuanyuan Guo* and Jianguo Sun, University of Missouri, Columbia
11:15	**Semi-Parametric Generalized Linear Model for Binary Count Data with Varying Cluster Sizes** Xinran Qi* and Aniko Szabo, Medical College of Wisconsin
11:30	**Drug Safety Evaluation Using Panel Count Model** Yizhao Zhou*, Ao Yuan and Ming Tan, Georgetown University
11:45	**Measurement Error Modeling for Count Data** Cornelis J. Potgieter*, Texas Christian University
12:00	**Conditional Mutual Information Estimation for Discrete and Continuous Data with Nearest Neighbors** Octavio César Mesner* and Cosma Rohilla Shalizi, Carnegie Mellon University

Denotes Student Award Winner

SHORT COURSES

	Short Course Registration Fees					
	By January 15			After January 15		
	Half Day	Second Half Day	Full Day	Half Day	Second Half Day	Full Day
Member	$250	$200	$350	$275	$225	$375
Non-Member	$325	$290	$425	$350	$315	$450

Sunday, March 22, 2020

SC 1. Implementing Bayesian Adaptive Designs: From Theory to Practice

Full Day | 8:00 am – 5:00 pm

Ying Yuan, University of Texas MD Anderson Cancer Center

J. Jack Lee, University of Texas MD Anderson Cancer Center

Description: As a statistical framework, a Bayesian approach is intuitive, logical, coherent, elegant, and adaptive in nature. It is uniquely suitable for the design and analysis of clinical trials. The learning curve of Bayesian methods, however, is steep and the complexity of Bayesian computation can be intimidating. To overcome these hurdles, this short course is designed to provide an overview of Bayesian theory and its application to adaptive clinical trials. The emphasis is on implementing such designs by turning theory into practice. Easy-to-use Shiny applications and downloadable standalone programs will be introduced to facilitate the study design, conduct, and analysis of Bayesian adaptive methods. The main application areas include adaptive dose finding, adaptive toxicity and efficacy evaluation, posterior probability and predictive probability for interim monitoring of study endpoints, outcome-adaptive randomization, hierarchical models, adaptive biomarker identification and validation, multi-arm, multi-stage designs, and platform designs, etc. Bayesian adaptive designs allow flexibility in clinical trial conduct, increase study efficiency, enhance clinical trial ethics by treating more patients with more effective treatments, increase the overall success rate for drug development and can still preserve frequentist operating characteristics by controlling type I and type II error rates. Lessons learned from real trial examples and practical considerations for conducting adaptive designs and will be given.

SC 2. Practical solutions for working with electronic health records data

Full Day | 8:00 am – 5:00 pm

Rebecca Hubbard, University of Pennsylvania

Description: The widespread adoption of electronic health records (EHR) as a means of documenting medical care has created a vast resource for the study of health conditions, interventions, and outcomes in the general population. Using EHR data for research facilitates the efficient creation of large research databases, execution of pragmatic clinical trials, and study of rare diseases. Despite these advantages, there are many challenges for research conducted using EHR data. To make valid inference, statisticians must be aware of data generation, capture, and availability issues and utilize appropriate study designs and statistical analysis methods to account for these issues.

This short course will introduce participants to the basic structure of EHR data and analytic approaches to working with these data through a combination of lecture and hands-on exercises in R. The first part of the course will cover issues related to the structure and quality of EHR data, including data types and methods for extracting variables of interest; sources of missing data; error in covariates and outcomes extracted from EHR data; and data capture considerations such as informative visit processes and medical records coding procedures. Participants will have the opportunity to explore a synthetic EHR-derived data set to gain familiarity with the structure of EHR data and data exploration and visualization tools for identifying data quality issues. In the second half of the course, we will discuss statistical methods to mitigate some of the data quality issues arising in EHR, including missing data and error in EHR-derived covariates and outcomes. R code will be provided for implementation of the presented methods, and hands-on exercises will be used to compare results of alternative approaches.

This short course is of interest to researchers without prior experience working with EHR data as well as more experienced individuals interested in learning practical solutions to some common analytic challenges. The overarching objective of this course is to provide participants with an introduction to the structure and content of EHR data as well as a set of practical tools to investigate and analyze this rich data resource.

SC 3. Design and Analysis of Sequential, Multiple Assignment, Randomized Trials for small and large samples

Full Day | 8:00 am – 5:00 pm

Kelley Kidwell, University of Michigan

Thomas Braun, University of Michigan

Roy Tamura, University of South Florida

Description: Sequential, multiple assignment, randomized trials (SMARTs) have been implemented in oncology, drug abuse, ADHD, obesity, depression, insomnia, autism, and smoking cessation, among other areas. A SMART is a multi-stage trial design that allows for individuals to be randomized at two or more stages based on intermediate outcomes. SMART design has primarily been focused on informing the construction of dynamic treatment regimens (DTRs) or adaptive interventions. DTRs are evidence-based treatment guidelines where treatment can be altered over time based on the individual. Most SMARTs are conducted in large samples and analyzed using frequentist methods to explore potential delayed effects and treatment interactions over time to estimate and compare DTRs. More recently, Bayesian and frequentist methods have been developed to apply the SMART design in rare diseases, or more generally, small samples to find the best overall treatment sharing information across stages. Thus, a SMART design can also be used to strengthen inference on the best single treatment. The Bayesian methods developed to analyze SMART data in small samples may also be extended to find the most effective DTRs. This short course will introduce SMART design for both large and small samples. Case studies will be used as examples and R code will be provided for practice.

SC 4. Programming with hierarchical statistical models: Using the BUGS-compatible NIMBLE system for MCMC and more

Half Day | 8:00 am – 12:00 pm

Christopher Paciorek, University of California, Berkeley

Description: NIMBLE (r-nimble.org) is a system for fitting and programming with hierarchical models in R that builds on the BUGS language for declaring models. NIMBLE provides analysts with a flexible system for using MCMC, sequential Monte Carlo, MCEM, and other techniques on user-specified models. It provides developers and methodologists with the ability to write algorithms in an R-like syntax that can be easily disseminated to users. C++ versions of models and algorithms are created for speed, but these are manipulated from R without any need for analysts or algorithm developers to program in C++. While analysts can use NIMBLE as a drop-in replacement for WinBUGS or JAGS, NIMBLE provides greatly enhanced functionality in a number of ways.

This hands-on tutorial will first show how to specify a hierarchical statistical model using BUGS syntax and fit that model using MCMC. Participants will learn how to customize the MCMC for better performance (choosing samplers and blocking schemes) and how to specify one's own statistical distributions and functions to extend the syntax of BUGS. We will demonstrate the use of NIMBLE for biostatistical methods such as semiparametric random effects models and clustering models using Bayesian nonparametric techniques. We will also demonstrate the use of NIMBLE's built-in reversible jump MCMC for variable selection and the use of NIMBLE's CAR-based spatial models.

SC 5. Multivariate meta-analysis methods

Half Day | 1:00 pm – 5:00 pm

Haitao Chu, University of Minnesota Twin Cities

Yong Chen, University of Pennsylvania

Description: Comparative effectiveness research aims to inform health care decisions concerning the benefits and risks of different prevention strategies, diagnostic instruments and treatment options. A meta-analysis is a statistical method that combines results of multiple independent studies to improve statistical power and to reduce certain biases compared to individual studies. Meta-analysis also has the capacity to contrast results from different studies and identify patterns and sources of disagreement among those results. The increasing number of prevention strategies, assessment instruments and treatment options for a given disease condition, as well as the rapid escalation in costs, have generated a need to simultaneously compare multiple options in clinical practice using innovative and rigorous multivariate meta-analysis methods.

This short course, co-taught by Drs. Chu and Chen who have collaborated on this topic for more than a decade, will focus on most recent developments for multivariate meta-analysis methods. This short course will offer a comprehensive overview of new approaches, modeling, and applications on multivariate meta-analysis. Specifically, this short course will discuss the contrast-based and the arm-based network meta-analysis methods for multiple treatment comparisons; network meta-analysis methods for multiple diagnostic tests; multivariate extension of network meta-analysis; and multivariate meta-analysis methods estimating complier average causal effect in randomized clinical trials with noncompliance.

Case studies will be used to illustrate the principles and statistical methods introduced in this course. R codes with real examples will also be provided. This application oriented short course should be of interest to researchers who would apply up-to-date multivariate meta-analysis methods and who are interested in developing novel methods for multivariate meta-analysis. We anticipate that it will be well-received by an interdisciplinary scientific community, and play an important role in improving the rigor and broadening the applications of multivariate meta-analysis.

SC 6. Statistical Network Analysis with Applications to Biology

Half Day | 8:00 am – 12:00 pm

Ali Shojaie, University of Washington

George Michailidis, University of Florida

Description: Networks and network analysis methods are increasingly used by biomedical scientists and computational biologists to glean insight into cellular functions and mechanisms of disease propagation and initiation. While many approaches have been recently proposed, statistical and machine learning tools commonly play a key role in such analyses. This course provides a practical introduction to statistical network analysis methods for biological application. This short course will cover the following classes of methods: (i) statistical methods for network-structured data analysis; (ii) inference methods for undirected networks. The course will primarily focus on methods that are widely used in biological applications and, in particular, in the analysis of -omics data, as well as recent developments in statistical machine learning. Throughout, the emphasis will be on practical applications of network analysis methods, as well as their limitations, including validation of results and tools for reproducible research. Case studies using publicly available -omics data will be used to describe various statistical network analysis methods.

SC 7. Trial Design and Analysis Using Multisource Exchangeability Models

Half Day | 1:00 pm – 5:00 pm

Joseph Koopmeiners, University of Minnesota

Brian Hobbs, Cleveland Clinic

Alex Kaizer, University of Colorado

Description: Modern biomedical applications often call statisticians to estimate the effect of a treatment or intervention in sub-groups defined by demographic, genetic, or other participant information. This results in increasingly smaller sample sizes, which reduces power. Hierarchical modeling allows sub-group specific effects to be "shrunk" together, thus borrowing strength and increasing precision. However, standard hierarchical approaches are limited because they lack the flexibility to model complex relationships between sub-groups, where some sub-groups are exchangeable, while others are not. In this short course, we discuss trial design using multi-source exchangeability models (MEMs), which provide a flexible approach to estimating sub-group-specific effects, while accounting for complex relationships between subgroups. We provide an overview of the methodology and a comparison with standard hierarchical modeling approaches. We then discuss multi-source modeling in the context of trial design, focusing specifically on platform and basket trial designs, illustrating the advantage of multi-source trial designs vs. standard designs. The ability to incorporate other adaptive elements, such as adaptive randomization, will also be discussed. Much of the course will be illustrated via the basket package in R.

TUTORIALS

	Tutorial Registration Fees	
	By January 15	After January 15
Member	$75	$85
Non-Member	$85	$95
Student	$40	$50

Monday, March 23 - Tuesday, March 24, 2020

T 1. Statistical methods for geometric functional data

Monday, March 23 | 8:30 am – 10:15 am

Karthik Bharath, University of Nottingham, UK

Sebastian Kurtek, The Ohio State University

Description: How can one quantify variation in Hippocampal shapes obtained from MRI images as 2D curves? How does one model intra-tumour heterogeneity using samples of pixel densities? Answers to such questions on functional data with rich geometric structure require methods that are at a nascent developmental stage, and are typically not part of the standard functional data toolbox.

In this tutorial, we shall introduce some modern statistical and computational tools for handling such functional data objects. The first part of the tutorial will focus on the representation of such data and computation of descriptive summaries such averages and PCA, with numerous references to existing works and computing resources. The focus then moves to understanding the challenges involved in developing regression models involving such data objects. The last part of the tutorial will present an overview of the current state-of-the-art, and suggest future directions of research with a view towards inference.

T 2. Disease Risk Modeling and Visualization using R

Monday, March 23 | 10:30 am - 12:15 pm

Paula Moraga, University of Bath, UK

Description: Disease risk models are essential to inform public health and policy. These models can be used to quantify disease burden, understand geographic and temporal patterns, identify risk factors, and measure inequalities. In this tutorial we will learn how to estimate disease risk and quantify risk factors using areal and geostatistical data. We will also create interactive maps of disease risk and risk factors, and introduce presentation options such as interactive dashboards. We will work through two disease mapping examples using data of malaria in The Gambia and cancer in Pennsylvania, USA. We will cover the following topics:

- Model disease risk in different settings,
- Manipulate and transform point, areal and raster data using spatial packages,
- Retrieve high resolution spatially referenced environmental data using the raster package,
- Fit and interpret spatial models using Integrated Nested Laplace Approximations (INLA) (http://www.r-inla.org/),
- Map disease risk and risk factors using leaflet (https://rstudio.github.io/leaflet/) and ggplot2 (https://ggplot2.tidyverse.org/),

The tutorial examples will focus on health applications, but the approaches covered are also applicable to other fields that use georeferenced data including epidemiology, ecology or demography. We will provide clear descriptions of the R code for data importing, manipulation, modeling and visualization, as well as the interpretation of the results. The tutorial materials are drawn from the book 'Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny' by Paula Moraga (2019, Chapman & Hall/CRC Biostatistics Series).

T 3. Integration of Genetics and Imaging Data in Scientific Studies

Monday, March 23 | 1:45 pm - 3:30 pm

Debashis Ghosh, Colorado School of Public Health

Description: In this tutorial, we will discuss issues and approaches in the consideration of combining genetics and imaging data in biological and biomedical studies. A variety of motivating examples will be described. A common life-cycle pipeline for analytics will be discussed, along with some emergent lessons that have been learned through the literature. I will also focus on the types of questions that typically asked with these data sources and the roles of regression modelling and machine learning in these contexts.

T 4. Causal Inference Using the R TWANG Package for Mediation and Continuous Exposures

Monday, March 23 | 3:45 pm - 5:30 pm

Donna Coffman, Temple University

Description: When randomized experiments are infeasible, analysts must rely on observational data in which treatment (or exposure) is not randomly assigned (e.g., in health policy research or when determining the effects of environmental exposures). In addition, knowing the mechanisms or pathways through which a treatment works requires causal inference methods because the mediator is not randomly assigned. This tutorial aims to promote the use of causal inference methods for mediation and continuous exposures using the R twang package. The twang package recently was expanded to handle mediation and continuous exposures. We will first introduce causal mediation using the potential outcomes framework and weighting methods for estimating the causal mediation effects. We then will illustrate the implementation of gradient (or generalized) boosting models (GBM) for estimating the weights using the R twang package. Next, we will introduce the generalized propensity score (GPS) for continuous exposures. We will illustrate the implementation of GBM for estimating the GPS using the R twang package. The tutorial will provide relevant statistical background knowledge of mediation, the GPS, GBM, and weighting but will focus on implementation rather than statistical theory. Attendees should have some familiarity with propensity score analysis (e.g., for binary treatments/ exposures) and regression models, but knowledge of causal mediation, GPS, and GBM is not necessary. Attendees will be provided with the R code.

T 5. Fundamentals of difference-in-differences studies

Tuesday, March 24 | 1:45 pm - 3:30 pm

Laura A. Hatfield, Harvard Medical School

Bret Zeldow, Harvard Medical School

Description: A popular design in policy and economics research, difference-in-differences contrasts a treated group's pre- to post-intervention change in outcomes to an untreated comparison group's change in outcomes over the same period. The difference between the changes in the treated and comparison groups may be interpreted as the causal effect of the intervention if one assumes that the comparison group's change is a good proxy for the treated group's counterfactual change if it had not been treated. In this tutorial, we review the fundamentals of difference-in-differences studies, including key causal assumptions and ways to assess their plausibility, selection of a good comparison group, matching and regression techniques, statistical inference, and robustness checks.

T 6. R package development

Tuesday, March 24 | 3:45 pm - 5:30 pm

John Muschelli, Johns Hopkins University

Description: The jump from R programming with scripts to packages can be quite large. We hope to answer some of the basic questions of getting you started with package development answering the questions of: How do you create a basic R package? What are some R package best practices? How do I know if I can install this package? How do I depend on other packages? The tutorial will go through a simple 2-function package and describe resources to use after the course, including the R Package Development YouTube series: https://www.youtube.com/watch?v=79s3z0gIuFU&list=PLk3B5c8iCV-T4LM0mwEyWIunIunLyEjqM&index=1

ROUNDTABLES

Registration is required. Roundtable Registration Fee: $45
Monday, March 23 12:15-1:30 p.m.

R 1. Statistical positions in government

Paul Albert, National Cancer Institute

Description: The federal government provides exciting career opportunities for biostatisticians. There are positions ranging from mathematical statisticians, postdoctoral fellows, and tenure-track investigators. We will discuss these different types of positions, including the different types of work and the citizen requirements. We will discuss how to locate positions and the application/interview process. Focus will be on positions at the National Institutes of Health and the Food and Drug Administration where most government biostatisticians work.

R 2. How Can We Improve Biostatistical Reviewing for Medical Journals?

Cynthia Garvan, University of Florida

Description: The scientific community is justifiably concerned about both the rigor and reproducibility of medical research. From Statistics Done Wrong (Reinhart, 2015) to findings from a recent National Academies of Sciences, Engineering, and Medicine workshop convened to address questions about the reproducibility of scientific research, lack of statistics education has been identified as a major culprit in the generation of poor science. Beyond a lack of statistics education for researchers, a lack of education for biostatistical reviewers is problematic. In this roundtable we will discuss steps needed to improve this vital contribution of the biostatistician to advance medical research.

R 3. Early career mentoring: What do I do now?

Lance A. Waller, Emory University

Description: A career in the field of Biostatistics can be rewarding but also a challenge to navigate early in one's career. Some parts of the field seem to be changing quickly, others seem to stay the same. Departments and research groups grow and shrink, scientific (and funding!) priorities shift with new technology, new discoveries, and new approaches. In this roundtable, we will consider multiple issues involved with beginning a career in Biostatistics. We will discuss the different "currencies of success" associated with careers in academic, industry, and government organizations. We will discuss communication skills, funding strategies, collaboration skills, and opportunities to contribute to the field in multiple ways. Please feel free to bring questions (or send them to the facilitator beforehand) to allow the discussion to address your needs as well as these guidelines.

R 4. Publish or Perish in Biostatistics

Geert Molenberghs, Hasselt University and KU Leuven, Belgium

Description: Like statistics and biostatistics itself, publishing in biostatistics journals is in full transition: from paper to also electronic to electronic only; what about open access? What about reproducibility and, relatedly, scientific integrity?; all of this against the background of privacy protection. Do we publish in a journal owned by a commercial publisher, in a society-owned journal, in a cooperative journal, or perhaps in no journal at all? Do we prefer a statistics or a data science journal – or is this a false dichotomy? What is the relative status of theory, theorems, methodology, modeling, data analysis, and simulations? Apart from being an author, what are the relative advantages and drawbacks of acting as referee or Associate Editor? Should we give weight to impact factors or are they ignorable?

R 5. Understanding the NIH Grant Review Process

Scarlett L. Bellamy, Drexel University

Description: Have you ever wondered what it's like to be member of an NIH study section? Have you ever wondered about the review process for grants that you have submitted or plan to submit? In this roundtable we will discuss the NIH review process, from the perspective of a current member of Biostatistical Methods and Research Design (BMRD) Study Section. Attendees should leave the discussion with a better understanding the grant review process to better inform how they might prepare future grants or as they consider service on future study sections.

R 6. Data Science Programs

Joel Greenhouse, Carnegie Mellon University

Description: Academic and online data science programs are popping up everywhere. Employers now post positions for data scientist and rarely for statisticians or data analyst. If statistical thinking is the bedrock of data science, how can we insure that statistics and good statistical thinking play a proper role in the training of the next generation of statistical scientist? What has your experience been with the emergence of data science at your University or your place of employment. These, as well as other participant generated questions will be the source of discussion for this roundtable.

R 7. Being a Biostatistician in a Medical Center

Bryan E. Shepherd, Vanderbilt University Medical Center

Description: Statisticians are in great demand in medical centers. This can be both exciting and daunting. We will discuss strategies for flourishing in a medical center, from gaining respect among medical collaborators, to identifying and pursuing interesting research projects, to protecting one's time.

R 8. How to navigate collaborative research

Andrea B. Troxel, NYU School of Medicine

Description: We will discuss best practices for working with collaborators to develop grant proposals, guidelines for effort allocation for both faculty and staff, and timelines for grant preparation. We will also discuss common roadblocks that arise, and offer tips for troubleshooting challenging situations.

R 9. Running a Statistical Consulting Business

Alicia Y. Toledano, Biostatistics Consulting, LLC

Description: Running your own consultancy has many benefits, such as choosing your clients and projects, setting your own hours, and possibly working from home. This roundtable will focus on meeting challenges and carrying out responsibilities associated with those benefits. We will discuss making decisions related to: incorporation, using an attorney to review contracts, accounting, insurance, SOPs including for quality control, and having subcontractors and/or employees. Based on time and attendees' interests, we may also discuss one or more of: 1) Deciding what projects to undertake, with respect to areas of statistical expertise and 2) project type, such as short- or long-term; papers, grants, and/or FDA submissions; 3) How to get clients; 4) Working with clients that are not local; and 5) Ensuring your continued professional development statistically, and in soft skills like working as part of an interdisciplinary team. Come with questions and/or suggestions!

WORKSHOP & STUDENT OPPORTUNITIES

Special Opportunities for Our Student Members

PARTICIPATE IN STUDENT-FOCUSED ELEMENTS OF THE SCIENTIFIC PROGRAM:

The Sunday night mixer presents an ideal opportunity to network and hear about emerging research at the annual ENAR Poster session. This year we will conduct our fifth Poster Competition for the session. Prizes will be announced within topical areas in the Tuesday morning Presidential Invited Address session. A student winner will be selected within each topical area. Watch for details on entering the competition on the website when the meeting registration goes live.

EDUCATIONAL AND PROFESSIONAL DEVELOPMENT OPPORTUNITIES:

Be sure to take advantage of the educational offerings to be held during the meeting – short courses, tutorials, and roundtable discussions (see pages 72-73).

Don't Forget the Popular ENAR Career Placement Services!

(See page 82.)

NETWORK WITH YOUR FELLOW STUDENTS

Back by popular demand, the **Council for Emerging and New Statisticians (CENS) Mixer** will be held the evening of Monday, March 23, 2020. This is a great way to meet and greet your fellow students from other graduate programs. Don't miss this opportunity to begin building connections with your future colleagues and friends.

Looking for more ways to plug in with other students? Check out additional CENS-sponsored activities on page 80.

CENS

Council for Emerging and New Statisticians

CENS EVENTS AT ENAR 2020

CENS Sponsored Session: The Three M's: Meetings, Memberships, and Money!

This panel will educate emerging and new statisticians on how to gain more from professional meetings and associations. Topics for discussion will include the benefits of joining a professional organization, means of navigating scientific sessions at a conference, developing a professional network, and obtaining funding (e.g., travel grants/awards, scholarships).

Networking Mixer: Monday, March 23, 2020 from 5:30 - 6:30pm

All students and recent graduates are invited to attend the CENS Networking Mixer. Registration is not required - so please plan to attend!

Networking Lunch: Tuesday, March 24, 2020 from 12:30 - 1:30pm at local restaurants

CENS will organize lunches for groups of attendees that share similar interests. The goal is to help attendees meet and network with each other. Although CENS will help to coordinate lunch at local restaurants, please note that lunch is at your own expense and CENS will not be able to cater to special dietary requirements. Closer to the meeting time, CENS will email all attendees interested in this networking event to request information to set up the groups and the lunch reservations. Participants meet at the CENS table in the Exhibition Hall at 12:15 PM before walking with their assigned group to a nearby restaurant for networking and lunch! Participation is open to all meeting attendees. If you would like to participate, please select the CENS lunch option on the registration form or email CENS at **enar.cens@gmail.com**.

About CENS

CENS was formed in 2012 by ENAR's Regional Advisory Board (RAB) to help ENAR better address the needs of students and recent graduates. CENS is composed of 10 graduate students, post-doctoral fellows, or recent graduates, who are ENAR members. With the help of the RAB Liaison, CENS members collaborate to bring student/recent graduate concerns to the attention of RAB and ENAR; work to help ENAR better serve all students/recent graduates; advise and help implement ideas to enhance the benefits of ENAR membership and to increase awareness of the benefits of ENAR membership to students; organize a CENS sponsored session at each ENAR Spring Meeting; assist in planning events that help advance students' and recent graduates' education and careers; and contribute to the development of ENAR's social media presence.

Join CENS

We are actively recruiting new members! Each member is appointed to a 2-year term. Within CENS, three or four people are chosen to participate in the steering committee, which reports to the RAB chair. Members of the steering committee will serve an additional year on CENS. CENS members meet in person yearly at the ENAR Spring Meeting and participate in conference calls throughout the year to plan events and address issues as they arise. If you are interested in joining CENS, please email **enar.cens@gmail.com**.

CENS seeks to advocate for the needs and concerns of students and recent graduates in collaboration with ENAR's Regional Advisory Board. Through annual events at the ENAR Spring Meeting, CENS strives to promote the benefits of participating in the ENAR community, support the advancement of students and recent graduates, and facilitate stronger connections within the statistical community.

CAREER PLACEMENT SERVICES

Hours of Operation:

Sunday, March 22	**4:00 pm – 6:30 pm**
Monday, March 23	**9:30 am – 4:30 pm**
Tuesday, March 24	**9:30 am – 3:30 pm**

General Information

The ENAR Career Placement Service helps match applicants seeking employment and employers. The service includes online registration and electronic uploading and distribution of applicant and employer materials through a password-protected online web-based facility. Visit the ENAR website at https://enar.org/meetings/spring2020/career/ to register for the placement center.

Job announcements and applicant information can be readily accessed electronically, applicant information will be opened prior to the meeting, and materials will remain available online after the meeting. ENAR provides separate large reading/planning rooms for employers and applicants to review materials, dedicated placement center personnel onsite, and optional private interview rooms available for employers. Employer and applicant reading/planning rooms are equipped with a small number of computers with internet connections, and printers. However, to make the most efficient use of the Placement Center, we recommend that participants register listings in advance of the meeting to maximize visibility, explore the database before the meeting, and, if attending, have a laptop computer on-site.

Employers

Each year numerous qualified applicants, many approaching graduation, look to the ENAR Placement Center to begin or further their careers. Organizations including government agencies, academic institutions, and private pharmaceutical firms all utilize the ENAR Career Placement Service. ENAR recognizes the value the Career Placement Service provides to members and, to make it more efficient and effective for both employers and applicants, uses an electronic registration process and an online database of applicant resumes. All registered employers will receive full access to the placement center for up to 3 company representatives, up to 4 job postings, pre-meeting access to the online applicant database of resumes, full conference registration for up to 3 representatives, and access to the employer placement center room. ENAR is also offering those organizations seeking private interview space the option to reserve a private room for interviews in 4-hour increments.

Employer Registration

The registration fee for employers includes full access for up to four position postings and up to 3 representatives, pre-meeting access to the online applicant database of resumes, up to 3 full conference registrations, and access to the employer placement center room.

Employer Resource Area: ENAR will provide internet access, laptops, and printers available in the employer resource room for viewing the applicant resume database. However, for most efficient use of the resource room, we recommend employers have on-site access to a personal laptop computer.

Interview Suites

For an additional fee, employers may reserve private interview suites each day on a first-come, first-served basis. There are a very limited number of private suites, so please reserve early.

Employer Registration Instructions, Deadlines, and Fees

ALL employers must FULLY complete an online Employer Form located at: https://enar.org/meetings/spring2020/career/ for each position listing. Attachments may be included.

Employer Registration Fees	By Jan. 15	After Jan. 15
Employer (3 reps/ 4 job postings)	$1,650	$1,725
Private Interview Room (Per 4-hour increments)	$275	n/a
Additional Representatives (Cost per person includes conference registration)	$520	$620
Additional Job Postings	$150	$250

Applicants

If you have an interest in a career in biometrics, you can utilize the ENAR Career Placement Center to get started or get ahead. Many employers attend the ENAR Spring Meeting each year seeking qualified applicants. All registered applicants may register for up to three job classification types and receive full access to the placement center applicant room and the online employer job posting database. Please note that to fully utilize the online database, we recommend applicants register in advance to maximize visibility, explore the database shortly before the meeting and, if attending, have a laptop computer on-site.

Applicant Registration

The ENAR Career Placement Center provides opportunities for qualified applicants to meet employers and learn about organizations employing biostatisticians.

Visibility to Employers

The Online Applicant database is made available to all employers prior to the opening of the placement center.

Applicant Resource Area

ENAR will have internet access, three laptops, and printers in the applicant room for viewing the employer job posting database. However, for most efficient use we recommend applicants have on-site access to a personal laptop computer.

Applicant Registration Instructions, Deadlines, and Fees

ALL applicants must FULLY complete an online Applicant Form located at: https://enar.org/meetings/spring2020/career/ for each job classification.

Applicant Registration Fees	By Jan. 15	After Jan. 15
Regular Registration	$60	$85
Student Registration	$25	$40

Applicants PLEASE NOTE: If you are planning to interview and participate on-site you must also register for the conference and pay the meeting registration fee.

FAMILY FRIENDLY ACCOMMODATIONS

ENAR maintains a family-friendly environment. However, to help with logistics/planning, all family members/guests, who wish to enter any of the meeting space, must be formally registered. Children 12 years and under may be registered for free. All other adults and children 13+ years will require a $100 guest registration (if registering prior to January 15) to attend any ENAR Spring Meeting event. If registering after January 15, the guest registration fee is $110.

Guest registration includes access to all Scientific Program sessions, exhibit space, Opening Mixer & Poster Session, refreshment break with exhibitors and the Presidential Invited Address. Guest registration does not include admission to Short Courses, Tutorials or Roundtables, or any invite only/user-pay events.

Child Care

Attendees with child-care needs may contact Sitter Scout (https://www.sitter-scout.com/) for arrangements during the ENAR Spring Meeting. Please contact Jaclyn at 860-508-766 or Cori at 802-540-0433 for arrangements.

ENAR assumes no responsibility for any child-care services and all policies are established by the child-care facility.

New Mothers Room

If you have a child and want a private space for nursing or other infant care, please visit the ENAR registration desk.

NOTES

ABSTRACTS & POSTER PRESENTATIONS

1.POSTERS: IMAGING DATA ANALYSIS

1a. Time Varying Estimation of Tensor-on-Tensor Regression with Application in fMRI Data

Pratim Guha Niyogi*, Michigan State University
Tapabrata Maiti, Michigan State University

In fMRI studies, data structure could be visualized as time-varying multidimensional arrays (tensor), collected at different time-points on multiple objects. We consider detection of neural activation in fMRI experiments in presence of tensor valued brain images and tensor predictors where both of them are collected over same set of time. A time-varying regression model with the presence of inherent structural composition of regressors and covariates is proposed. B-spline technique is used to model the regression coefficient and the coefficients of the basis function are estimated using CP decomposition based on Lock (2018) algorithm by minimizing a penalized loss function. In this article, we have generalized the varying coefficient model from vector-valued covariates and responses, as well as, the tensor regression model. Hence, it is a logical and nontrivial extension of function-on-function concurrent linear models in complex data structure where the inherent structures of the data is considered. Efficacy of the proposed model is studied based on both simulated data for different setups and a real data set.

email: guhaniyo@msu.edu

1b. Estimation of Fiber Orientation Distribution through Blockwise Adaptive Thresholding

Seungyong Hwang*, University of California, Davis
Thomas Lee, University of California, Davis
Debashis Paul, University of California, Davis
Jie Peng, University of California, Davis

Due to recent technological advances, large brain imaging data sets can now be collected. Such data are highly complex so extraction of meaningful information from them remains challenging. Thus, there is an urgent need for statistical procedures that are computationally scalable and can provide accurate estimates that capture the neuronal structures and their functionalities. We propose a fast method for estimating the fiber orientation distribution (FOD) based on diffusion MRI data. This method models the observed dMRI signal at any voxel as a convolved and noisy version of the underlying FOD, and utilizes the spherical harmonics basis for representing the FOD, where the spherical harmonic coefficients are adaptively and nonlinearly shrunk by using a James-Stein type estimator. To further improve the estimation accuracy by enhancing the localized peaks of the FOD, as a second step a super-resolution sharpening process is then applied. The resulting estimated FODs can be fed to a fiber tracking algorithm to reconstruct the white matter fiber tracts. We illustrate the overall methodology using both synthetic data and data from the Human Connectome Project.

email: syhwang@ucdavis.edu

1c. Estimating Dynamic Connectivity Correlates of PTSD Resilience Using MultiModal Imaging

Jin Ming*, Emory University
Suprateek Kundu, Emory University
Jennifer Stevens, Emory University

Recently, the potential of dynamic brain networks as neuroimaging biomarkers for mental illnesses is being increasingly recognized. We develop a novel approach for computing dynamic brain functional connectivity (FC), that is guided by brain structural connectivity (SC) computed from diffusion tensor imaging (DTI). The proposed approach involving dynamic Gaussian graphical models decomposes the time course into non-overlapping state phases determined by change points, each having a distinct network. We develop an optimization algorithm to implement the method such that the estimation of both the change points and the state-phase networks are fully data-driven and unsupervised, and guided by SC information. An application of the method to a posttraumatic stress disorder (PTSD) study reveals important dynamic resting-state connections in regions of the brain previously implicated in PTSD. We also illustrate that the dynamic networks computed under the proposed method can better predict psychological resilience among trauma-exposed individuals compared to existing dynamic and stationary connectivity approaches, which highlights its potential as a neuroimaging biomarker.

email: jming2@emory.edu

1d. Towards an Automatic Detection Method of Chronic Active Lesions

Carolyn Lou*, University of Pennsylvania
Jordan D. Dworkin, University of Pennsylvania
Alessandra Valcarcel, University of Pennsylvania
Martina Absinta, National Institute of Neurological Disorders and Stroke, National Institutes of Health
Pascal Sati, National Institute of Neurological Disorders and Stroke, National Institutes of Health
Kelly Clark, University of Pennsylvania
Daniel Reich, National Institute of Neurological Disorders and Stroke, National Institutes of Health

Recent developments in magnetic resonance imaging (MRI) have shown that chronic active multiple sclerosis lesions can be assessed in vivo. These lesions are typically characterized by a dark rim indicating increased iron-laden microglia/macrophages at their edge, and their presence is associated with worse disease outcomes. Here, we present two candidate methods for the automatic detection of these rims on 3T MRI images. Our first method conducts a radiomics image analysis directly on the images, and our second quantifies the covariance structure via inter-modal coupling analysis of multi-modal images. Both analyses show promising results for detecting iron rims in lesions.

email: louc@upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

1e. Bayesian Mixture Model for Lesion Detection and Clustering in MS

Jordan D. Dworkin*, University of Pennsylvania
Melissa L. Martin, University of Pennsylvania
Arman Oganisian, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania

Neuroimaging research in multiple sclerosis often centers around the detection and analysis of white-matter lesions. Despite the importance of lesions in MS, the total extent of lesion tissue in the brain (referred to as 'lesion load') is only modestly associated with clinical outcomes. More advanced analysis that investigates the number and individual characteristics of distinct lesions has been hampered by extreme spatial overlap of lesions that makes them hard to separate visually or computationally. Here, we propose a method for statistically estimating the number and location of distinct lesions in MS. This method models the lesion probability map as a mixture of Gaussian densities, and parameters for lesions' centers, spreads, and intensities are fit in a Bayesian manner. Following model fit, voxels can then be clustered probabilistically, allowing for analysis at the level of individual lesions. Using a longitudinal study of patients with MS, the validity and reliability of this method are quantified and compared to previously proposed techniques.

email: jdwor@pennmedicine.upenn.edu

1f. Seeing Very Small Things: Applications of Mixture Modeling and Extreme Value Distributions in Microscopic Image Analysis

Miranda L. Lynch*, Hauptman-Woodward Medical Research Institute
Sarah E.J. Bowman, Hauptman-Woodward Medical Research Institute

Image data from microscopy imaging modalities provide critical information about cellular processes, enable diagnostic evaluation at tissue-level resolution, and are used in a host of nanoscale physical processes. These imaging modalities are highly specialized and tailored to specific scientific applications, and differ with regards to form and frequency of energy, in how images are recorded, and in detector capabilities. Multiple optical pathologies can impact information content in microscopic images: various aberrations, limits of detection, or missing data. We focus on applying statistical modeling to problems of nanoparticle detection in nonlinear optical spectroscopy methods for protein crystallization screening trials. A challenging problem in this arena is characterizing sizes of smallest particles in the images. In our application area of nanocrystal detection, we make use of Generalized Extreme Value Distributions (GEVD) and mixture models to address questions about size distributions of crystals under varying experimental conditions, subject to zero inflation due to lack of crystal formation and/or detection limits, and under different imaging methods.

email: mlynch@hwi.buffalo.edu

2.POSTERS: SURVIVAL ANALYSIS/COMPETING RISKS

2a. Functional Additive Cox Model

Erjia Cui*, Johns Hopkins University
Andrew Leroux, Johns Hopkins University
Ciprian Crainiceanu, Johns Hopkins University

We propose a Functional Additive Cox Model to flexibly quantify the association between high dimensional functional predictors and time to event data. The model extends the linear functional proportional hazards model and introduces flexible transformations of the functional covariates to address problems of weak identifiability due to information sparsity. The model is implemented in R, is very fast, publicly available, and was applied to the National Health and Nutrition Examination Survey (NHANES). The results of this application are scientifically relevant as new and interpretable patterns of physical activity that affect time to death have been identified and described. We also describe an efficient and easy-to-use simulation framework for both the functional covariate and survival data.

email: ecui1@jhmi.edu

2b. Gene-Based Association Analysis of Survival Traits via Functional Regression based Mixed Effect Cox Models for Related Samples

Ruzong Fan*, Georgetown University Medical Center
Chi-yang Chiu, University of Tennessee Health Science Center
Bingsong Zhang, Georgetown University Medical Center
Shuqi Wang, Georgetown University Medical Center
Jingyi Shao, Georgetown University Medical Center
M'Hamed Lajmi Lakhal-Chaieb, Universite Laval
Richard J. Cook, University of Waterloo
Alexander F. Wilson, Computational and Statistical Genomic Branch of the National Human Genome Research Institute, National Institutes of Health
Joan E. Bailey-Wilson, Computational and Statistical Genomic Branch of the National Human Genome Research Institute, National Institutes of Health
Momiao Xiong, University of Texas Health Science Center at Houston

The importance to integrate survival analysis into genetics and genomics is widely recognized, but only a small number of statisticians have produced relevant work toward this research direction. For unrelated population data, functional regression models have been developed to test for association between a trait and genetic variants in a gene region. We extend this approach to analyze censored traits for family data or related samples using functional regression based mixed effect Cox models (FamCoxME). The FamCoxME model the effect of major gene as fixed mean via functional data analysis techniques, the local gene or polygene variations or both as random, and the correlation of pedigree members by kinship coefficients. The association between the censored trait and the major gene is tested by likelihood ratio tests (FamCoxME FR LRT). Simulation results indicate that the LRT control the type I error rates conservatively and have good power levels when both local gene or polygene variations are modeled. The proposed methods were applied to analyze a breast cancer data set from the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA).

email: rf740@georgetown.edu

ABSTRACTS & POSTER PRESENTATIONS

2c. Regression Model for the Lifetime Risk using Pseudo-Values

Sarah C. Conner*, Boston University School of Public Health
Ludovic Trinquart, Boston University School of Public Health

The lifetime risk measures the cumulative risk for developing a disease over the lifespan. Statistical methods for the lifetime risk must account for left truncation, semi-competing risk of death, and inference at a fixed timepoint. Covariates may be associated with the cumulative incidence function but not with the lifetime risk. We introduce a novel method to predict the lifetime risk based on individual characteristics. We define pseudo-values of the Aalen-Johansen estimator at a fixed time point, allowing delayed entry. We regress transformed pseudo-values on covariates with generalized estimating equations. We derive the difference in lifetime risk and 95% confidence intervals with the multivariate delta method. To compare its performance with the Fine-Gray model, we simulated semi-competing risks data with non-proportional subdistribution hazards. We measured bias, mean squared error, coverage, Type I error, and power. Our method performed well in all criteria in most scenarios, and demonstrated appropriate Type I error and high power. We illustrate our model with a prediction model for the lifetime risk of atrial fibrillation in the Framingham Heart Study.

email: sconner@bu.edu

2d. Proportional Subdistribution Hazards Model with Covariate-Adjusted Censoring Weight for Clustered Competing Risks Data

Manoj Khanal*, Medical College of Wisconsin
Soyoung Kim, Medical College of Wisconsin
Kwang Woo Ahn, Medical College of Wisconsin

A competing risk is an event which precludes the occurrence of main event of interest. Competing risks data often suffer from cluster effects such as center effect and matched pairs design. Zhou et. al (2012) proposed a competing risks regression for clustered data with the covariate-independent censoring assumption. They considered a Kaplan-Meier-estimator-based inverse probability of censoring weighting approach. However, the censoring distribution often depends on covariates in practice. In such cases, ignoring covariate-dependent censoring may lead to bias in parameter estimation. We propose an inverse probability of censoring weighted estimation method that fits a marginal Cox model to adjust for covariate-dependent censoring for clustered competing risks data. Our simulation results show that, in the presence of covariate-dependent censoring, the parameter estimates are unbiased with approximately 95% coverage rates.

email: mkhanal@mcw.edu

2e. A Unified Power Series Class of Cure Rate Survival Models for Spatially Clustered Data

Sandra Hurtado Rua*, Cleveland State University
Dipak Dey, University of Connecticut

We propose a unified power series transformation class of cure rate survival models for spatially clustered data. We take a Bayesian approach and our model includes as a special case the mixture and the promotion time cure models while accounting for spatial heterogeneity. The primary advantage of our model is the estimation of spatial frailties that identify regional characteristics or spatial disparities affecting the cure fraction. We apply our methodologies to breast cancer survival times from the state of Utah (1990-2017) extracted from the National Cancer Institute SEER database, yet the proposed work offers useful contributions to general time-to-event analysis and methodology. We compare a broad collection of high-dimensional hierarchical models by implementing widely used model selection criteria. In addition to the usual inference for cure rate survival model, we also obtain smooth maps of spatial frailties over space.

email: s.hurtadorua@csuohio.edu

2f. Optimizing Incremental Cost-Effective Ratios for Censored Survival Time and Cost

Xinyuan Dong*, University of Washington

Individualized treatment rules recommend treatments based on individual patient characteristics in order to maximize clinical outcome. In this paper, we aim at deriving individual treatment rules that can minimize incremental cost-effectiveness ratio (ICER), which is defined by the difference in cost between two treatment options, divided by the difference in their effect. Our goal is to optimize ICER, while setting minimum threshold on the improvement in benefit. We propose two algorithms (adaptive learning and non-adaptive learning) to optimize constrained optimization problem. Adaptive learning first determines the activeness of the constraint by evaluating the performance of unconstrained optimal solution, and automatically adjusts the parameter estimates if the unconstrained optimal solution does not satisfy the constraint requirement. Non-adaptive learning uses d.c (difference of convex) algorithm to directly solve a non-convex optimization problem by performing a sequence of convex problems until convergence.

email: xd23@uw.edu

2g. An EM Algorithm in Fitting the Generalized Odds-Rate Model to Right Censored Data

Ennan Gu*, University of South Carolina

The generalized odds-rate model which is a general class of semiparametric regression models including the proportional hazards model and proportional odds model as special cases has been studied and most of the existing approaches assume the transformation parameter to be known. In this paper, we propose a gamma-gamma data augmentation approach to estimate the transformation parameter together with other parameters. The proposed EM algorithm is robust to initial values and has variance estimation in closed form by Louis method. The performance of the proposed method is evaluated by comprehensive simulation studies and illustrated by a real data application.

email: egu@email.sc.edu

ABSTRACTS & POSTER PRESENTATIONS

3.POSTERS: MACHINE LEARNING AND HIGH-DIMENSIONAL DATA

3a. Distributed Quadratic Inference Functions for Integrating Studies with High-Dimensional Repeated Measures

Emily C. Hector*, University of Michigan
Peter X.-K. Song, University of Michigan

We consider integrating studies that collect the same high-dimensional repeated outcomes with different but overlapping sets of covariates. The primary goal of this integrative analysis is to estimate the effects of the overlapping covariates, while adjusting for study-specific covariates, through a marginal regression model. To estimate the effects of covariates of interest on the high-dimensional outcomes, we develop a divide-and-conquer procedure for statistical estimation and inference of regression parameters, which is implemented in a fully distributed and parallelized computational scheme. To overcome computational and modeling challenges arising from the high-dimensional likelihood of the repeated outcome, we propose to analyze small batches of data using Qu, Lindsay and Li (2000)'s Quadratic Inference Functions, and then to combine these estimators using a one-step meta-estimator in a similar spirit to Hansen (1982)'s Generalized Method of Moments. We show both theoretically and numerically that the proposed method is efficient and computationally fast. We develop an R package for ease of implementation.

email: ehector@umich.edu

3b. Statistical Inference for the Word2vec Natural Language Processing Algorithm Applied to Electronic Health Records

Brian L. Egleston*, Fox Chase Cancer Center
Stan Taylor, Fox Chase Cancer Center
Michael Lutz, Fox Chase Cancer Center
Richard J. Bleicher, Fox Chase Cancer Center
Slobodan Vucetic, Temple University

Word2vec is a natural language processing algorithm that has been widely cited, and has spurred further innovation in the computer and information sciences fields. Often word2vec is used to investigate which sets of words in a document might be written in neighborhoods of each other. We have investigated the probability theory underlying the method and its relationship to the pointwise mutual information statistic. The pointwise mutual information statistic provides a metric of closeness of words in a document. Word2vec can be computationally intensive. Examination of a mathematical proof gives insight into steps that can be taken to speed estimation of word2vec. We also describe how one can obtain standard errors for the resulting statistics. We use our methods to distinguish those with and without diabetes mellitus in electronic health record free text data using hundreds of thousands of clinical notes from an academic medical center.

email: brian.egleston@fccc.edu

3c. Neural Network Survival Model for Cardiovascular Disease Prediction

Yu Deng*, Northwestern University
Lei Liu, Washington University, St. Louis
HongMei Jiang, Northwestern University
Kho Abel, Northwestern University
Yishu Wei, Northwestern University
Norrina Allen, Northwestern University
John Wilkins, Northwestern University
Kiang Liu, Northwestern University
Donald Lloyd-Jones, Northwestern University
Lihui Zhao, Northwestern University

Multiple algorithms have been developed to predict the risk of cardiovascular disease (CVD), e.g., the Framingham risk score and the pooled cohort equation. However, most of these algorithms are based on traditional statistical model such as cox regression. Neural network models have shown great success in image classification and text classification in recent years, but its utility in clinical structured data are less known. In this study, we applied various state-of-the-art neural network-based survival models (deep-surv, cox-nnet, nnet) in CVD risk prediction. We compared the performance these models with pooled cohort equation. For comparison purpose, all the models have the same predictors as the pooled cohort equation. We then derived our own neural network-survival model based on the architecture of deep-surv. We fine-tuned several parameters including: number of epoch, batch size, number of hidden layers, regularization term and activation function. Our results in the testing set show that the existing neural network survival models have comparable AUC score compared to cox regression model, while our proposed model has slightly superior AUC score.

email: yudeng2015@u.northwestern.edu

3d. Applying Statistical Learning Algorithms on the Prediction of Response to Immune Checkpoint Blockade Therapy

Tiantian Zeng*, University of Kentucky
Chi Wang, University of Kentucky

Immune checkpoint blockade (ICB) therapy could bring long-lasting clinical gains for the treatment of cancer. However, studies show that only a fraction of patients respond to the treatment. In this regard, the statistical modeling, which constructs classification algorithms to predict patients' response to the ICB therapy, could help explore the complexity of immune response. In this study, we used several published melanoma datasets with RNA-seq data and clinical response, and built prediction models using random forest and Lasso methods. We found that the specific pairwise relations of the expression of immune checkpoint genes performed the best in predicting the treatment response. In addition, we compared the prediction performance using combined datasets versus each single dataset. Our finding demonstrated that the utilization of statistical modeling and data integration is of high value to identify ICB response biomarkers in future studies.

email: t.zeng@uky.edu

ABSTRACTS & POSTER PRESENTATIONS

3e. Integrative Biclustering for Characterization of Biomarker and Phenotype Associations

Weijie Zhang*, University of Minnesota

Integration of multiple data types allows for new discoveries by combining the strengths from different but related data. In biomedical research where multiple high-dimensional omics data are available, a particular interest is to detect omics clusters or pathways that are associated with subgroups of a population. For instance, a crucial need in chronic obstructive pulmonary (COPD) research is to identify new COPD phenotypes and their associated clusters of biomarkers based on multiple omics data. Biclustering is a technique for identifying clusters of associated variables and observations simultaneously for one data type. Such methods have been successfully used to detect coregulated genes for different tumor types, but they are not directly applicable to more than one data type. In this paper, we propose an integrative bi-clustering method for multiple data types based on singular value decomposition. Our method requires no distributional assumption, performs dimension reduction with data specific sparsity, and results in interpretable row-column associations. An efficient algorithm is also proposed. We will assess our algorithm using synthetic and real data.

email: weijie25zh@gmail.com

3f. Testing Presence-Absence Association in the Microbiome Using LDM and PERMANOVA

Andrea N. Lane*, Emory University
Glen Satten, Centers for Disease Control and Prevention
Yijuan Hu, Emory University

The difference between two groups of microbial composition profiles can be characterized by the presence-absence status of certain microbes. But there is a lack of methods that provide a global test of the presence-absence difference and tests of individual operational taxonomic units (OTUs) while accounting for biases induced by variation in sampling depth (i.e. library size). PERMANOVA is a commonly used distance-based method for testing the global hypothesis of any microbiome effect. The linear decomposition model (LDM) includes the global test and tests of individual OTU effects. We propose to rarefy the OTU table so that all samples have the same depth and then apply the LDM to the rarefied table or PERMANOVA to a presence-absence distance based on the rarefied table. We repeat the process for a number of randomly rarefied tables and sum up the test statistics over rarefactions. Our simulations indicate that the proposed strategy is robust to all systematic differences in library size. We also explored the optimal number of rarefactions that balance statistical power and computational cost and provide practical guidelines on how to select the rarefaction depth.

email: andrea.nicole.lane@emory.edu

3g. Feature Selection for Support Vector Regression Using a Genetic Algorithm

Shannon B. McKearnan*, University of Minnesota
David M. Vock, University of Minnesota
Julian Wolfson, University of Minnesota

Support vector regression (SVR) is particularly beneficial when the outcome and predictors have a non-linear relationship. However, when many covariates are available, the method's flexibility can lead to overfitting and an overall loss in predictive accuracy. To overcome this, we develop a feature selection method for SVR based on a genetic algorithm that iteratively searches across potential subsets of variables to find those that yield the best performance according to a user-defined fitness function. We evaluate the performance of our feature selection method for SVR, comparing it to alternate methods including LASSO and random forest, in a simulation study. We find that our method yields higher predictive accuracy than SVR without feature selection. Our method outperforms LASSO when the relationship between covariates and outcome is non-linear. Random forest performs equivalently to our method in some scenarios, but more poorly in the case of correlated covariates. In addition, we apply our method to predict forced expiratory volume at one year after lung transplant using data from the United Network for Organ Sharing national registry.

email: mckea018@umn.edu

3h. Statistical Inferences for F1-scores in Multi-Class Classification Problems

Kouji Yamamoto*, Yokohama City University
Kanae Takahashi, Osaka City University
Aya Kuchiba, National Cancer Center, National Institutes of Health
Tatsuki Koyama, Vanderbilt University Medical Center

A binary classification problem is common in medical field, and we often use sensitivity, specificity, accuracy, negative and positive predictive values as measures of performance of a binary predictor. In computer science, a classifier is usually evaluated with precision (positive predictive value) and recall (sensitivity). As a single summary measure of a classifier's performance, F1-score defined as the harmonic mean of precision and recall, is widely used in the context of information retrieval and information extraction evaluation. This measure is rarely used in diagnostic studies in medicine; however, it possesses favorable characteristics, especially when the prevalence is low. Some statistical methods for inference have been developed for the F1-score in binary classification problems. In this presentation, we extend the problem to multi-class classification. In this context, there are two types of measures; macro-averaged F1-score and micro-averaged F1-score exist, and statistical properties of those F1-scores have hardly ever been discussed. We present methods for estimating F1-scores with confidence intervals.

email: yamamoto.phd@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

4.POSTERS: PERSONALIZED MEDICINE AND BIOMARKERS

4a. Individualized Treatment Effect Estimation using Auto-Encoder and Conditional Generative Adversarial Networks

Yuanyuan Liu*, University of Texas Health Science Center at Houston
Momiao Xiong, University of Texas Health Science Center at Houston

Patient heterogeneity has caused many uncertainties in medical research and clinical practice. In this project, we developed a novel framework to estimate individualized treatment effects (ITEs), which quantifies variation in response to the same treatment among patients with heterogenous profiles. Our work builds on conditional generative adversarial networks (CGANs), with auto-encoders to reduce the dimension of covariates. CGANs are used to infer the unseen, individual counterfactuals based on factual outcomes and covariates profiles, which later can be used to obtain ITE estimation. Simulation studies showed that our approach outperformed other state-of-art methods. In addition, we will apply this method to The Cancer Genome Atlas (TCGA) lung cancer dataset to search for best treatment strategy given individual genetic profiles. ITE estimation will have the potential to replace the one-size-fits-all average treatment effects (ATEs) commonly used in clinical practice and provide patient-specific treatment guidance.

email: Yuanyuan.Liu@uth.tmc.edu

4b. Weighted Sparse Additive Learning for ITR Estimation under Covariate Space Sparsity

Jinchun Zhang*, New York University

Individual treatment rule (ITR) estimation is a rapidly growing area in precision medicine; various parametric and semi-parametric methods have been proposed. Though methods have been proposed to estimate ITR estimation when covariate space sparsity presents, no method is suitable for non-linear ITR. To balance this shortcoming, we propose a residual weighted sparse additive learning model which essentially aims to optimize an instance weighted hinge loss with an additive form of decision function. Unlike the linear classifier, the additive form of decision boundary function allows flexibility in estimating ITR, and the lasso-type of penalty enables the model to perform variable selection and ITR estimation simultaneously. Our model performance is examined through extensive simulation studies and the results suggested the new model improved both ITR value and variable selection accuracy.

email: jz2516@nyu.edu

4c. One-Step Value Difference Test for the Existence of a Subgroup with a Beneficial Treatment Effect Using Random Forests

Dana Johnson*, North Carolina State University
Wenbin Lu, North Carolina State University
Marie Davidian, North Carolina State University

Subgroup identification techniques are often used to distinguish patients who may benefit from a particular treatment from those who may not. While post hoc or unprincipled approaches to subgroup identification may lead to spurious results, theoretically justified approaches are essential to precision medicine. We propose a one-step value difference estimator to test for the existence of a subgroup that benefits from an investigative treatment. The test statistic is valid under the exceptional law and converges in distribution to a standard normal random variable as the sample size goes to infinity. If the null hypothesis is rejected, subgroups are identified using a readily available estimated treatment decision rule. We consider four versions of the test statistic, which arise from allowing two forms of an estimator of the value difference, and two options for how observations are partitioned into smaller data sets, which we call “chunks”. In certain simulation settings, the choice of chunking method drastically affected the power of the test. We apply our test to AIDS clinical trial data and observe that the results agree with previous findings.

email: danajo1893@gmail.com

4d. Selecting Optimal Cut-Points for Early-Stage Detection in K-class Diseases Diagnosis Based on Concordance and Discordance

Jing Kersey*, Georgia Southern University
Hani Samawi, Georgia Southern University
Jingjing Yin, Georgia Southern University
Haresh Rochani, Georgia Southern University
Xinyan Zhang, Georgia Southern University

An essential aspect of medical diagnostic testing using biomarkers is to find an optimal cut-point that categorizes a patient as diseased or healthy. This aspect can be extended to the diseases which can be classified into more than two classes. For diseases with general k (k>2) classes, well-established measures include hypervolume under the manifold and the generalized Youden Index. Another two diagnostic accuracy measures, maximum absolute determinant (MADET) and Kullback-Leibler divergence measure (KL), are recently proposed. This research proposes a new measure of diagnostic accuracy based on concordance and discordance (CD) for diseases with k (k>2) classes and uses it as a cut-points selection criterion. The CD measure utilizes all the classification information and provides more balanced class probabilities. Power studies and simulations show that the optimal cut-points selected with CD measure may be more accurate for early-stage detection in some scenarios compared with other available measures. As well, an example of an actual dataset from the medical field will be provided using the proposed CD measure.

email: jingkersey@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

4e. Designing and Analyzing Clinical Trials for Personalized Medicine via Bayesian Models

Chuanwu Zhang*, University of Kansas Medical Center
Matthew S. Mayo, University of Kansas Medical Center
Jo A. Wick, University of Kansas Medical Center
Byron J. Gajewski, University of Kansas Medical Center

Patients with different properties may have different responses to the same medicine. We investigate three Bayesian response-adaptive models for subgroup treatment effect identification: pairwise independent, hierarchical, and cluster hierarchical achieved via Dirichlet Process. The impact of interim analyses and longitudinal data modeling on the enrichment study design is also explored. The performance of designs in terms of power for the subgroup treatment effects and overall treatment effect, sample size, and study duration is investigated via simulation. We apply an innovative cluster hierarchical model for subgroup analysis, integrated two-component prediction method for longitudinal data simulation, and simple linear regression for longitudinal data imputation. Interim analyses are also considered since they can accelerate enrichment studies in cases where early stopping rules for success or futility are met. We found the hierarchical model with interim analyses is an optimal approach to identifying subgroup treatment effects, and the cluster hierarchical model is an alternative approach in cases where insufficient information is available for specifying hyperpriors.

email: chwzhang09@gmail.com

4f. Some Improved Tests for the Assessment of Bioequivalence and Biosimilarity

Rabab Elnaiem*, University of Maryland, Baltimore County
Thomas Mathew, University of Maryland, Baltimore County

Bioequivalence testing deals with assessing the similarity of two drug products. A common bioequivalence criterion is that of average bioequivalence (ABE). A popular test for ABE is the two one-sided t-test (TOST), introduced by Schuirmann (1981). However, the TOST is very conservative when the variability becomes large; I noticed that the conservatism of the TOST can be easily fixed by applying a bootstrap calibration. The type I error then becomes close to the nominal level, giving significant gain in power and giving a considerable reduction in sample size. For assessing similarity, instead of focusing on a criterion based on averages, we can evaluation of the similarity between the distributions of the responses. We propose the use of the overlap coefficient (OVL), which represents the area of overlap between two probability distributions, as a measure of the similarity between distributions. Using a fiducial approach, we have explored the computation of confidence limits for the OVL value. The confidence limits can be used to decide if the OVL value is large enough in order to declare similarity. All the results will be illustrated with practical examples.

email: relnaiem1@umbc.edu

4g. Fusing Continuous and Time-Integrated Data for Estimating Personal Air Pollution Exposures

Jenna R. Krall*, George Mason University
Anna Z. Pollack, George Mason University

Exposure to fine particulate matter air pollution (PM2.5) is associated with increased mortality and morbidity. PM2.5 is a complex mixture of chemical constituents including silicon, zinc, and copper, which are associated with anthropogenic sources such as motor vehicles. Although personal exposure to total PM2.5 can be measured continuously using personal monitors, personal exposure to PM2.5 chemical constituents is frequently only available integrated over time (e.g., over 48 hours). We develop an approach to utilize both continuous PM2.5 and integrated PM2.5 chemical composition data to estimate exposure to source-specific PM2.5. We leverage time-activity data to partition continuous PM2.5 data into commute and non-commute PM2.5. The partitioned data can then be linked to integrated PM2.5 chemical composition data to determine source-specific exposures. We apply our approach to a study of 49 commuters in the DC metro area with continuous PM2.5 and integrated PM2.5 chemical constituent concentrations. Using our approach, we estimate personal exposures to traffic-related PM2.5.

email: jkrall@gmu.edu

4h. Value of Biostatistical Support in a Hospital Quality Improvement Department

Henry John Domenico*, Vanderbilt University Medical Center
Daniel W. Byrne, Vanderbilt University Medical Center
Li Wang, Vanderbilt University Medical Center

Hospital quality improvement groups rarely fund full-time biostatisticians. At Vanderbilt University Medical Center, the quality improvement group has supported two biostatisticians for the past nine years. As this is a substantial investment, many hospitals will be interested in the return on investment before making such a commitment. One benefit of this collaboration is developing rigorous study designs to evaluate the impact of interventions on outcomes. The biostatisticians can design randomized pragmatic trials that evaluate these interventions without interrupting usual care. These evaluations avoid problems associated with traditional before-after designs. Another benefit of biostatistical collaboration is raising the level of sophistication of the graphics and analysis produced by the quality improvement group. Modern graphical techniques allow hospital leaders to focus on signal rather than the noise from traditional graphs. Predictive modeling using electronic health record data allows for deeper understanding of our patient population. The biostatisticians also provide a bridge between academic silos enabling researchers to implement findings after publication.

email: henry.domenico@vumc.org

ABSTRACTS & POSTER PRESENTATIONS

4i. Prediction of Intervention Effects in Healthcare Systems

Emily A. Scott*, Johns Hopkins Bloomberg School of Public Health
Zhenke Wu, University of Michigan
Elizabeth Colantuoni, Johns Hopkins Bloomberg School of Public Health
Sarah Kachur, Johns Hopkins HealthCare
Scott L. Zeger, Johns Hopkins Bloomberg School of Public Health

Waste in the US healthcare system places undue financial burden on patients. We propose a quantitative decision-support tool to evaluate behavioral and clinical interventions on their ability to improve population health at affordable costs. We use clinical and claims data from Johns Hopkins Health Care (JHHC) to investigate members' health states and health state trajectories. We then use a simulation-based approach to predict intervention effects and their associated uncertainty. Our proposed prediction of intervention effects (PIE) model is composed of constituent models corresponding to health state, medical utilization and expenditure, and disenrollment from JHHC. The health state model uses latent class analysis to identify groups of diagnostic codes that commonly co-occur. Each patient's class membership probability distribution is then allowed to change smoothly through time as their underlying health state changes. The constituent models can be adapted to the scientific question of interest. This quality enables evaluation of interventions that target myriad morbidities and comorbidities; users need only tailor model inputs to their specific aims.

email: escott29@jhu.edu

5.POSTERS: CANCER APPLICATIONS

5a. Comparison of Several Bayesian Methods for Basket Trials when a Control of Subgroup-Wise Error Rate is Required

Gakuto Ogawa*, National Cancer Center, Japan
Shogo Nomura, National Cancer Center, Japan

Subgroup-specific analysis (SS) is the conventional approach in basket trials that assess the efficacy of a new agent across multiple histological subtypes in one trial. The notable power gain is expected if one assumes homogeneity of response rates in each subtype and borrows information across subtypes by using a hierarchical Bayesian model (HBM). However, the power gain is seriously lost (Freidlin and Korn, 2013) when "subgroup-wise" (type-I) error rate (SWER) needs to be controlled. This is because one must consider a situation where responses in each subtype were heterogeneous. To better meet a balance between potential homogeneity and heterogeneity, there have been proposed several alternative methods: EXNEX model (Neuenschwander et al., 2016) and multisource exchangeability model (MEM) (Hobbs and Landin, 2018). In this study, via numerical simulation with a control of SWER, we will report a power gain with HBM and EXNEX compared to MEM and SS. We will also report the impact on power when such methods are applied only in enriched subtypes by a clustering of response rates. All simulation studies are motivated by an actual basket trial in oncology.

email: gogawa@ncc.go.jp

5b. Gene Profile Modeling and Integration for EWOC Phase I Clinical Trial Design while Fully Utilizing all Toxicity Information

Feng Tian*, Rollins School of Public Health, Emory University
Zhengjia (Nelson) Chen, Rollins School of Public Health, Emory University

The personalized medicine using gene information is important in modern medicine. This is especially profound for cancer considering that lots of gene mutations are associated with cancer progression and the efficacy of cancer therapy. To apply gene profile in personalized medicine, personalized Maximum Tolerated Dose(pMTD) estimation in phase I clinical trial can be a key step. In this study, we compare four methods for gene profile integration: model selection with logistic regression, regularization (LASSO, RIDGE and Elastic Net), principle components analysis and random forest. Genes or extract components with significance are integrated to the EWOC-NETS model (escalation with overdose control using normalized equivalent toxicity score), a model to estimate pMTDs with a Bayesian adaptive phase I design based on all toxicity information. In our study, most methods show high similarity in selecting important gene predictors for treatment response. The common elements recommended by different methods can recap the major whole gene profiles to be incorporated to EWOC-NETS. This incorporation will have great potential to improve treatment precision and trial efficacy.

email: ftian9@emory.edu

ABSTRACTS & POSTER PRESENTATIONS

5c. A Pan-Cancer and Polygenic Bayesian Hierarchical Model for the Effect of Somatic Mutations on Survival

Sarah Samorodnitsky*, University of Minnesota
Katherine A. Hoadley, University of North Carolina, Chapel Hill
Eric F. Lock, University of Minnesota

We built a novel Bayesian hierarchical survival model based on the somatic mutation profile of patients across 50 genes and 27 cancer types. The pan-cancer quality allows for the model to borrow information across cancer types, motivated by the assumption that similar mutation profiles may have similar (but not necessarily identical) effects on survival across different tumor types. The effect of a mutation at each gene was allowed to vary by cancer type while the mean effect of each gene was shared across cancers. Within this framework we considered four parametric survival models (normal, log-normal, exponential, Weibull), and we compared their performance via a cross-validation approach in which we fit each model on training data and estimate the log-posterior predictive likelihood on test data. The log-normal model gave the best fit, and we investigated the partial effect of each gene on survival via a forward selection procedure. We determined that mutations at TP53 and FAT4 were together the most useful for predicting patient survival. We validated the model via simulation to ensure that our algorithm for posterior computation gave nominal coverage rates.

email: samor007@umn.edu

5d. A Novel GEnomic NEtwork COrrelation Merging System (GENECOMS) to Investigate the Relation between Differentially Expressed Methylation Regions and Gene Modules in Bladder Cancer

Shachi Patel*, University of Kansas Medical Center
Jeffrey Thompson, University of Kansas Medical Center

Bladder cancer (BCa) is one the most common cancers identified in the U.S. veterans. Agent Orange (AO) and Agent Blue (AB), which have been reported as carcinogens, were used during the Vietnam War. We hypothesize that AO/AB exposure causes unique epigenetic alterations, resulting in long-term changes to gene expression. Current methods for investigating the relationship among environmental exposure, epigenetic alterations, gene expression, and disease lack the ability to consider all sources of variation. We propose a novel method, which we call the GEnomic NEtwork COrrelation Merging System (GENECOMS), that uses RNA-seq data to build networks of genes related to both exposure and disease merged with networks of differentially methylated regions (DMRs) based on overall spatial correlation. The result is a unique approach that can indicate how DMRs and genes interrelate with both exposure and disease. The motivation for this approach is to explore the alterations that occur in urothelium cells of Vietnam veterans with exposure to AO/AB. GENECOMS opens the possibility of discovering novel mechanisms through which exposure to toxic chemicals can influence long-term disease risk.

email: spatel14@kumc.edu

5e. Comparing the Performance of Phase I/II Oncology Trial Designs in Low-Toxicity Rate Situations

Ryo Takagi*, Hokkaido University Hospital
Isao Yokota, Hokkaido University Hospital

Main objectives of the phase I study are to find the maximum tolerated dose (MTD) of a drug and those of the phase II study is to assess the efficacy of the drug. Also, there are phase I/II designs that carry out them seamlessly. It is important to search the correct MTD with fewer subjects and to assign more subjects for efficacy assessment in the phase I/II studies. There are many studies to evaluate the MTD, such as 3 + 3 designs, continuous reassessment method (CRM) and Bayesian optimal interval (BOIN), and these designs are based on the high/moderate toxicity rate. However, there are few studies for the assessment the performance for finding the MTD in the drugs with low toxicity rate, such as Metformin. Metformin is widely and safely used as a diabetic drug, but has recently been investigated to have an anticancer effect. The objective of this study is to compare and assess the performance of 3 + 3 design, CRM and BOIN for finding the optimal MTD in the phase I/II study under the situations for low toxicity.

email: r-takagi@pop.med.hokudai.ac.jp

5f. Advantage of Using a Finite-Sample Correction when Designing Clinical Trials in Rare Diseases

Audrey Mauguen*, Memorial Sloan Kettering Cancer Center

Clinical trials are challenging in rare diseases like pediatric cancers, where the potential accrual is limited. Parameter estimation in this context usually assumes that the sample comes from an infinite population. This assumption leads to overestimating the variance when the population is small. While the finite-population correction factor is often used in prevalence and survey context, its use in clinical trials has been limited. This simulation study aims at determining the impact of the finite-population correction on confidence intervals for a binary endpoint from a single-arm clinical trial, when varying the sample and population sizes, and the true proportion in the population. The impact on power and sample size of a comparative trial is also investigated. Depending on the scenario, the correction factor increased the power by up to 3.8%, 6.4% and 8.1% to detect a difference from 30 to 50% with 50, 70 and 100 patients. This translated into a 17% decrease (168 to 139) in accrual to ensure 80% power. This result shows that the same level of confidence can be reached with fewer patients when the disease is rare and the population size is approximately known.

email: mauguena@mskcc.org

ABSTRACTS & POSTER PRESENTATIONS

5g. Implementation of Clusterability Testing Prior to Clustering

Naomi Brownstein*, Moffitt Cancer Center

Cluster analysis is utilized in numerous biometrics applications, such as genomics and cancer to find and study subpopulations of interest. Thus, clustering is useful when the population under study is known to contain multiple distinct subgroups. On the other hand, the interpretation and properties of clustering methods are less clear when the population consists of a single homogeneous population. Clusterability testing enables the user to measure evidence of multiple inherent clusters and signals when such evidence is lacking, potentially rendering cluster analysis inappropriate. There is a need for user-friendly software with clusterability testing to serve as a sanity check before subsequent clustering. In this talk, we first provide a brief introduction to clusterability. Then, we discuss a new package to implement clusterability tests, including a brief sketch of the package requirements and setup. We conclude with example applications.

email: naomi.brownstein@moffitt.org

5h. A Probabilistic Model for Leveraging Intratumor Heterogeneity Information to Enhance Estimation of the Temporal Order of Pathway Mutations during Tumorigenesis

Menghan Wang*, University of Kentucky
Chunming Liu, University of Kentucky
Arnold Stromberg, University of Kentucky
Chi Wang, University of Kentucky

Cancer arises through accumulation of somatically acquired genetic mutations. An important question is to understand the temporal order of mutations during tumorigenesis. We develop a statistical method to estimate the temporal order of mutations in biological pathways while simultaneously accounting for intratumor heterogeneity information of patients and the difference in mutations' functional impacts. A probabilistic model is constructed for each pair of biological pathways to characterize the probability of mutational events from those two pathways occurring in a certain order. Intratumor heterogeneity information, which was inferred by subclonal reconstruction tool PhyloWGS, is incorporated into the model to weigh more on the mutation orders which accord with phylogenic relationship among mutations in each patient. A maximum likelihood method is used to estimate model parameters and infer the probability of one pathway being mutated prior to the other. Analysis of mutation data from The Cancer Genome Atlas demonstrate that our method is able to accurately estimate the temporal order of pathway mutations.

email: mwa287@g.uky.edu

5i. Functional Clustering via Weighted Dirichlet Process Modeling with Breast Cancer Genomics Data

Wenyu Gao*, Virginia Tech
Inyoung Kim, Virginia Tech

Model-based clustering techniques are commonly considered for functional clustering analyses. Parametric model-based clustering methods usually require strong model and distribution assumptions. Nonparametric Bayesian approach, i.e. the Dirichlet process mixture (DPM) modeling requires no assumption for prior distributions, and is able to cluster automatically. The weighted Dirichlet process mixture (WDPM) model further relaxes the homogeneity assumption from DPM model by allowing for multiple candidate priors. The prior assignments are based on weight functions, which are constructed using data information. Thus, the WDPM model would have more promising results when the cluster assignments depend on the data. However, not many literatures have studied in detail about WDPM model, especially when applying to functional clustering. Thus, in this project, we would like to investigate and compare the clustering behaviors from different weight functions proposed by the literatures. Meanwhile, to improve the computation efficiency, the variational Bayes method is evaluated as well. We examined the results through simulation studies and applications with breast cancer genomics data.

email: wenyu6@vt.edu

6.POSTERS: CLINICAL TRIALS

6a. Sample Size Determination Method that Accounts for Selection Probability of the Maximum Tolerated Dose in Phase I Oncology Trials

Yuta Kawatsu*, Tokyo University of Science
Jun Tsuchida, Tokyo University of Science
Shuji Ando, Tokyo University of Science
Takashi Sozu, Tokyo University of Science
Akihiro Hirakawa, The University of Tokyo

In phase I oncology clinical trials, the maximum tolerated dose (MTD) is selected from candidate doses as the recommended dose for subsequent trials. The required number of participants is usually determined based on the probability of correctly selecting the MTD. A time-consuming calculation is necessary to obtain the required sample size. Recently, a time-saving sample size determination method accounting for the posterior probability of toxicity included in the assumed tolerance interval is proposed. This method does not account for the probability of correctly selecting the MTD. In this study, we proposed a time-saving sample size determination method that accounts for the probability of correctly selecting the MTD. We then compared the performance of the proposed method with those of the conventional time-consuming method and the time-saving method. The sample sizes of the proposed method were nearly the same as those of the conventional method. The proposed method may be a time-saving alternative to the conventional method.

email: 4418509@ed.tus.ac.jp

ABSTRACTS & POSTER PRESENTATIONS

6b. The Scale Transformed Power Prior with Applications to Studies with Different Endpoints

Brady Nifong*, University of North Carolina, Chapel Hill
Matthew A. Psioda, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

We develop a scale transformed version of the power prior to accommodate settings in which the historical data and the current data involve different data types, such as binary and survival data. This situation arises often in clinical trials, for example, when the historical data involves binary response data and the current data may have a time-to-event outcome. The traditional Power Prior proposed by Ibrahim and Chen (2000) does not account for different data types in the context discussed here. Thus, a new type of power prior needs to be formulated in these settings, which we call the scale transformed power prior. The scale transformed power prior is constructed so that the information matrix based the current data likelihood is appropriately scaled by the information matrix from the power prior, thereby shifting the scale of the parameter vector from the historical data to the new data. Several examples are presented to motivate the scale transformation and several simulation studies are presented to show the advantages in performance of the scale transformed power prior over the power prior and other priors.

email: bsmelton@live.unc.edu

6c. Design and Analysis for Three-Arm Clinical Trials with Intra-Individual Right-Left Data

Ryunosuke Machida*, National Cancer Center, Japan
Kentaro Sakamaki, Yokohama City University, Japan
Aya Kuchiba, National Cancer Center, Japan

Chemotherapy-induced peripheral neuropathy can appear in both hands (R: right hand, L: left hand). To evaluate the effect of treatment (e.g., cryotherapy) on such symptoms, the intra-individual R vs L comparison design can be considered. When multiple treatments are available, 3-arm 2-period cross-over design, which is the intra-individual comparison design for 3 treatments, can be a potential design for this situation. However, since outcome of one hand might be affected by the treatment effect on the other hand, especially for subjective outcome, it would be difficult to appropriately obtain outcome data when each hand receives different treatments. Thus, we propose the design randomized to 4 groups for 3-arm comparison: (1) no treatment for R; treatment A for L, (2) treatment A; no treatment, (3) no treatment; treatment B, and (4) treatment B; no treatment. This design allows to compare the effects of each treatment to no treatment and between experimental treatments. We compared the power of the proposed design with the randomized 3-arm parallel design where each patient is randomized to either of 3 treatments, for several situations of R-L correlations.

email: rmachida@ncc.go.jp

6d. An Estimation of Efficacy of Potential Drug in Multiple Diseases with Discriminating Heterogeneity in Treatment Effects in Basket Trials

Shun Hirai*, Tokyo University of Science
Jun Tsuchida, Tokyo University of Science
Shuji Ando, Tokyo University of Science
Takashi Sozu, Tokyo University of Science
Akihiro Hirakawa, The University of Tokyo

The development of drugs for patients with specific markers has accelerated. A basket trial, where multiple types of patients with certain diseases are enrolled, is useful for the development of such drugs. A Bayesian hierarchical model is commonly used for facilitating information sharing among diseases. This model assumes that each disease parameter is exchangeable with parameters of other diseases. However, this assumption is not always satisfied. An exchangeability-nonexchangeability (EXNEX) model that consists of two models with and without exchangeability assumptions has been proposed. However, the mixture ratio in EXNEX model is often estimated to be non-zero, even if a disease that is significantly different from other diseases is included in the trial. The EXNEX model is not appropriate when non-exchangeable diseases are grouped together in the trial, even if a single exchangeability component is extended to more than one component. We have proposed a method for estimation of efficacy of potential drug on each disease, with discriminating heterogeneity in treatment effects. In the proposed method, the exchangeability assumption of each disease is evaluated.

email: 4418519@ed.tus.ac.jp

6e. Longitudinal Study of Opioid Treatment on Hamilton Depression Rating Scale Using a Negative Binomial Mixed Model

Kesheng Wang*, West Virginia University
Wei Fang, West Virginia Clinical and Translational Science Institute
Toni DiChiacchio, West Virginia University
Chun Xu, University of Texas Rio Grande Valley
Ubolrat Piamjariyakul, West Virginia University

Hamilton Depression Rating Scale (HAM-D) can monitor changes in depression. However, the HAM-D scale is a count outcome. This study aimed to evaluate the opioid treatment effect on depression using repeated measures mixed model in a randomized clinical trial. Data was obtained from National Drug Abuse Treatment CTN protocol-0051. Opioid-dependents (N=570) were randomly assigned to receive Buprenorphine-naloxone (BUP-NX) and extended-release naltrexone (XR-NTX). The HAM-D scores were completed at weeks 1, 2, 3, 4, 8, 12, 16, 20, 24, 28 and 36. Both AIC and BIC statistics revealed a negative binomial mixed model (NBMM) was better than Poisson regression and the unstructured (UN) covariance structure showed more suitable than other covariance structures. Significant interactions exist between treatment and time at weeks 1, 2, 3, and 4; when XR-NTX increased HAM-D scores.

ABSTRACTS & POSTER PRESENTATIONS

Baseline anxiety, major depressive, and amphetamines use disorder were associated with increased HAM-D scores. The NBMM addresses the issue of over-dispersion in the count data and covariance structures of repeated measures. When using XR-NTX, it is essential to monitor for changes in depressive symptoms.

email: kesheng.wang@hsc.wvu.edu

6f. Group Sequential Analysis for Sequential Multiple Assignment Randomized Trials

Liwen Wu*, University of Pittsburgh
Junyao Wang, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

Sequential multiple assignment randomized trial (SMART) facilitates simultaneous comparison of multiple adaptive treatment strategies (ATSs), where patients are sequentially randomized to treatments given their response to their own treatment history. Previous studies have established a framework to test the homogeneity of multiple ATSs by a global Wald test on the inverse probability weight-normalized estimators. SMART trials are generally lengthier than classical trials due to the sequential nature of treatment randomization in multiple stages. Thus, it would be beneficial to add interim analyses allowing for early stop if overwhelming efficacy is observed. We extend the traditional group sequential methods to SMARTs and propose a method to perform interim analysis based on an approximation of bivariate chi-square distribution. Simulation studies demonstrate that the proposed method maintains desired type I error and power with reduced expected sample size. Lastly, we apply our method to real data from a SMART trial assessing the effects of cognitive behavioral and physical therapies in patients with knee osteoarthritis and comorbid subsyndromal depressive symptoms.

email: liw88@pitt.edu

6g. Incorporating Truncation Information from Phase I Clinical Studies into Phase II Designs

Li-Ching Huang*, Vanderbilt University Medical Center
Fei Ye, Vanderbilt University Medical Center
Yi-Hsuan Tu, Independent Scholar
Chia-Min Chen, Nanhua University, Taiwan
Yu Shyr, Vanderbilt University Medical Center

A large body of literature has sought to develop better Phase I/II study for oncology trials. Commonly used Phase I designs include conventional and modified “3+3” type of designs, continual reassessment method (CRM), and modified toxicity probability interval design (mTPI). Simon’s two-stage design is likely the most widely implemented design in Phase II clinical trials. Some of studies tend to allow the patients treated at maximum tolerated dose (MTD) in Phase I study to be included in Phase II analysis. A common assumption in oncology trials is that the higher the dose, the greater the likelihood of efficacy and toxicity. In a conventional 3+3 design, because number of dose limiting toxicities (DLTs) in the MTD cohort of six patients is restricted to be no more than 2, we jointly model the toxicity and efficacy outcomes from phase I study and use the marginal distribution of efficacy outcome to determine design characteristics. A modified Simon’s two-stage design is proposed to incorporate evidence from the Phase I trial MTD cohort in Phase II design. The proposed method is evaluated with comparisons to Simon’s optimal and minimax designs.

email: li-ching.huang.1@vumc.org

6h. Replicability of Treatment Effects in Meta-Analyses

Kirsten R. Voorhies*, Brown University
Iman Jaljuli, Tel-Aviv University
Ruth Heller, Tel-Aviv University
Orestis A. Panagiotou, Brown University

Introduction: Replicability, i.e. the property that multiple trials find a treatment effect, is important for identifying interventions immune to type I error. Meta-analysis does not ensure replicability as 1 study may drive statistical significance. Methods: We tested the replicability hypothesis that the effect is replicated in at least u studies in a meta-analysis among Cochrane meta-analyses of >=5 studies. We applied the partial conjunction hypothesis test which calculates the r-value as the largest p-value of all subsets after removing u-1 studies. Results: Eligible were 20,470 meta-analyses. Replicability for u=2 was not met (r>0.05) in 12,821 (63%) and for u=3 in 14,485 (73%) meta-analyses. Of 9,665 meta-analyses with p<0.05, replicability for u=2 was not met in 2,418 (25%) with 1 study driving significance. For u=3, replicability was not met in 4,131 (43%) with 2 studies driving significance. Of 5,093 random-effects meta-analyses with p>0.05, 402 (8%) rejected the replicability null for u=2 and 91 (2%) for u=3. Discussion: Effects are replicated in at least 2 trials in 75% and at least 3 trials in 57% of statistically significant Cochrane meta-analyses.

email: kirsten_voorhies@brown.edu

ABSTRACTS & POSTER PRESENTATIONS

7.POSTERS: DIAGNOSTICS/PREDICTION/AGREEMENT

7a. A Resampling Perspective on Evaluation of Diagnosis Accuracy: An Appendicitis Example

Calvin S. Elder*, St. Jude Children's Research Hospital
Yousef El-Gohary, Center of Colorectal and Pelvic Reconstruction
Hui Zhang, Northwestern University
Li Tang, St. Jude Children's Research Hospital

Despite recent controversy over the value of p in science, calculating a p-value is often necessary in addressing key scientific questions, as a p-value is still considered an objective metric when performing pre-specified hypothesis tests. When the underlying distribution of study data is known, classical formulas can be applied. In practice, the study data does not always follow a known distribution and is bound by sample size, so relying on formulas to calculate p-values is not always effective. As an example, one could compare diagnostic accuracy of classification techniques, quantified by both sensitivity and specificity, over a small sample size. We present bootstrapping in combination with ideas borrowed from ROC curves as a method of comparing the performance of multiple classification techniques. By resampling the data, we can visualize the empirical distribution of interest and show that a reliable estimate can be attained. Simulation examples are provided to demonstrate the value of the proposed resampling method. We showcase the utility of our method on an appendicitis clinical research study, comparing two clinical centers with differing diagnostic techniques.

email: eldcs-21@rhodes.edu

7b. Improving the Performance of Polygenic Risk Score with Rare Genetic Variants

Hongyan Xu*, Augusta University
Varghese George, Augusta University

Polygenic risk score combines the genetic contribution from multiple genetic variants across the genome and has the potential clinical utilities in predicting disease risk for common human diseases. Current polygenic risk scores are based on the results from genome-wide association studies, where the information is from common genetic variants. With the availability of genome sequencing data, many rare genetic variants have been shown to be associated with the risk of common human diseases. In this study, we build a polygenic risk score based on both common and rare genetic variants. We incorporate the effects of rare genetic variants by dividing them into subgroups according to the signs of the effect and combine the rare genetic variants in one subgroup with a genetic load. Results from our simulation study show that our polygenic risk score has improved predictability measured by the C-statistic. Our polygenic risk score also has lower prediction error from cross-validation than the risk score without rare genetic variants. We applied our polygenic risk score method to the breast cancer data from the Cancer Genome Atlas to predict the risk of developing breast cancer.

email: hxu@augusta.edu

7c. A Domain Level Index to Enhance the Prediction Accuracy of Pathogenic Variants

Hua-Chang Chen*, Vanderbilt University Medical Center
Qi Liu, Vanderbilt University Medical Center

Sequencing is becoming more accurate and affordable, a vast number of mutations could be detected in a single individual, however only handful of them are disease-causing, hence the effective distinguishing of pathogenic from benign variants remains the main obstacle of the clinical utility of sequencing. There are many prioritization tools available now to facilitate the interpretation, including the gene-level and some variant level metrics. To better illustrate the heterogeneity within a given gene, we built the domain damage index (DDI), a protein domain level variants prioritizing tool by depicting the distribution pattern of low frequency missense mutations in domains. Our result shows that, the domains with higher DDI which devoid of missense mutations are significantly enriched with pathogenic variants. Compared to the existing gene level metrics, DDI have higher accuracy to predict the pathogenic variants, also DDI outperforms the variant level methods in identifying the pathogenic variants within the known disease-causing genes. Finally, DDI provides novel and robust evidences to prioritizing variants via highlighting those regions with high constraint.

email: hua-chang.chen@vumc.org

7d. The Cornelius Project - Randomizing Real-Time Predictive Models Embedded in the Electronic Health Record to Assess Impact on Health Outcomes

Daniel W. Byrne*, Vanderbilt University
Henry J. Domenico, Vanderbilt University
Li Wang, Vanderbilt University

To improve health outcomes, healthcare systems are beginning to use real-time predictive models in the electronic health record. The problem is that currently there is more hype than evidence that these models benefit patients. Before this approach becomes standard of care, robust proof is required, with randomized controlled pragmatic trials, that implementation of the models result in improved health outcomes. Healthcare systems are notoriously resistant to change and yet success of these projects requires a significant redesign of job roles and allocation of prevention resources. To overcome this resistance and provide the infrastructure to support pragmatic trials we created a Learning Healthcare System platform as part of our CTSA grant. Over the past 7 years, as part of the Cornelius predictive modeling project, we have discovered how to overcome many of the implementation challenges. Healthcare systems that lead the way in creating, implementing, and assessing outcomes with randomization are the more likely to thrive in the future. Randomization is essential in not only assessing if the model can improve outcomes but also ensuring that it does not cause harm.

email: daniel.byrne@vanderbilt.edu

ABSTRACTS & POSTER PRESENTATIONS

7e. Privacy-Preserving Outcome Prediction

Lamin Juwara*, McGill University

In machine learning, data mining and statistics, constructing predictive models sometimes require the use of sensitive data stored in different study centers/nodes (e.g. medical or financial data). It is often challenging to obtain individual-level data due to privacy concerns, which has led to increased interest in privacy-preserving data analysis. To construct predictive models under data privacy constraints, we propose a new approach for constructing predictive models (at a central analytical node) using pooled/aggregate data provided by the different nodes. More specifically, we construct a pooled logistic model for disease diagnosis using aggregate covariate data. The performance of the pooled model is then compared to a standard logistic model constructed from individual-level data. Using biomarker data as a case study, we show that the pooled model results in comparable performance to the standard model under various distributional forms of the biomarker data.

email: lamin.juwara@mail.mcgill.ca

7f. Interpretable Clustering of Hierarchical Dependent Binary Data: A Doubly-Multi-Resolution Approach

Zhenke Wu*, University of Michigan
Yuqi Gu, University of Michigan
Mengbing Li, University of Michigan
Gongjun Xu, University of Michigan

We consider clustering dependent binary data observed at internal and leaf nodes over multiple rooted trees. For example, hierarchical diagnosis codes characterize increasingly specific conditions and may reveal patient subgroups that guide care. Existing methods are unable to fully characterize multi-level variations, resulting in poor clustering accuracy. Building on structured latent attribute models (SLAM) that perform interpretable clustering, we propose a Doubly-Multi-Resolution (DMR) framework to match multi-level observations and latent attributes. We prove partial identifiability and statistical consistency in the selection and prevalence estimation of the subset of latent attribute patterns. Simulation studies show that DMR SLAM significantly outperforms state-of-the-art flattened analyses in terms of clustering accuracy. We apply the proposed model to an administrative claims data set for latent disease phenotyping ~15,000 privately-insured cancer-associated thrombosis patients using hundreds of hierarchical International Classification of Diseases-Version 9 (ICD-9) codes.

email: zhenkewu@umich.edu

7g. Estimation and Construction of Confidence Intervals for the Cutoff-Points of Continuous Biomarkers Under the Euclidean Distance in Trichotomous Settings

Brian Mosier*, University of Kansas Medical Center
Leonidas Bantis, University of Kansas Medical Center

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal forms of cancer with a 5-year survival rate between 5-7%. Thus, its early detection could be key to reducing PDAC mortality. Novel biomarkers are currently being studied for the early detection of PDAC. The ROC based Youden index is a popular method for choosing the optimal cutoff values of biomarkers. It was only recently that it was generalized to the three-class case. However, in such trichotomous settings it has the drawback of ignoring one of the three classes when estimating the optimal cutoff values. When moving to more than three classes this issue is further compounded. We provide new parametric and nonparametric approaches based on the Euclidean distance from the perfection corner that can take advantage of the full sample size. We further generalize our approaches to address k-class problems. Our results indicate estimated cutoffs with smaller variance and as a result, narrower confidence intervals compared to Youden index-based ones. Our approach is illustrated using real data from pancreatic cancer patients from a study conducted at the MD Anderson Cancer Center.

email: bmosier@kumc.edu

7h. Confidence Interval of the Mean and Upper Tolerance Limit for Zero-Inflated Gamma Data

Yixuan Zou*, University of Kentucky
Derek S. Young, University of Kentucky

In practice, it is not uncommon to observe count data that possess excessive zeros (i.e., zero-inflation) relative to the assumed discrete distribution. When data are semicontinuous, the log-normal and gamma distributions are often considered for modeling the positive part of the model. The problems of constructing a confidence interval for the mean and calculating an upper tolerance limit of a zero-inflated Gamma population are considered using generalized fiducial inference. Our simulation studies indicate that the proposed method is very satisfactory in terms of coverage properties and precision. An application to medical data also highlights the utility of the proposed method.

email: yixuanzou@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

7i. Predictive Performance of Physical Activity Measures for 1-year up to 5-year All-Cause Mortality in NHANES 2003-2006

Lucia Tabacu*, Old Dominion University
Mark Ledbetter, Lynchburg University
Andrew Leroux, Johns Hopkins University
Ciprian Crainiceanu, Johns Hopkins University

We study the predictive performance of physical activity measures for each year, starting with 1-year up to 5-year all-cause mortality in adults between 50 and 85 years old. In NHANES 2003-2006 the physical activity data is collected by accelerometers at every minute and for 7 days. In univariate logistic regression the total activity count was the best physical activity predictor in each year all-cause mortality model. In multivariate logistic regression the physical activity covariates were selected as top predictors using cross-validated AUC. This is joint work with M. Ledbetter, A. Leroux and C. Crainiceanu.

email: ltabacu@odu.edu

8.POSTERS: ADAPTIVE DESIGN/EXPERIMENTAL DESIGN

8a. An Empirical Bayesian Basket Trial Design Accounting for Uncertainties of Homogeneity and Heterogeneity of Treatment Effect among Subpopulations

Junichi Asano*, Pharmaceuticals and Medical Devices Agency
Akihiro Hirakawa, The University of Tokyo

A basket trial in oncology enrolls patient populations with a specific molecular status among several cancer types and often evaluates the response rate of investigational therapy across cancer types. Recently developed statistical methods for evaluating response rate in basket trials can be roughly categorized into two groups: those that account for the degrees of homogeneity/heterogeneity of response rates among subpopulations and those using information borrowing of response rate among subpopulations using Bayesian hierarchical models. In this study, we developed a new basket trial design that accounts for the uncertainties of homogeneity and heterogeneity of response rates among subpopulations using the Bayesian model averaging approach. We demonstrated the utility of the proposed method by comparing against other methods in the two methodological groups using simulated data. The proposed method would be useful in the practical setting of “signal-finding” basket trials without prior information on the treatment effect of the investigational drug because it does not require specifications of prior distribution on homogeneity of response rates among subpopulations.

email: asano-junichi@pmda.go.jp

8b. Lessons Learned in Developing an Interdisciplinary Collaboration Between Biostatistics and Forensic Nursing

Yesser Sebeh*, Georgia State University
Katherine Scafide, George Mason University
Matthew J. Hayat, Georgia State University

A recent National Institute of Justice funded study of detection and visibility of cutaneous bruises provided an opportunity for a meaningful and rich biostatistics collaboration and learning experience. The study entailed a crossover randomized controlled trial with repeated measures designed to compare different alternate light wavelengths and filters to white light. The data were complex and multilevel. I will discuss my learning experience as a graduate student about the collaborative process, including developing 3-level multilevel models, products of this work that have a real-world clinical impact, and lessons learned about interdisciplinary collaboration.

email: Yessersebeh1@gmail.com

8c. Response-Adaptive Randomization in a Two-Stage Sequential Multiple Assignment Randomized Trial

Junyao Wang*, University of Pittsburgh

Sequential multiple assignment randomized trials (SMARTs) are systematic and efficient media for comparing dynamic treatment regimes (DTRs), where each patient is involved in multiple stages of treatment with the randomization at each stage depending on that patient’s previous treatment history and interim outcomes. However, in standard SMART designs the differential benefits of treatments observed during the previous stages are neglected while assigning treatment at the current stage. In this paper, we propose a response-adaptive SMART (RA-SMART) design that incorporates a technique of unbalancing the allocation probabilities at the current stage based on the accumulated information on treatment efficacy from previous stages. A simulation study is conducted to assess the operating characteristics of RA-SMART design, including the consistency and efficiency of response rate estimates for each DTR, and the power of identifying the optimal DTR. Some practical suggestions on design parameters are discussed at the conclusion.

email: juw55@pitt.edu

ABSTRACTS & POSTER PRESENTATIONS

8d. Integrated Multiple Adaptive Clinical Trial Design Involving Sample Size Re-Estimation and Response-Adaptive Randomization for Continuous Outcomes

Christine M. Orndahl*, Virginia Commonwealth University
Robert A. Perera, Virginia Commonwealth University

Single adaptive clinical trial designs are utilized most often, where only one adaptive design is used within the clinical trial. Consequently, only one pitfall of a fixed clinical trial design is addressed. Increased interest has been developed in the area of multiple adaptive designs, incorporating more than one adaptive design within a single clinical trial. However, multiple adaptive designs are typically performed independently and sequentially. The goal of this project is to integrate multiple adaptive designs, specifically sample size re-estimation and response-adaptive randomization, into a clinical trial with a continuous outcome. To accomplish this, the weighted sum method for multi-objective optimization with a constraint to maintain statistical power is used to combine two objective functions. The first minimizes the total expected treatment response while the second minimizes the sample size required. These objective functions serve to adaptively adjust the allocation ratio and sample size. Results for applying these new methods to a clinical trial are presented.

email: orndahlc@vcu.edu

8e. Design of a Calibrated Experiment

Blaza Toman*, National Institute of Standards and Technology (NIST)
Michael A. Nelson, National Institute of Standards and Technology (NIST)

Efficient experimental design is a critical aspect of practical scientific planning and measurement execution. Achieving fit-for-purpose measurement results using limited resources is a significant priority for laboratories. In this presentation we show how to optimally construct fit-for-purpose measurement schemes that achieve appropriate confidence. Specifically, we plan a two-stage experiment with a calibration phase followed by the measurement of an unknown via LC-IDMS, using an isotopically-enriched internal standard. An experimental design for the two phased procedure consists of the following quantities: the number of calibration standards, the number of replications of the measurements for each calibration standard, the set of nominal values of the standards, the number of samples of the unknown in the second experiment, and the number of replicates per sample. We will show how to optimally select the experimental design which guarantees that the expected relative measurement uncertainty is at most a pre-specified percentage.

email: blaza.toman@nist.gov

8f. Modified Q-learning with Generalized Estimating Equations for Optimizing Dynamic Treatment Regimes with Repeated-Measures Outcomes

Yuan Zhang*, University of Minnesota
David Vock, University of Minnesota
Thomas Murray, University of Minnesota

Dynamic treatment regimes (DTRs) are of increasing interest in clinical trials and personalized medicine because they allow tailoring decision making based on a patient's treatment and covariate history. In some sequential multiple assignment randomized trials (SMARTs) for children with developmental language disorder (DLD), investigators monitor a patient's performance by collecting repeated-measures outcomes at each stage of randomization as well as after the treatment period. Standard Q-learning with linear regression as Q-functions is widely implemented to identify the optimal DTR, but fails to provide point estimates of average treatment effect (ATE) at every time point of interest. Moreover, Q-learning in general is susceptible to misspecification of outcome model. To address these problems, we propose a modified version of Q-learning with a generalized estimating equation (GEE) as Q-functions. Simulation studies demonstrate that the proposed method performs well in identifying the optimal DTR and is also robust to model misspecification.

email: zhan5817@umn.edu

8g. Development of a Spatial Composite Neighborhood SES Measure

Shanika A. De Silva*, Drexel University
Melissa Meeker, Drexel University
Yasemin Algur, Drexel University
Victoria Ryan, Drexel University
Leann Long, University of Alabama at Birmingham
Nyesha Black, Noire Analytics
Leslie A. McClure, Drexel University

Neighborhood socioeconomic status (NSES) is an important variable to consider when attempting to reduce health inequalities. Most research in this area has focused on using global measures of NSES to make inferences at the local level. However, socioeconomic processes seldomly occur constantly and independently over geographic space. Therefore, it is important to take into account the collinearity and spatial dimensions of variables in the construction of NSES measures. The purpose of this study is to assess the benefit of adapting a global composite NSES index to integrate spatial effects. Principal components analysis (PCA) and geographically weighted principal components analysis (GWPCA) were used to develop non-spatial and spatial NSES indices for census tracts in Pennsylvania. The GWPCA approach highlighted tracts that were most socioeconomically disadvantaged. The spatial composite NSES index could allow decision-makers to identify hotspots within the state where priority actions need to be taken to reduce health inequalities.

email: sad345@drexel.edu

ABSTRACTS & POSTER PRESENTATIONS

8h. Estimating Disease Prevalence with Potentially Misclassified Dorfman Group Testing Data

Xichen Mou*, University of Memphis
Joshua M. Tebbs, University of South Carolina
Dewei Wang, University of South Carolina

When screening a large population for an infectious disease, such as chlamydia or HIV, public health laboratories often first perform diagnostic tests on pools of individual specimens (e.g., blood, urine, swabs, etc.). Those pools which test positively are subsequently resolved to determine which individuals are positive--usually by retesting each individual separately. Testing pools of individuals, which is also known as group testing, saves laboratories time and money because numerous individuals can be classified negatively with a single test. In this article, we consider the problem of estimating the population-level probability of disease when this two-stage (Dorfman) group testing algorithm is used. For maximum flexibility, we regard assay accuracy probabilities as unknown. Our estimation methods utilize maximum likelihood, so we are able to develop large-sample inference for various case identification characteristics in group testing, including the expected number of tests and classification accuracy probabilities. We illustrate the techniques in this article using chlamydia data collected by the State Hygienic Laboratory at the University of Iowa.

email: xmou@memphis.edu

9.POSTERS: BAYESIAN METHODS

9a. Bayesian Spatial Analysis of County-Level Drug Mortality Rates in Virginia

Jong Hyung Lee*, Virginia Commonwealth University
Derek A. Chapman, Virginia Commonwealth University

Over the past decade, the U.S. drug mortality rate has shown a two-fold increase. Mortality rates can vary by different geographic areas due to differences in socioeconomic and environmental characteristics. The primary objective of this study was to investigate the spatial distribution of drug mortality rates across Virginia through Bayesian spatial modeling by an approximation method, the integrated nested Laplace approximation (INLA). This study also aimed to identify potential risk and protective factors that were associated with drug mortality outcome. In this study, counties in Virginia (n=133) were the geographical units. The aggregated (2012-2016) death counts for each county were used to compute the drug mortality rate and its county-specific log relative risk. Based on the posterior mean estimates of the model, five covariates indicated statistical significance. The exceedance probability indicated eleven counties had above 50% excess risk of mortality. The results of this study could provide insights for which specific areas need to be targeted for interventions in the context of community-level risk factors.

email: leejh47@vcu.edu

9b. Robust Partial Reference-Free Cell Composition Estimation in Tissue Expression Profiles

Ziyi Li*, Emory University
Zhenxing Guo, Emory University
Ying Cheng, Yunnan University
Peng Jin, Emory University
Hao Wu, Emory University

High cost, intensive labor requirements and technical limitations hinder the cell composition quantification using cell sorting or single-cell technology. As alternatives, reference-based deconvolution algorithms are limited if no appropriate reference panel from purified tissues is available, while reference-free deconvolution algorithms are suffered from low accuracy and difficulty in cell type label assignment. Here, we introduce TOAST, a partial reference-free algorithm for estimating cell composition of heterogeneous tissues from their gene expression profiles. Guided by prior information obtainable from various sources, the proposed method can capture the cell composition significantly better than existing methods in extensive simulation studies and real data analyses. We evaluate the markers obtained from different high-throughput modalities and existing repositories, confirming the proposed method performs better than existing methods wherever the prior knowledge types is obtained. Finally, the analyses of two Alzheimer's disease datasets show consistency of the proposed method with cell compositions from single cell technology and existing knowledge.

email: ziyi.li@emory.edu

9c. Multivariate Space-Time Disease Mapping via Quantification of Disease Risk Dependency

Daniel R. Baer*, Medical University of South Carolina
Andrew B. Lawson, Medical University of South Carolina

There is an ongoing need to characterize progression in Alzheimer's disease (AD) risk. As such, we focus on characterizing how at-risk patients with a precursor to AD, called mild cognitive impairment (MCI), transition from MCI to AD via novel variants of the Bayesian space-time mixture (STM) model for disease mapping. In particular, we make use of a non-parametric quantification of the dependence between area-specific MCI and AD risk trajectories called the maximal information coefficient (MIC). The MIC then allows us to conditionally share latent temporal disease risk components within the STM modelling framework in order to best capture area-specific temporal evolution of disease risk between MCI and AD. Novel aspects of this research include a comparison of model performance between the original STM model and our STM model variants, as well as a novel comparison of the continuous and discrete formulations of STM model. Preliminary results suggest that our novel STM model variants outperform the original STM model in terms of goodness of fit to AD spatiotemporal incidence data.

email: baerd@musc.edu

ABSTRACTS & POSTER PRESENTATIONS

9d. Bayesian Envelope in Logistic Regression

Minji Lee*, University of Florida
Zhihua Su, University of Florida

We propose a new approach to Bayesian logistic regression using the envelope method, which aims to gain estimation efficiency in multivariate analysis. We provide a data augmentation Metropolis with in Gibbs sampler for implementation of our model. In addition, we illustrate the applicability of the Bayesian envelope in logistic regression on simulations and real datasets.

email: mlee9@ufl.edu

9e. Bayesian Kinetic Modeling for Tracer-Based Metabolomic Data

Xu Zhang*, University of Kentucky
Ya Su, University of Kentucky
Andrew N. Lane, University of Kentucky
Arnold Stromberg, University of Kentucky
Teresa W-M. Fan, University of Kentucky
Chi Wang, University of Kentucky

Kinetic modeling of the time dependence of metabolite concentrations including the stable isotope labeled species is an important approach to simulate metabolic pathway dynamics. It is also essential for quantitative metabolic flux analysis using tracer data. However, as the metabolic networks are complex including extensive compartmentation and interconnections, the parameter estimation for enzymes that catalyze individual reactions needed for kinetic modeling is challenging. We propose a Bayesian approach that specifies an informative prior distribution for kinetic parameters. This prior knowledge prioritizes regions of parameter space that encompass the most likely parameter values, thereby facilitating robust parameter estimation. A component-wise adaptive Metropolis algorithm is used to generate the posterior samples of the kinetic parameters and conduct hypothesis tests under different treatments. Simulation studies using defined networks are used to test the performance of this algorithm under conditions of variable noise.

email: xzh323@g.uky.edu

9f. Forecasting Glaucoma Progression using Bayesian Structural Time Series Analysis

Manoj Pathak*, Murray State University

Glaucoma, a group of progressive optic neuropathies, permanently damages vision and progress to complete irreversible blindness if left untreated. Early detection of glaucoma and prior knowledge of its future progression profile help clinicians adopt appropriate treatment modalities to halt or slow down glaucoma progression. A global index such as Mean Deviation (MD) from standard automated perimetry is often used to assess the disease. However, forecasting the progression using MD is difficult because MD measurements from glaucomatous eyes are considerably noisy and that there are rarely sufficient data to build and validate a predictive model. This study uses Bayesian Structural Time Series (BSTS) method to reveal a hidden trend of glaucoma progression and get an improved estimate of the disease progression profile by incorporating uncertainties into the forecast.

email: mpathak@murraystate.edu

9g. A Three-Groups Bayesian Approach to GWAS Data with Application to Parkinson's Disease

Vivian Cheng*, The Pennsylvania State University
Daisy Philtron, The Pennsylvania State University
Ben Shaby, Colorado State University

Alongside the development of gene expression technology as a gateway to genetic disease, the need to extract useful information of genetic data has gone up. To understand the potential role of genes associated with Parkinson's disease, we examine the data from a genome-wide association study. Our proposed statistical model is a hierarchical three-group mixture of distributions, which classifies genes related to Parkinson's disease into three groups as null, deleterious or beneficial. We examine the model performance based on simulation studies and apply the model to Parkinson's disease.

email: xzc90@psu.edu

9h. Improving Estimation of Gene Expression Differences via Integrative Modeling of Transcriptomic and Genetic Data

Xue Zou*, Duke University
William H. Majoros, Duke University
Andrew S. Allen, Duke University

High-throughput genetic sequencing technologies have revolutionized genetic diagnoses of rare genetic disease as they allow a more comprehensive characterization of variation in patient genomes. One potential reason for the existence of patients with undiagnosed disease is that, noncoding variants that disrupt the expression of a gene, leading to a disease phenotype, are typically not considered in current diagnosis. To identify such cases, we propose to use allele-specific-expression (ASE) as a marker for cis-regulatory gene disruption. When multiple heterozygous variants are present in a gene, the allelic phase becomes an important factor in estimating ASE. In this study, we construct a Bayesian hierarchical model to integrate prior information and explicitly account for phasing error estimating ASE. We also propose to integrate family data, both to improve phasing as well as to allow the identification of genes displaying characteristic patterns of ASE across the family that are informative of the inheritance pattern of the underlying causal regulatory variation.

email: xue.zou@duke.edu

ABSTRACTS & POSTER PRESENTATIONS

9i. Reliable Rates and the Effect of Prior Information with an Application to the County Health Rankings & Roadmaps Program

Guangzi Song*, Drexel University
Harrison Quick, Drexel University
Loni Philip Tabb, Drexel University

When analyzing rare event and mortality data, the criteria for declaring rate estimates as “reliable” is still a matter of dispute. What these varying criteria have in common, however, is that they are often infeasible in most small area analysis settings. In this study, we provide a general definition for a “reliable” estimate in a Bayesian framework in which a rate is defined as reliable at the 1- alpha level if its posterior median is larger than the width of its (1- alpha)*100% equal-tailed credible interval, thereby allowing prior information to improve the precision of our estimates and thus yield more reliable estimates. After describing the properties of this definition via a Poisson-gamma conjugate pairing, we extend these criteria to the more general lognormal prior specification commonly used in the disease mapping literature. Following an illustration using data from the County Health Rankings & Roadmaps program, we describe how this work can be used to ensure that prior distributions do not overpower the data in our analyses and to provide guidance with respect to the aggregation of data.

email: gs556@drexel.edu

10. POSTERS: CAUSAL INFERENCE AND CLINICAL TRIALS

10a. The Importance of Propensity Score Estimation to Achieve Balance in Covariates

Hulya Kocyigit*, University of Georgia

The propensity score is a balancing score: conditional on the propensity score, treated and untreated subjects have the same distribution of observed baseline characteristics. Such lack of covariate balance creates some problems. Such as lack of balance in covariates make the inferences imprecise. I use SUPPORT data, Study to Understand Prognoses and Preferences, is used to investigate trimming to improve balance in covariates distributions. However, there are some alternatives to improve balance in covariates without applying propensity score trimming by using different techniques: propensity score matching with logistics regression model, generalized boosted model and covariate balancing propensity score and illustrate how to estimate the propensity score such that the resulting covariate balance. Then, I evaluate the performance of those techniques using Monte Carlo simulations study. So, findings about performance of different methods to estimate propensity score for matching method are supported to understand whether or not propensity score trimming has the necessary qualifications to achieve balance in covariates.

email: hk20902@uga.edu

10b. Performance of Instrumental Variable and Mendelian Randomization Estimators for Count Data

Phillip Allman*, University of Alabama at Birmingham
Hemant Tiwari, University of Alabama at Birmingham
Inmaculada Aban, University of Alabama at Birmingham
Dustin Long, University of Alabama at Birmingham
Todd MacKenzie, Dartmouth College
Gary Cutter, University of Alabama at Birmingham

Instrumental variable (IV) methods are causal inference techniques to mitigate unobserved confounding; often in observational data. When some risk factor and outcome of interest are confounded, we may use IV methodology to assess causality in certain scenarios. A valid IV is to observational data what the random treatment group assignment is to clinical trials. If a valid IV is found, then these are the only known methods to account for unobserved confounders in observational data. Originating in econometrics, IV methods have recently been adopted by epidemiology and genetics. Genetic variants in particular provide promising sources of potential IVs. The use of genetic variants as IVs has been dubbed Mendelian randomization (MR). Numerous statistical methods have been proposed for IV or MR analyses. The practical performance of these methods is well characterized for normally distributed data; but much less so for count data. This project implements Monte Carlo simulations to assess bias and variance in causal effect estimates obtained from a variety of MR methods when the data are count variates. These MR methods include several existing and one novel method.

email: allman@uab.edu

10c. Improve Power Analysis in Clinical Trials with Multiple Primary Endpoints: An Application of Parametric Graphical Approaches to Multiple Comparison

Zhe Chen*, Biogen
Ih Chang, Biogen

The motivation of this work is from the design of a randomized, placebo-controlled, multiple-dose clinical trial with the objective to demonstrate the pharmacodynamic (PD) efficacy of a novel molecule for treating the neuropathic pain. The study has multiple primary neurophysiology endpoints. The PD efficacy will be considered as demonstrated if a statistically significant effect is observed in all of those endpoints in any of the treatment dose groups as compared to placebo. To power the trial, a crude way is to assume the endpoints are independent and then test each of them by making the pairwise comparison between each treatment dose group and placebo, with the multiplicity adjusted by the Dunnett’s method. The limitations of this approach are the conservativeness of controlling the type I error and the unrealistic among-endpoint independence assumption, which can make the power analysis sub-optimal. Here we investigate how to address this issue to improve the power analysis in the trial design by using the graphical approaches to the parametric multiple testing and present the simulation results.

email: vincent.chen@biogen.com

ABSTRACTS & POSTER PRESENTATIONS

10d. Two-Stage Randomized Trial for Testing Treatment Effect for Time to Event Data

Rouba A. Chahine*, University of Alabama at Birmingham
Inmaculada Aban, University of Alabama at Birmingham
Dustin Long, University of Alabama at Birmingham

A standard component of randomized clinical trials is informed consent, in which participants are made aware of all procedures or treatments. This knowledge may decrease compliance or reduce tolerance for inconveniences or difficulties when participants receive non-preferred treatments. A 2-stage randomized design incorporates participants preference. First, participants are randomized to the random (RT) or choice (CT) group. Second, RT group participants are randomized to one of two treatments; and those in the CT group choose between treatments. Current power analysis methods to calculate appropriate sample size to estimate effects of treatment, selection and preference are based on the ANOVA approach, which assumes asymptotic normality of test statistics. We used simulations to investigate performance of current methods for time-to-event outcomes following the Exponential and Weibull distributions, for complete-case and right-truncated data. Results indicate the ANOVA approach performs well with non-censored Exponential data, but not with Weibull and right-truncated Exponential data, particularly if there is an imbalance in treatment choice in the CT group.

email: chahine@uab.edu

10e. Estimating Power for Clinical Trials with Patient Reported Outcomes Endpoints using Item Response Theory

Jinxiang Hu*, University of Kansas Medical Center
Yu Wang, University of Kansas Medical Center

Background: Patient reported outcomes (PRO) are important in patient-centered health outcomes research and quality of life (QOL) studies. Item Response Theory (IRT) with latent scores is used to develop the Patient Reported Outcomes Measurement Information System (PROMIS). However, Classical Test Theory (CTT) using observed scores are routinely used to estimate power for clinical trials with PRO endpoints. The purpose of this paper is to compare power results between CTT and IRT approaches. Methods: Motivated from PROMIS depression scales (4, 6, 8 items) we conducted a simulation study in order to estimate power differences between IRT- and CTT-based scoring for a two- armed prospective randomized clinical trial (control vs active arm). We simulated data using various sample size, allocation ratio, number of items, effect sizes. Three models were fit to each simulation: IRT with MLE, IRT with Bayesian estimator, and CTT. Results and conclusion: Our results showed sample size, effect size, and number of items are important indicators of IRT power with PRO as endpoints. CTT overestimates power when sample size and effect size are small.

email: jhu2@kumc.edu

10f. Bayesian Multi-Regional Clinical Trials Using Model Averaging

Nathan W. Bean*, University of North Carolina, Chapel Hill
Matthew A. Psioda, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Multi-regional clinical trials (MRCTs) provide the benefit of rapidly introducing drugs to the global market, however, current statistical methods pose limitations to the control of information sharing and estimation of regional treatment effects. With the recent publication of the ICH E17 guideline in 2017, the MRCT design is recognized as a viable strategy that can be accepted by regional regulatory authorities, necessitating new statistical methods that overcome the challenge of information sharing. We develop novel methodology for estimating regional and global treatment effects from MRCTs using Bayesian model averaging. This approach can be used for trials that compare two treatment groups with respect to a continuous outcome, and it allows for the incorporation of patient characteristics through the inclusion of covariates. Posterior model probabilities provide a natural assessment of consistency between regions that can be used by regulatory authorities for drug approval. We compare our method to existing methods, including linear regression with common treatment effect and a Bayesian hierarchical random effects model, and the results from simulations are presented.

email: nbean@live.unc.edu

10g. Constructing Causal Methylation Network by Additive Noise Model

Shudi Li*, University of Texas Health Science Center at Houston
Rong Jiao, University of Texas Health Science Center at Houston
Momiao Xiong, University of Texas Health Science Center at Houston

Alzheimer's disease (AD) is a chronic neurodegenerative disease that causes problems with memory, thinking and behavior. Alzheimer's Disease Neuroimaging Initiative (ADNI) is a multi-study that aim to improve clinical trials for the prevention and treatment of Alzheimer's disease (AD). Methylation plays an important role in the development of AD. To uncover how the DNA methylation affect development of AD, we first conduct genome-wide causation studies of methylation for AD using additive noise models (ANMs) to identify methylated genes that cause AD. Then, we infer methylation causal networks using the identified methylation causing genes across three time points: baseline, 12 months, and 24 months. Finally, we construct the methylation causal networks that influence AD at three time points and look at the progression of the networks. The developed methods are applied to ADNI methylation dataset with 423 samples and 866,840 CpG cites.

email: shudili2015@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

10h. Detecting Intervention Effects in a Randomized Trial within a Social Network

Shaina J. Alexandria*, University of North Carolina, Chapel Hill
Michael G. Hudgens, University of North Carolina, Chapel Hill
Allison E. Aiello, University of North Carolina, Chapel Hill

Studies within social networks provide opportunities to study effects of population-level interventions that may reach more of the population than was intervened on directly. Such spillover effects will exist in the presence of interference, i.e., when one subject's treatment can affect another subject's outcome. Randomization-based inference (RI) methods provide a theoretical basis for inference in randomized studies, even in the presence of interference. Here we consider constructing RI tests and confidence intervals for analysis of data from the eX-FLU trial, a randomized study designed to assess the effect of a social isolation intervention on influenza-like-illness transmission in a connected network of college students. Existing and proposed methods for hypothesis testing and confidence interval construction are compared empirically via simulation studies. The different approaches are then applied to the eX-FLU trial.

email: shaina.mitchell22@gmail.com

11. POSTERS: GENOMICS/PROTEOMICS

11a. Kernel-Based Genetic Association Analysis for Microbiome Phenotypes

Hongjiao Liu*, University of Washington
Michael C. Wu, Fred Hutchinson Cancer Research Center

Understanding human genetic influences on the microbiome offers clues as to the mechanisms by which genetics influences traits. To assess the relationship between host genetic variation and microbiome composition, we consider using the kernel RV coefficient (KRV) to evaluate the association between groups of SNPs at gene level and microbiome composition at community level. In KRV test, kernel matrices are constructed from genetic and microbiome data, respectively, to describe the genotypic and microbiome similarity between pairwise subjects. Such analyses have been shown to improve power. However, in genetic association studies, it is important to also control for covariates and confounders. Due to the restriction to discrete inputs for microbiome-based kernels, it is not feasible to adjust for covariates on raw microbiome data in the KRV framework. Therefore, we propose to perform a kernel PCA on the microbiome kernel and adjust the covariates on the derived finite sample basis. Simulation studies show that this approach effectively controls type I error while maintaining power. We demonstrate the covariate-adjusted KRV test in a real microbiome-host genetics study.

email: liuhj1224@gmail.com

11b. True Source of Inflated Zeros in Single Cell Transcriptomics

Tae Kim*, University of Chicago
Mengjie Chen, University of Chicago

Single-cell transcriptomic data generated from 10X protocol has been proven difficult to properly analyze. One of the main challenges is to separate the biological signal from the technical noise. Contrary to the traditional belief that the difficulty lies in the excessive technical noise from different sources, we argue that the technical noise for most genes follows a Poisson distribution without any over-dispersion after proper pre-processing. We show a new perspective of analyzing the single-cell data through examples when the cell population is truly homogeneous or heterogeneous, respectively. We then propose a novel method of sequential clustering to reflect our prior knowledge that certain genes are more heterogeneous than others. By allowing each gene to have a different number of clusters, this method not only offers a method for multi-level cell type clustering but also provides a natural representation of cell lineage and trajectory inference.

email: taehyun0313@gmail.com

11c. Estimating Cell Type Composition Using Isoform-Level Gene Expression Data

Hillary M. Heiling*, University of North Carolina, Chapel Hill
Douglas R. Wilson, University of North Carolina, Chapel Hill
Wei Sun, Fred Hutchinson Cancer Research Center
Naim Rashid, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Human tissue samples are often mixtures of heterogeneous cell types, which can confound gene expression data analyses derived from such tissues. The cell type composition of a tissue sample may be of interest itself and is needed for proper analysis of differential expression. A variety of deconvolution statistical methods have been proposed to address this issue, but most of the proposed methods are designed to analyze gene-level expression only. However, RNA isoforms can also be differentially expressed across cell types, and they can be more informative for cell type origin than gene expression. We propose an isoform deconvolution method, IsoDeconvMM, that models the exon-set read counts within genes with a multinomial structure and the gene and isoform expression parameters with Dirichlet distributions. IsoDeconvMM demonstrates good performance in simulations when tested against a range of the following factors: variability in isoform expression across subjects, number of purified reference samples per cell type, and the number of genes utilized for estimation. We also illustrate the utility of our method with a real data analysis.

email: hheiling@live.unc.edu

ABSTRACTS & POSTER PRESENTATIONS

11d. EWAS of Kidney Function Identifies Shared and Ethnic-Specific Loci

Anna Batorsky*, University of North Carolina, Chapel Hill
Mi Kyeong Lee, National Institute of Environmental Health Sciences, National Institutes of Health
Stephanie J. London, National Institute of Environmental Health Sciences, National Institutes of Health
Josyf C. Mychaleckyj, University of Virginia
Andrew Marron, University of North Carolina, Chapel Hill
Eric A. Whitsel, University of North Carolina, Chapel Hill
Nora Franceschini, University of North Carolina, Chapel Hill
Charles E. Breeze, Altius Institute for Biomedical Sciences & University College London

We perform an epigenome-wide association study to examine associations between eGFR (estimated glomerular filtration rate, a quantitative measure of kidney function) and DNA methylation (DNAm) patterns in whole blood from 3049 participants (1058 African, 585 Hispanic, 1406 European ancestry) from the Women's Health Initiative. DNAm beta values, obtained using Illumina 450K beadchip arrays, were used in ethnicity-stratified linear models with robust standard errors, adjusting for age, smoking, technical and study-specific variables. Meta-analyzed results across ethnicities show 12 differentially methylated positions (DMPs) at a genome-wide significance level of p<10-7 (Bonferroni). Functional annotation (eFORGE) of top eGFR-associated DMPs across ethnicities shows associations with chromatin accessibility in developmental cell types, and with transcription factor binding motifs strongly implicated in developmental pathways. Dysregulation of these sites may mediate disease risk. Further examination of ethnic-specific DMPs may elucidate differences in kidney disease risk across ancestries.

email: abatorsk@live.unc.edu

11e. Deconvolutional Mixture Modeling to Account for Cell Type Composition in Tissue Samples

Zachary P. Brehm*, University of Rochester
Marc K. Halushka, Johns Hopkins University
Matthew N. McCall, University of Rochester

Tissue samples are heterogeneous mixtures of numerous cell types exhibiting unique functions and expression signatures. In analyses of tissue level gene expression data, it is uncertain whether differences between samples should be attributed to changes in cellular function or composition. Deconvolution methods allow us to control for this uncertainty by estimating the proportions of component cell types and their contribution to tissue level gene expression. We estimate cell type specific gene expression profiles from publicly available sequencing data of experimentally purified cell types in order to develop a preliminary deconvolutional mixture model that accounts for compositional differences in tissue. Specifically, we curate a unified dataset containing cell type specific mRNA profiles from samples that are representative of the components of coronary artery tissue in order to deconvolute coronary artery sequencing data provided by GTEx. We present a preliminary deconvolution of these GTEx data alongside comparisons with existing methods. Using histology images provided by GTEx, we assess the performance of these methods in detecting atherosclerotic tissue.

email: zachary_brehm@urmc.rochester.edu

11f. Developing a Computational Framework for Precise TAD Boundary Prediction using Genomic Elements

Spiro C. Stilianoudakis*, Virginia Commonwealth University
Shumei Sun, Virginia Commonwealth University

Chromosome conformation capture combined with high-throughput sequencing experiments (Hi-C) have revealed that chromatin undergoes layers of compaction through DNA looping and folding, forming dynamic 3D structures. Among these are Topologically Associating Domains (TADs), which are known to play critical roles in cell dynamics like gene regulation and cell differentiation. Precise TAD mapping remains difficult, as it is strongly reliant on Hi-C data resolution. Obtaining genome-wide chromatin interactions at high-resolution is costly resulting in variability in true TAD boundary location by TAD calling algorithms. To aid in the precise identification of TAD boundaries we developed a computational framework that leverages the spatial relationship of many high-resolution ChIP-seq defined genomic elements. Our framework precisely predicts chromosome-specific TAD boundaries on multiple cell types. We show that known molecular drivers of 3D chromatin including CTCF, RAD21, and SMC3 are more enriched at our predicted TAD boundaries compared to the boundaries identified by the popular ARROWHEAD TAD caller. Our results provide useful insights into the 3D organization of the human genome.

email: stilianoudasc@vcu.edu

ABSTRACTS & POSTER PRESENTATIONS

11g. Parsing Latent Factors in High-Dimensional Classification on Genomic Data

Yujia Pan*, University of Michigan
Johann Gagnon-Bartsch, University of Michigan

Classification of high-throughput biological data is challenging because the signal is often weak and sparse. Incorporating additional covariates (e.g. smoking status, age) can yield better predictive accuracy; however, it is often the case that such information is unobserved. To this end, we introduce a classifier which adaptively leverages both observed variables as well as inferred latent ones. Including these latent variables tends to improve accuracy, sometimes substantially, as illustrated on several simulated and real genomic datasets. Our method is broadly applicable to many different types of high-throughput biological data, as illustrated on a diverse collection of microarray, methylation, and sequencing datasets covering a wide range of disease phenotypes.

email: yujiap@umich.edu

11h. Estimation of Metabolomic Networks with Gaussian Graphical Models

Katherine Hoff Shutta*, University of Massachusetts, Amherst
Subhajit Naskar, University of Massachusetts, Amherst
Kathryn M. Rexrode, Harvard Medical School
Denise M. Scholtens, Northwestern University
Raji Balasubramanian, University of Massachusetts, Amherst

Network-based metabolomic analyses have high potential to capture signatures of complex biological processes. Gaussian graphical model (GGM) estimation is one approach for such analyses. Modeling metabolomic data with GGMs enables biologically meaningful interpretation of the estimated edge set as a map of functional dependence between metabolites, conditioned on biological state. GGM estimation is an active area of research; several open-source R packages have been developed for this purpose. GGM estimation involves several choices of scoring criteria and estimation algorithms. Our studies suggest that the estimated GGM may be highly sensitive to these choices, and the effectiveness of each method may depend on structural characteristics of the true network. We characterize these variations by simulating data from networks of scale-free, random, and mixture topologies. We assess the sensitivity and specificity of each method in recovering the original network and discuss implications in the practical context of drawing conclusions from estimated networks. We illustrate the approaches on data from a cardiovascular disease metabolomics study nested in the CATHGEN repository.

email: kshutta@umass.edu

11i. Weighted Kernel Method for Integrative Metabolomic and Metagenomic Pathway Analysis

Angela Zhang*, University of Washington
Michael C. Wu, Fred Hutchinson Cancer Research Center

Dysbiosis of the microbiome can lead to abnormal levels of microbe-produced metabolites, which has been linked to a variety of diseases and conditions. Innovations in high-throughput technology now allow rapid profiling of the metabolome and metagenome – the gene content of the bacteria – for characterizing microbial metabolism. Due to the small sample sizes and high-dimensionality of the data, pathway analysis (wherein the effect of multiple genes or metabolites on an outcome is cumulatively assessed) of metabolomic data is commonly conducted and also represents the standard for metagenomic analysis. However, how to integrate both data types remains unclear. Recognizing that a metabolic pathway can be complementarily characterized by both metagenomics and metabolomics, we propose a weighted kernel framework to test if the joint effect of genes and metabolites in a biological pathway is associated with outcomes. The approach allows analytic p-value calculation, correlation between data types, and optimal weighting. Simulations show that our approach often outperforms other strategies. The approach is illustrated on real data.

email: azhang2@uw.edu

---

12. POSTERS: FUNCTIONAL DATA/HIGH DIMENSIONAL

12a. Dimension Reduction Methods for Multilevel Neural Firing Rate Data

Angel Garcia de la Garza*, Columbia University
Jeff Goldsmith, Columbia University

Recent advances have allowed high-resolution observations of firing rates for a collection of individual neurons; these observations can provide insights into patterns of brain activation during the execution of tasks. Our data come from an experiment in which mice performed a reaching motion following an auditory cue, and contain measurements on firing rates from neuron in the motor cortex before and after the cue. We focus on appropriate dimension reduction techniques for this setting, in which sharp increases in firing rates after the cue are expected; we also allow for correlation across neurons in each of the trials, and within a neuron across trials. Initial results suggest differing patterns of activation, perhaps representing the involvement of different motor cortex functions at different times in the reaching motion.

email: agarciadelagarza@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

12b. Amplitude-Phase Separation of Trace-Variogram and its Applications in Spatial Functional Data Analysis

Xiaohan Guo*, The Ohio State University
Sebastian Kurtek, The Ohio State University
Karthik Bharath, University of Nottingham

We propose a novel method for statistical analysis of functional data with spatial correlation that considers amplitude and phase of the functions as separate data objects. Methods developed for spatial functional data during the past decade use the trace-variogram based on the L2-norm as a core statistic that captures the spatial correlation between functions; this statistic is then applied for different tasks including clustering and kriging. In contrast, we propose a new notion of the trace-variogram that uses amplitude-phase separation in a dataset of functions. This allows us to capture the spatial correlation of the amplitude and phase components in functional data separately. We show the utility of the proposed amplitude-phase trace-variograms in kriging and clustering of spatial functional data. We assess the proposed approach using simulations and real data examples.

email: guo.1280@osu.edu

12c. Free-Living Walking Strides Segmentation in Wrist-Worn Accelerometry Data

Marta Karas*, Johns Hopkins Bloomberg School of Public Health
Ryan T. Roemmich, Johns Hopkins School of Medicine
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health
Jacek K. Urbanek, Johns Hopkins School of Medicine

Wearable accelerometry data have been widely used to estimate characteristics of in-the-lab walking that are known to be related to human health. For example, gait speed is associated with survival, neurodegenerative disorders, stroke recovery, and obesity. Further studies suggest there might be even more information in free-living, accelerometry-derived patterns of walking. However, obtaining these patterns requires a well-crafted analytical approach for the detection and identification of walking events, which is especially challenging in noisy, wrist-accelerometry data. To the best of our knowledge, there are no open-source, well-documented and validated methods available for precise, high-specificity segmentation of walking in such data. To address that, we build upon adaptive empirical pattern transformation (ADEPT) - an existing method for pattern segmentation in high-frequency data - to detect accelerometry patterns that resemble walking strides, and then fine-tune the results with a binary smoother. We validate our approach with data collected on 6 participants who wore 5 sensors simultaneously (2x wrist, 2x ankle, lower back) in in-the-lab and free-living settings.

email: mkaras2@jhu.edu

12d. Variable Selections for High-Dimensional Unsupervised Learning with Applications in Genomics and Regulatory Pathway Analysis

Zhipeng Wang*, Rice University
David Scott, Rice University

In this work, we focus on a newly-developed unsupervised machine learning algorithms and investigate their applications in genomics and regulatory pathway analysis. The algorithm is based on the idea of variable selection for unsupervised learning and density-based clustering. We developed a statistical metric called Projection Density Score (PDS), which quantifies how important a subset of variables is for the clustering task from the information theory perspective. The PDS is derived from the shape of the density function (e.g. number of modes, disconnections between modes etc.) and the level of locality preserving in the projected subspace. We can perform dimension reduction by picking the subset of variables with high PDS, then perform a variety of clustering algorithms (K-means, spectral clustering, tree-based clustering algorithms etc.) in the low-dimensional subspace. Our method increases the interpretability of the dimension reduction, thus is particularly useful in genomics research. We applied our algorithm to a variety of real-world gene expression datasets and achieved good performance of identifying important set of genes for pathway analysis.

email: Zhipeng.Wang@alumni.rice.edu

12e. Integrative Analysis and Prediction Method for Identifying Subgroup-Specific Omics Biomarkers

Jessica Butts*, University of Minnesota
Sandra Safo, University of Minnesota

In many settings, it is desirable to combine information from several different sources of data such as genomics, proteomics, and metabolomics in addition to clinical data. Integrating multiple data types is valuable because it allows us to use all of the available information simultaneously. Another important consideration is subgroup heterogeneity because there may be variables important in predicting the outcome of interest in some subgroups but not others. For example, different subgroups may have similar symptoms but different disease courses and may respond to treatment differently. Thus, considering subgroup heterogeneity in an analysis is critical. We introduce a novel statistical method that capitalizes on the benefits of both integrative analysis and consideration of subgroup heterogeneity in prediction models. By using latent models with hierarchical penalties, we propose integrative analysis and prediction methods to select subgroup-specific biomarkers predictive of clinical outcomes. The utility of the proposed method will be explored with extensive simulations and real-world data applications.

email: butts029@umn.edu

ABSTRACTS & POSTER PRESENTATIONS

12f. A Novel FWER Controlling Procedure for Data with Generalized Reduced Rank Correlation Structure

Jiatong Sui*, University of Rochester
Xing Qiu, University of Rochester

Different types of multiple testing procedures have been proposed in order to overcome the low power issue while controlling the family-wise error rate for data with arbitrary dependence structures. However, none of them has produced a balanced solution among type I error, power, stability and computational efficiency under general assumption of dependence. In this study, we have proposed a new method that successfully controls the FWER and produces powerful results under general correlation structure, without any heavy computational burden. Our proposed method, rrMTP, is constructed based on the foundation of the decomposition of the correlation matrix using the generalized reduced rank structure model. The distribution theory for the extreme spherical distance has been utilized to appropriately approximate the null distribution of the most extreme test statistics. We also account for the noise part so that rrMTP is robust to model misspecifications. The proposed method is flexible and stable for a wide range of real data applications with different sample sizes and number of hypotheses. Our results are demonstrated by simulation studies and some real data applications.

email: sui930703@hotmail.com

12g. Analyzing Accelerometer Data with Probability Magnitude Graphs

Margaret Banker*, University of Michigan
Peter X.K. Song, University of Michigan

Accelerometry data is a promising avenue to provide individual-specific data that can be beneficial in precision health frameworks. Other studies analyze accelerometry data by utilizing summary statistics (eg single-axis count data) and applying regression-based cutoff thresholds to determine Physical Activity Levels (i.e. classify activity levels as Vigorous, Moderate, Light, and Sedentary). However, these cutoffs are often not generalizable across populations, devices, or studies. Thus, a more functional-data approach would be useful in analyzing accelerometer data. We consider at subject's probability magnitude graphs to holistically represent their activity profile; these "probability magnitude graphs" measure what percentage of an individual's time is spent at or below a continuum of activity levels. Using a fused lasso approach, we use these functional activity curves as covariates to determine what patterns of activity affect health outcomes, such as BMI and other anthropometric body measurements. We also consider these curves on an hour-by-hour basis in order to maintain the important time-specific information.

email: mbanker@umich.edu

12h. Normalization of Minute-Level Activity Counts from Chest- and Wrist-Worn Accelerometers: An Example of Actiheart, Actiwatch, and Actigraph

Vadim Zipunnikov*, Johns Hopkins University
Jiawei Bai, Johns Hopkins Bloomberg School of Public Health

Wearables provide a much more complete view of physical activity, sleep, and circadian rhythmicity and their association with health and disease. However, different studies use different devices placed at different parts of the body. Thus, there is a critical need of harmonizing data across clinical and population samples to facilitate more comprehensive cross-study analyses. We will present three novel normalizations including global, local, and latent normalizations that efficiently addresses this challenge. The main idea is to use Semiparametric Gaussian Copula approach that enforces a normalization invariant up to any monotone transformation and takes into account the temporal correlation over 24-hour period. We demonstrate the approach using data from 3 devices simultaneously worn by 624 participants of Baltimore Longitudinal Study of Aging.

email: vadim.zipunnikov@gmail.com

13. POSTERS: BAYESIAN, CLUSTERED DATA, HYPOTHESIS TESTING

13a. Bayesian Mechanism for Categorical Data with Data Augmentation Strategy

Arinjita Bhattacharyya*, University of Louisville
Subhadip Pal, University of Louisville
Riten Mitra, University of Louisville
Shesh N. Rai, University of Louisville

Bayesian shrinkage priors have emerged as a popular and flexible method of variable selection in regression settings. Here we present a Bayesian hierarchical framework that can accommodate regression of categorical responses with many variable selection priors like Horseshoe, Dirichlet Laplace, and Double Pareto. The key feature of our approach is a representation of the likelihood using a Polya-gamma data augmentation approach that leads to a very general Gibbs scheme and hence can be adapted for several prior settings. To assess performance, we considered various measures of predictive accuracies including AUC and Brier loss. Using these measures, we conducted extensive simulation studies to compare these priors under different settings of sample sizes, parameter values, and covariate dimensions. We validated the performance of our method in a data set.

email: a0bhat09@louisville.edu

ABSTRACTS & POSTER PRESENTATIONS

13b. False Coverage Rate-Adjusted Smoothed Bootstrap Simultaneous Confidence Intervals for Selected Parameters

Jing Sun*, Augusta University
Santu Ghosh, Augusta University
Varghese George, Augusta University

Benjamini and Yekutieli (2005) proposed the false coverage-statement rate (FCR) method for multiplicity correction when constructing confidence intervals for only selected parameters. We suggest a novel procedure to construct simultaneous confidence intervals for the selected parameters by using a smoothed bootstrap procedure to control the FCR. A pertinent problem associated with the smoothed bootstrap approach is how to choose the unknown bandwidth in some optimal sense. We derive an optimal choice for the bandwidth and the resulting smoothed bootstrap confidence intervals asymptotically in order to give better control of the FCR than its competitors. Finite sample performances of our method are illustrated based on empirical studies. Through these empirical studies, it is shown that the proposed method can be successfully applied in practice.

email: sun60527@gmail.com

13c. A State-Space Approach in Handling Challenges Associated with Longitudinal Continuous Neuropsychological Outcomes

Alicia S. Chua*, Boston University School of Public Health
Yorghos Tripodis, Boston University School of Public Health & Boston University School of Medicine

Neuropsychological assessments are important in evaluating the disease state of patients with cognitive issues. The distribution of these outcomes is often skewed and does not exhibit a linear trajectory. We introduced the adjusted local linear trend model to handle these challenges. The model involves two equations - measurement and state. Covariates of interest are estimated via the state equation with the flexibility to adjust for the time elapsed between visits via a transition matrix. This model has a maximum likelihood estimation step for the unknown variances prior to feeding the input values into the recursive Kalman Filter and Kalman Smoother algorithms to obtain unbiased model estimates. When data from 297 subjects who completed the Boston Naming Test were used in our simulation study, our proposed model appeared to perform just as well as the linear mixed-effects models within the equally spaced-time intervals analysis but the model was superior to the mixed-effects models in the unequally spaced-time intervals analysis. Our proposed model was able to attain the lowest variance for the estimates versus other models compared (<0.09 and <0.13, respectively).

email: aschua@bu.edu

13d. Combining Dependent P-values with a Quantile-Based Approach

Yu Gu*, University of Rochester
Michael P. McDermott, University of Rochester
Xing Qiu, University of Rochester

In many applications, practitioners need to combine a large number of dependent p-values. These observed p-values may contain outliers due to technical issues. We proposed a novel p-value combination method based on the quantiles of the transformed p-values. It is robust to outliers and capable of incorporating the correlation structure of the data. We derived the asymptotic distribution of overall test statistic, the theoretical type I error and statistical power of our test. We also implemented a method to find the optimal quantile based on prior information about the sample. In simulation studies, our method was compared with several competing methods, including Fisher's method, Stouffer's method, Tippet's method and Kost & McDermott's method for combining dependent p-values. We showed that even a very small correlation can have profound effects on p-value combination tests, and our proposed method was the only one that controlled the type I error at the nominal level, when outliers were present and the p-values were dependent. Furthermore, this robustness did not come with significant loss of statistical power in classical situations.

email: yu_gu@urmc.rochester.edu

13e. Bayesian Estimation for Parameters of Nonlinear Multilevel Models under Burr Distributions

Mohan D. Pant*, Eastern Virginia Medical School
Ismail E. Moudden, Eastern Virginia Medical School
Jiangtao Luo, Eastern Virginia Medical School

Data in healthcare analytics are often multilevel, e.g., patients nested within hospitals. The multilevel modeling approaches, used for modeling such data, are based on both patient- and hospital-level covariates. Most of these modeling approaches assume normal distribution for the outcome variable and error terms. However, several of the patient- and hospital-level covariates (e.g., length of stay, hospital charges, etc.) are known to follow nonnormal distributions. Thus, multilevel modeling under the assumption of normal distribution may not accommodate variables with outlying observations and is prone to making inaccurate inferences. Therefore, we consider a class of nonlinear multilevel models where the outcome variable and error terms are assumed to follow Burr distributions, a family of distributions suitable for modeling nonnormal data. Bayesian framework utilizing No-U-Turn-Sampler (NUTS) is used for the estimation of parameters associated with the proposed models. We provide model comparison using different criteria and contrast results from our models with that based on the normal assumption, using both simulated and real-world data.

email: pantmd@Evms.edu

ABSTRACTS & POSTER PRESENTATIONS

13f. A Flexible and Nearly Optimal Sequential Testing Approach to Randomized Testing: QUICK-STOP

Julian Erik Hecker*, Brigham and Women's Hospital and Harvard Medical School
Ingo Ruczinski, Johns Hopkins Bloomberg School of Public Health
Michael M. Cho, Brigham and Women's Hospital and Harvard Medical School
Edwin Silverman, Brigham and Women's Hospital and Harvard Medical School
Brent Coull, Harvard T.H. Chan School of Public Health
Christoph Lange, Harvard T.H. Chan School of Public Health

In the analysis of current life science datasets, we often encounter scenarios in which the application of asymptotic theory to hypothesis testing can be problematic. Besides improved asymptotic results, permutation/simulation-based tests are a general approach to address this issue. However, these randomized tests can impose a massive computational burden, e.g., in scenarios in which large numbers of statistical tests are computed, and the specified significance level is very small. Stopping rules aim to assess significance with the smallest possible number of draws while controlling the probabilities of errors due to statistical uncertainty. In this communication, we derive a general stopping rule, QUICK-STOP, based on the sequential testing theory that is easy to implement, controls the error probabilities rigorously, and is nearly optimal in terms of expected draws. In a simulation study, we show that our approach outperforms current stopping approaches for general randomized tests by factor 10 and does not impose an additional computational burden.

email: hecker.julian@gmail.com

13g. A Weighted Jackknife Approach Using Linear Model-Based Estimates for Clustered Data

Yejin Choi*, University of New Mexico
Ruofei Du, University of New Mexico Comprehensive Cancer Center

For cluster data analysis, standard statistical analysis methods (e.g. linear mixed-effects modeling) require the assumption that the number of clusters collected is sufficiently large. However, it has been realized that such an assumption rarely occurs in real-world problems. Additionally, it is common to observe heterogeneity across the clusters, which leads the inference less stable. Thus, we propose a weighted delete-one-cluster Jackknife-based framework for the influence of each cluster to be balanced to get more reliable inference. In this scheme, a cluster affected by the heterogeneity is weighted less in estimating the mean of the outcome variable for a study condition. Compared with a published approach for the same question setting that working with the mean outcome values of the clusters, the proposed framework employs Least Squares or Generalized Least Squares estimators involving the data at individual level directly. It is shown by simulation studies that the proposed framework produces more precise mean estimates, and is more accurate for statistical testing with respect to its power than other statistical methods.

email: yjchoi@unm.edu

14. POSTERS: HIGH-DIMENSIONAL DATA, MISSING DATA AND MORE

14a. Predicting Latent Contacts from Self-Reported Social Network Data via Outcome Misclassification Adjustment

Qiong Wu*, University of Maryland
Tianzhou Ma, University of Maryland
Shuo Chen, University of Maryland

Social network has become a valuable tool to study the role of the contact patterns in epidemics. However, social networks in epidemiological research are often estimated based on self-reported data and study subjects may not report all their contacts in survey questionnaires. The unreported links can lead to partially observed social networks instead of the complete social networks which gives rise to biased estimates of the effects of risk factors. To address this challenge, we propose a parametric social network model to predict latent contacts. We show that the social network model with latent links under the assumption of conditional independence is related to the logistic regression with misclassified outcomes. We develop new algorithms to optimize model parameters and yield robust and accurate prediction of latent links without validation data. We perform extensive simulation studies and apply the method to a partially observed social network data and incomplete brain network data. The results demonstrate the improved performance of latent-contact prediction comparing with existing machine learning and nonparametric models for partially observed social networks.

email: qwu1221@terpmail.umd.edu

14b. Validate Surrogate Endpoints with Continuous and Survival Setup

Idris Demirsoy*, Florida State University
Helen Li, Regeneron Pharmaceutical

Surrogate endpoint is a biomarker which is planned to take the place of clinical outcomes. For example, controlling blood pressure to prevent stroke. As is in this example, studying stroke might take very long time. Biomarkers use in clinical trials and drug development, using surrogate endpoint will save time to get result quicker, will save time and money. However, before replacing true endpoint with surrogate endpoint, we need to study and show that surrogate is validated to use. As Fleming (1996) said, "a correlate does not a surrogate make." However, correlation is necessary for surrogacy but how to evaluate perfectly correlated surrogate endpoint whether or not validated surrogate endpoint. We have simulated continuous true end point and survival time surrogate end point for different correlation such as 0.5 and 0.8. We have extended Buyse (2000)'s copula method trial level and individual level measure check however, we realized that we need to add a restriction to this method before applying to data and show that higher correlation necessary to show validation of surrogacy but checking mean difference and hazard ratio difference is mandatory.

email: idris.demirsoy@fsu.edu

ABSTRACTS & POSTER PRESENTATIONS

14c. New Two-Step Test for Mediation Analysis with Sets of Biomarkers

Andriy Derkach*, Memorial Sloan Kettering Cancer Center
Joshua Sampson, National Cancer Institute,
National Institutes of Health
Simina Boca, Georgetown University Medical Center

We consider the scenario where there is an exposure, multiple biologically-defined sets of biomarkers, and an outcome. We propose a new two-step procedure that tests if any of the sets of biomarkers mediate the exposure/outcome relationship, while maintaining a prespecified Family-Wise Error Rate. The first step of the proposed procedure is a screening step that removes all groups that are unlikely to be strongly associated with both the exposure and the outcome. The second step conducts post-selection inference to test if there are true mediators in each of the remaining, candidate sets. Our simulation results show that this simple two-step procedure has higher statistical power to detect true mediating sets when compared with existing procedures. We then use our two-step procedure to identify a set of Lysine-related metabolites that potentially mediate the known relationship between increased BMI and the increased risk of ER+ breast cancer in post-menopausal women.

email: derkacha@mskcc.org

14d. Meta-Analysis of Binary or Continuous Outcomes Combining Individual Patient Data and Aggregate Data

Neha Agarwala*, University of Maryland, Baltimore County
Anindya Roy, University of Maryland, Baltimore County

Often both aggregate or meta-analysis (MA) studies and Individual Patient Data (IPD) studies are available for specific treatments. Combining these two sources of data could improve the overall meta-analytic estimates of treatment effects. We propose a method to combine treatment effects across trials where the response can be binary or continuous. For some studies with MA data, the associated IPD maybe available, albeit at some extra effort or cost to the analyst. We consider the case when treatment effects are fixed and common across studies and evaluate the wisdom of choosing MA when IPD is available by studying the relative efficiency of analyzing all IPD studies versus combining various percentages of MA and IPD studies. For many different model design constraints under which the MA estimators are the IPD estimators, and hence fully efficient, are known. For such models we advocate a selection procedure that chooses MA studies over IPD studies in a manner that force least departure from design constraints and hence ensures a fully efficient combined MA and IPD estimator.

email: agneha1@umbc.edu

14e. A Post-Processing Algorithm for Building Longitudinal Medication Dose Data from Extracted Medication Information Using Natural Language Processing from Electronic Health Records

Elizabeth McNeer*, Vanderbilt University Medical Center
Cole Beck, Vanderbilt University Medical Center
Hannah L. Weeks, Vanderbilt University Medical Center
Michael L. Williams, Vanderbilt University Medical Center
Nathan T. James, Vanderbilt University Medical Center
Leena Choi, Vanderbilt University Medical Center

Extracting detailed longitudinal medication dose data from electronic health records is challenging and may require a specialized program such as a natural language processing system. However, the raw output from many of these systems is not directly usable and must be converted to a more appropriate form for analysis. We have developed a post-processing algorithm to address this issue. The algorithm consists of two parts: Part I parses the raw output and connects entities together so that they can be used to calculate appropriate dose amount in the next step, and Part II removes redundant data and calculates dose amount taken at a given time and daily dose. We used the output from two natural language processing systems, MedXN and medExtractR, to develop the algorithm, and we tested the algorithm using gold standard datasets for two test drugs, tacrolimus and lamotrigine, which have different prescribing patterns. Compared to gold standard datasets, both parts of the algorithm performed well (all F-measures > 0.9).

email: elizabeth.mcneer@vumc.org

14f. Power and Sample Size Analysis using Various Statistical Methods in a Tumor Xenograft Study

Sheau-Chiann Chen*, Vanderbilt University Medical Center
Gregory D. Ayers, Vanderbilt University Medical Center
Rebecca L. Shattuck-Brandt, Vanderbilt University, Department of Veterans Affairs and Tennessee Valley Healthcare System
Ann Richmond, Vanderbilt University, Department of Veterans Affairs and Tennessee Valley Healthcare System
Yu Shyr, Vanderbilt University Medical Center

In a tumor xenograft study, the aim is to compare differences in tumor volume between control and treated mice over time. For sample size and power estimation, tumor volumes at the final timepoint are usually analyzed, with and without transformation using the independent two sample t-test. However, with limited resources, it is worthwhile to know which statistical method(s) have the greatest power for fixed sample size. In our simulation study, several statistical methods are included to evaluate the probability of detecting a true difference between groups at fixed significance levels under true null and alternative hypothesis tests. These results will illustrate to statisticians and non-statisticians alike the relative performance of standard and non-standard analysis models in this common basic science research setting with implications for sample size and power estimation.

email: sheau-chiann.chen.1@vumc.org

ABSTRACTS & POSTER PRESENTATIONS

14g. Estimation and Outliers for Overdispersed Multinomial Data

Barry William McDonald*, Massey University

Suppose a number of geographical locations are surveyed, with each location giving rise to a multinomial response. We wish to estimate distributional parameters common across locations, allowing for overdispersion, compare the summary statistics for each location to their sampling distribution, and visualise and identify locations with unusual multinomial response patterns.

email: b.mcdonald@massey.ac.nz

14h. Partial Least Squares Regression-Based Framework for Incomplete Observations in Environmental Mixture Data Analysis

Ruofei Du*, University of New Mexico Comprehensive Cancer Center

Challenges are seen in the statistical analysis of environmental mixture data. In addition to the multicollinearity and high dimensionality of the covariates from the mixture exposure, a study may also be complicated by the relatively low association effect and a large proportion of incomplete observations. A statistical method needs to be capable of using all available information in capturing the weak association signal. We previously proposed methodology adjustments on the general PLSR approach so that the exposure information from the participants without outcomes can also contribute to the analysis. In the current study, we additionally code 'semi-dummy' variables to handle the missing exposure values due to technical limitations. For example, a heavy metal concentration tested lower than the limit of detection (< LOD) is conventionally treated as a missing value, but in the studied method the < LOD values become 'visible' to analysis and the related observation no longer be excluded. We further developed a procedure that can leverage the fittings of the rotated matrix of the mixture and the vectors of other variables.

email: rdu@salud.unm.edu

14i. Marginalized Zero-Inflated Negative Binomial Regression Model with Random Effects: Estimating Overall Treatment Effect on Lesion Counts among Multiple Sclerosis Patients (CombiRx Trial)

Steve B. Ampah*, University of Alabama at Birmingham
Lloyd J. Edwards, University of Alabama at Birmingham
Leann D. Long, University of Alabama at Birmingham
Byron C. Jaeger, University of Alabama at Birmingham
Nengjun Yi, University of Alabama at Birmingham

Correlated and overdispersed count outcome data that exhibit many zeros are most common in follow-up studies, where researchers often use the zero-inflated negative binomial (ZINB) regression model with random effects to assess the relationship between covariates of interest and dependent count responses over time. However, the regression parameters from this model do not provide direct interpretations for the overall population mean count because of its latent class specification. This makes regression coefficients from the ZINB models not appropriate for drawing inferences that relates the overall population mean. In this paper, we propose a marginalized zero-inflated negative binomial regression model with random effects that jointly models the marginal mean and the inflated zero process to provide estimates that allow direct quantification of covariates effects on the overall population mean while excess zeros and over-dispersion are accounted for. Finite sample characteristics of the new model are examined through simulations and used to analyze lesion counts in 1008 multiple sclerosis patients from the CombiRx study.

email: sbampah@uab.edu

---

15. POSTERS: CONSULTING, EDUCATION, POLICY, EPIDEMIOLOGY

15a. Semiparametric Shape Restricted Mixed Effect Regression Spline with Application on US Urban Birth Cohort Study Data and State-Wide Prenatal Screening Program Data

Qing Yin*, University of Pittsburgh

The linear model has been widely used in epidemiology to model the relationship between maternal exposure and fetal/infant outcome. When researchers suspect nonlinear relationship exists, some nonparametric techniques, including regression splines, smoothing splines and penalized regression splines, can be used to model the relationship. Applying these nonparametric techniques, researchers can relax the assumptions of linear model and capture more shapes underlying the data other than linearity. In this paper, we focus on the regression spline technique and develop a method to help researchers select the most suitable shape to describe their data among increasing, decreasing, convex and concave shapes. The method is developed using mixed effect regression spline by extending fixed effect regression spline. We illustrate the method using a U.S. urban birth cohort study data set and a state-wide prenatal screening program data set.

email: qiy25@pitt.edu

15b. Development and Validation of Models to Predict Foodborne Pathogen Presence and Fecal Indicator Bacteria Levels in Agricultural Water using GIS-Based, Data-Driven Approaches

Daniel L. Weller*, University of Rochester
Tanzy Love, University of Rochester
Alexandra Belias, Cornell University
Martin Wiedmann, Cornell University

The FDA recently proposed microbial standards for the use of surface water for produce production. Since water quality varies based on environmental conditions, approaches that account for this variation may improve growers' ability to identify and address food safety risks associated with preharvest water use. Due to the availability of spatial data, GIS-based analyses can be used to develop such approaches for individual water sources. This study used datasets collected in 2018 (68 streams, 196 samples) and 2017 (6 streams, 181 samples) to train and test, respectively, models that predict likelihood of foodborne pathogen presence and fecal indicator bacteria (FIB) levels in NY streams. At each sampling, separate samples were collected and tested for pathogens and used to enumerate FIB levels. Various machine learning approaches (e.g., random forest, support vector machines) were used to develop models for each outcome (e.g., Salmonella detection). To characterize the utility of each model type, each model was compared against a set of baseline models where E. coli levels were the sole covariate.

email: wellerd2@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

15c. Accounting for Competing Risks in Estimating Hospital Readmission Rates

John D. Kalbfleisch*, University of Michigan
Kevin Zhi He, University of Michigan
Douglas E. Schaubel, University of Pennsylvania
Wenbo Wu, University of Michigan

There has been substantial effort to reduce unnecessary hospital readmissions as one way to reduce medical costs in the Medicare and other patient populations. Standardized 30-day readmission rates or ratios for each provider have been frequently applied methods to monitor hospitals in this regard. These measures typically are based on a binary logistic regression model for each hospital discharge, with adjustment made for patient characteristics and comorbidities. In general, competing risks (e.g., planned readmissions, death, transplant, or admission to long-term care or rehabilitation facilities) can also occur. However, existing facility-profiling methods often account for competing risks in an ad hoc manner. Focusing on 30-day hospital readmission, we develop novel facility-profiling methods that appropriately account for competing risks. In the models used for covariate adjustment, time-to-readmission is represented by a discrete failure time variable. Based on patients admitted to acute care hospitals from dialysis facilities, we illustrate this approach and compare the results to methods currently in use.

email: jdkalbfl@umich.edu

15d. A New Framework for Cost-Effectiveness Analysis with Time-Varying Treatment and Confounding

Nicholas A. Illenberger*, University of Pennsylvania
Andrew J. Spieker, Vanderbilt University Medical Center
Nandita Mitra, University of Pennsylvania

To make informed health policy decisions, we must consider both a treatment's effectiveness and its cost. We previously developed the Net Benefit Separation (NBS) as a novel probabilistic measure of cost-effectiveness using the potential outcomes framework. This parameter describes the probability that a patient receiving one treatment will have a more cost-effective outcome than a patient receiving another treatment, at a particular willingness-to-pay value. Previous estimators of the NBS have focused on the case of point-exposure with time-invariant confounding. However, medical interventions often vary as a function of a patient's response to treatment or to changes in their underlying health status. Ignoring this time-varying structure can lead to biased estimates of cost-effectiveness. We propose a g-computation based procedure to estimate the NBS in the presence of time-varying exposure and confounding. Through simulation studies, we explore the finite sample properties of our estimator. Using data from a large observational cancer registry, we apply our method to determine the cost-effectiveness of adjuvant radiation therapy in endometrial cancer patients.

email: nillen@upenn.edu

15e. Rethinking the Introductory Biostatistics Curriculum for Non-Biostatisticians

Emily Slade*, University of Kentucky

Traditionally, introductory biostatistics courses rely on the foundation of probability and distributional theory. This foundation is necessary for students seeking degrees/jobs as biostatisticians. However, many students enrolling in intro biostatistics at the graduate level are training to be researchers in other fields. We redesigned the intro biostatistics curriculum for these students to streamline focus to data analysis, removing as much statistical theory as possible. While stripping away traditional statistical foundations from an intro course may seem radical, we found that it is possible to teach how and why specific statistical methods work with carefully-designed thought exercises. Significant time is also spent ensuring that the students understand the limits of their statistical knowledge. This enables the students to become smart users of basic statistics as well as effective team scientists when collaborating with biostatisticians. We will present the framework for our redesigned intro biostatistics course along with examples of transforming traditional lecture instruction and exercises to meet the needs of applied researchers in the modern team science landscape.

email: slademily@gmail.com

15f. Establishing Successful Collaboration in a Competitive Environment: Case Studies from a Healthcare Setting

Jay Mandrekar*, Mayo Clinic

As a statistical consultant at an academic health care center, we often get contacted to collaborate on intramural projects or extramural grant applications. In current environment when the institutional funding is shrinking and grant funding is also limited, one needs to have alternative mechanisms to support the infrastructure. In addition to this, with rapid advances in statistical methodology such as data science and artificial intelligence, statisticians are also expected to keep up with new methodology and software packages. The focus of this talk is to provide a brief overview of issues arising with limited funding and strategies to establish successful collaborations while sustaining in today's competitive environment. Real life examples and strategies used will be presented. Audience participations will be encouraged.

email: mandrekar.jay@mayo.edu

ABSTRACTS & POSTER PRESENTATIONS

15g. Likelihood Ratios to Compare the Statistical Performance of Multiple Tests in Simulation Studies

Qiuxi Huang*, Boston University School of Public Health

Simulation studies commonly aim at comparing the power of several tests. However, the empirical size can differ between tests, and comparing tests on power alone may lead to spurious conclusions. We propose positive and negative likelihood ratios (LR) to compare test performance while allowing for the discrepancy in test size. The positive LR is the ratio of the empirical power to the empirical size. It reflects the likelihood that a significant test result is found under the alternative as opposed to under the null. The negative LR is the ratio of 1 minus empirical power to 1 minus empirical size. It reflects the likelihood that a non-significant test result is found under the alternative as opposed to under the null. We discuss 3 publication bias tests in meta-analysis: rank correlation, linear regression, and a test based on observed and expected cell frequencies. Using positive and negative LR, we explored 144 scenarios. Under certain conditions, the linear regression has the largest empirical power but rank correlation is preferable after accounting for empirical size.

email: qiuxi@bu.edu

15h. Impact of a Biostatistics Department on an Academic Medical Center

Li Wang*, Vanderbilt University Medical Center
Henry Domenico, Vanderbilt University Medical Center
Daniel W. Byrne, Vanderbilt University Medical Center

A biostatistics department can have a positive impact on an academic medical center, but this effect is rarely described. At Vanderbilt University we created a Department of Biostatistics in 2003 and have seen the benefits of this investment but never formally measured the impact. This information may be useful for academic medical centers that do not have a true Department of Biostatistics or do not have leadership support to grow their department to the optimal size. The primary benefit of a Department is that it enables a medical center to hire and retain a large number of highly-skilled biostatisticians. We increased the number of biostatisticians from 6 (in 2003) to 73 (in 2019). This enabled researchers to become more successful at obtaining NIH funding, which increased from $221 million (in 2003) to $340 million (in 2017). The biostatisticians also make it possible to publish more scientific papers in high-profile journals. Our biostatisticians have also been instrumental at improving: hospital operations, health outcomes, and medical centers rankings. A final benefit is having the faculty for a graduate program and to train hundreds of physician-scientists.

email: li.wang@vumc.org

16. POSTERS: GENETICS, COMPUTATION

16a. Heterogeneity-Aware and Communication-Efficient Distributed Statistical Analysis

Rui Duan*, University of Pennsylvania
Yang Ning, Cornell University
Yong Chen, University of Pennsylvania

In multicenter research, individual level data are often protected against sharing across different sites. To overcome the barrier of data sharing, many distributed algorithms, which only require sharing the aggregated information, have been developed. The existing distributed algorithms usually assume the data are homogeneously distributed across sites. This assumption ignores the important fact that the samples collected in different sites may come from different sub-populations, which leads to substantial amount of heterogeneity across sites. In this paper, we propose a distributed algorithm which accounts for the heterogeneous distributions by allowing site specific nuisance parameters. The proposed method extends the surrogate likelihood approach to heterogeneous setting through a modified score function. We establish the non-asymptotic risk bound of the estimator and its limiting distribution in the two-index asymptotic setting. In addition, the asymptotic variance of the estimator attains the Cramer-Rao lower bound. Finally, the simulation study showed the proposed methods reach higher estimation accuracy compared to several existing methods.

email: ruiduan@upenn.edu

16b. False Discovery Rate Computation and Illustration

Megan C. Hollister*, Vanderbilt University Medical Center
Jeffrey D. Blume, Vanderbilt University Medical Center

False discovery rates (FDR) are an essential component of statistical inference. They reveal the propensity for an observed result to be mistaken. A FDR should accompany all observed results to help contextualize the relevance and potential impact of findings. In this talk, we introduce a user-friendly R function for computing FDRs for observed p-values. A variety of methods for FDR estimation and FDR control are available, and the user can select the approach most appropriate for their setting. These include Efron's Empirical Bayes FDR, Benjamini-Hochberg FDR control for multiple testing, Lindsey's method for smoothing empirical distributions, estimation of the mixing proportion, and central matching. We will illustrate the important difference between estimating the FDR for a particular finding and adjusting a hypothesis test to control the false discovery propensity. Finally, we will end with a quick comparison of the capabilities of the popular p.adjust function from the stats-package, which does not provide the desired FDR statistics. We aim to encourage wider reporting of false discovery rates for observed findings.

email: megan.c.hollister@vanderbilt.edu

ABSTRACTS & POSTER PRESENTATIONS

16c. A Modified Genomic Control Method for Genetic Association Analysis Using a Stratified, Cluster Sample

Donald Malec*, National Center for Health Statistics
John Pleis, National Center for Health Statistics
Rong Wei, National Center for Health Statistics
Bill Cai, National Center for Health Statistics
Yulei He, National Center for Health Statistics
Hee-Choon Shin, National Center for Health Statistics
Guangyu Zhang, National Center for Health Statistics

Genetic association analysis using a representative sample may have a potential benefit over using non-random observational data because confounding due to sample selection is eliminated. However, if correlation exists in the population it can still cause overdispersion in resulting tests of association. Here, the combined effects of overdispersion and the effects due to sampling on a trend test of association are examined together. These effects are identified in both the variability of the test of trend and in the variability of the standard genomic control method. Possible modifications based on conditional variances are suggested. These modifications are illustrated using a re-analysis of a candidate gene study of the genetic association with Hepatitis E resistance using data from the 1991-1994 National Health and Nutrition Examination Survey (NHANES III, phase 2) Genetic Data Repository.

email: dmalec@cdc.gov

16d. Semiparametric Functional Approach with Link Uncertainty

Young Ho Yun*, Virginia Tech

Many studies have reported associations between ambient temperature and daily deaths in cities. The regional impacts of climate change will vary widely depending on the vulnerability of the population. The nonlinear relations (U, J, or V shaped) have been observed with increased mortality at both high and low temperatures. However, this relationship was investigated under the strong model and distribution assumption which are often not satisfied in real applications that involve a relatively small number of subjects, have heterogeneous measurement errors, or have large variation among subjects. This traditional approach may easily turn out to be mis-specified, and overdispersion problems may arise, influencing the statistical inferences. Therefore, in this paper, we develop a nonparametric setting which can include the traditional model with the canonical link and also cover the noncanonical link and nonlinear model in general. We demonstrate our approach to study the ambient temperature–mortality association for four capital cities in East Asia.

email: yyun1@vt.edu

16e. Multi-Ethnic Phenotype Prediction via Effective Modeling of Genetic Effect Heterogeneity

Lina Yang*, The Pennsylvania State University College of Medicine
Dajiang Liu, The Pennsylvania State University College of Medicine

Genetic prediction is critical for understanding the risk for developing diseases and achieving precision medicine. Properly analyzing the multi-ethnic samples while accommodating the heterogeneities in the LD patterns and genetic effects can improve prediction accuracy. Existing methods for multi-ethnic analysis often group studies into discrete ancestry groups, which ignores the fact that ancestry may vary continuously. Another multi-ethnic analysis method uses meta-regression to model genetic effects as a function of the ancestry groups, but it does not model the non-ancestral heterogeneity of effects. We developed a novel mixed effect meta-regression approach which decomposes genetic effect heterogeneity into a component that is determined by ancestry (fixed effect) and a non-ancestral component (random effect). Our simulation study shows our methods have improved accuracy of genetic prediction over genetic predictions based upon 1) fixed effect meta-analysis of all ancestries, 2) ancestry specific analysis of the training data, and 3) random effect meta-analysis of all ancestries. The dataset is benchmarked using UK Biobank dataset and smoking/drinking phenotypes.

email: lzy51@psu.edu

16f. High Dimensional Sparse Regression with Auxiliary Data on the Features

Constanza Rojo*, University of Wisconsin, Madison
Pixu Shi, University of Wisconsin, Madison
Ming Yuan, Columbia University
Sunduz Keles, University of Wisconsin, Madison

Regulatory genomic information has been recognized as a potential source of information that can improve the detection and biological interpretation of single-nucleotide polymorphism (SNP) in genome-wide association studies (GWAS). Current methodologies that aim to integrate annotation information focus mainly on fine-mapping and overlook the importance of its use on earlier GWAS stages. Moreover, there is a lack of development on proper statistical frameworks that provide annotation and SNP selection simultaneously. In this paper, we develop annoReg, a statistical method that enables the integration of high-dimensional and sparse auxiliary information into high-dimensional regression. This method seeks to improve the simultaneous detection of relevant genetic variants and annotation data. Simulation experiments indicate that annoReg can improve the identification of genetic variants by increasing the average area under the precision-recall curve by up to 60%. Real data applications show that annoReg can select relevant genetic variants while detecting several transcription factors involved on specific phenotypical changes.

email: mrojo@wisc.edu

ABSTRACTS & POSTER PRESENTATIONS

16g. A Unified Linear Mixed Model for Simultaneous Assessment of Familiar Relatedness and Population Structure

Tao Wang*, Medical College of Wisconsin
Paul Livermore Auer, University of Wisconsin, Milwaukee
Regina Manansala, University of Wisconsin, Milwaukee
Andrea Rau, GABI, INRA, AgroParisTech and Université Paris-Saclay, France
Nick Devogel, Medical College of Wisconsin

There are two broad sources for genetic correlation: familial relatedness and population structure. Each source can further break down into additive and dominance components to account for potential additive and dominance genetic effects, respectively. In this study, a simultaneous assessment of familial relatedness and population structure including both the additive and dominance components is considered under a unified linear mixed model. First, the additive and dominance genomic relationship matrices are introduced and linked to the kinship (or cocancestry) and double coancestry coefficients among individuals. Then, a unified linear mixed model is proposed with separate correlation matrices for the familial relatedness and population structure including both the additive and dominance components. How to fit this unified linear mixed model and test for the variance components of additive and dominance effects from familial relatedness and population structure are also explored. Finally, this unified linear mixed model is applied to analyze a real data set from UKBiobank.

email: taowang@mcw.edu

16h. Cubic Kernel Method for Implicit T Central Subspace

Weihang Ren*, University of Kentucky
Xiangrong Yin, University of Kentucky

The T-central subspace, introduced by Luo, Li and Yin (2014), allows one to perform sufficient dimension reduction for any statistical functional of interest. We propose a general estimator using (third) moment kernel to estimate the implicit T-central subspace. In particular, we focus on central quantile/expectile subspace via the implicit functionals of quantile/expectile. Theoretical results are established and simulation studies demonstrate the efficacy of our proposed methods.

email: weihang.ren@uky.edu

16i. ODAH: A One-Shot Distributed Algorithm for Estimating Semi-Continuous Outcomes using EHR Data in Multiple Sites

Mackenzie J. Edmondson*, University of Pennsylvania
Chongliang Luo, University of Pennsylvania
Rui Duan, University of Pennsylvania
Mitchell Maltenfort, Children's Hospital of Philadelphia
Christopher Forrest, Children's Hospital of Philadelphia
Yong Chen, University of Pennsylvania

EHR data are widely used in modern healthcare research, containing useful information describing patients' clinical visits. Due to privacy concerns surrounding patient-level data sharing, most clinical data analyses are performed at individual sites. This leads to underpowered studies specific to a certain population, creating a need for methods which perform analyses across sites without sharing patient-level data. To address this, distributed algorithms have been developed to conduct analyses across sites by sharing only aggregated information, preserving patient privacy. We propose a communication-efficient distributed algorithm for performing hurdle regression on data stored in multiple sites. By modeling zero and positive counts separately, we account for zero-inflation in the outcome, which is common in characterizing patient hospitalization frequency. Our simulations show that our algorithm achieves high accuracy comparable to the oracle estimator using all patient-level data pooled together. We apply our algorithm to data from the Children's Hospital of Philadelphia to estimate how often a patient is likely to be hospitalized given data collected during clinical visits.

email: macjohn@pennmedicine.upenn.edu

---

17. POSTERS: META-ANALYSIS, MISSING DATA AND MORE

17a. Multiple Imputation of Missing Covariates in Meta-Regression using Multivariate Imputation by Chained Equations

Amit K. Chowdhry*, University of Rochester Medical Center
Michael P. McDermott, University of Rochester Medical Center

When conducting study-level meta-regression analyses, it is not uncommon to have missing covariates for some of the studies. Often, covariates are aggregated patient characteristics that may not be reported in the publication of all studies included in the meta-analysis. Methods for handling missing covariates include multiple imputation assuming monotone missingness, joint distribution modeling, and fully conditional specification, one implementation of which is multivariate imputation by chained equations (MICE). We propose a method adapting MICE to missing covariates in meta-analysis by using meta-regression to impute missing covariates that are aggregate quantities. This method weights studies by the inverse of the variance, which is a function of sample size (W-MICE). We performed simulation studies to evaluate the proposed approach in terms of standard operating characteristics. W-MICE was compared with the following methods: standard MICE using linear regression that weights all aggregate statistics equally, mean imputation (ME), and complete case analysis (CC). W-MICE and standard MICE were found to have similar properties and to be generally superior to ME and CC.

email: amit.chowdhry@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

17b. Test-Inversion Confidence Intervals for Estimands in Contingency Tables Subject to Equality Constraints

Qiansheng Zhu*, University of Iowa
Joseph B. Lang, University of Iowa

We focus on constructing test-inversion confidence intervals for estimands in contingency tables subject to equality constraints. Recommended test statistics include the difference in G2 statistic and nested versions of power-divergence statistics. Efficient and robust algorithms are proposed, and they broaden the applicability in both estimands and constraints: (i) Compared with existing methods where only likelihood-explicit estimands are applicable, our algorithms can also deal with likelihood-implicit estimands, where the log-likelihood cannot be reparameterized in terms of the estimand of interest and nuisance parameters; (ii) Only mild conditions on constraints are required, and it is unnecessary to re-express constraints as a generalized linear model. Simulation studies highlight the advantages of using likelihood-ratio intervals rather than bootstrap and Wald intervals, especially when table counts are small and/or the true estimand is close to the boundary. Finally, examples are presented to illustrate the appropriateness of imposing constraints and the utility of test-inversion intervals.

email: qiansheng-zhu@uiowa.edu

17c. Bayesian Cumulative Probability Models for Continuous and Mixed Outcomes

Nathan T. James*, Vanderbilt University
Bryan E. Shepherd, Vanderbilt University
Leena Choi, Vanderbilt University
Yuqi Tian, Vanderbilt University
Frank E. Harrell, Jr., Vanderbilt University

Ordinal cumulative probability models (CPMs) such as the proportional odds regression model are typically used for discrete ordered outcomes, but can readily accommodate a wide range of continuous and mixed discrete/continuous outcomes. Several recent papers have described the advantages of ordinal CPMs in this setting using non-parametric maximum likelihood estimation (NPMLE), however the extension to Bayesian inference has not been thoroughly explored. We formulate a Bayesian CPM for continuous or mixed outcome data. Bayesian CPMs have advantages over frequentist CPMs with regard to interpretation and flexibility/extendibility. These models also allow exact inference (within MCMC sampling error) when the log-likelihood is not well-approximated by a quadratic function. We detail progress on several extensions including model averaging across various link function specifications and use for non-linear mixed effects modeling. We evaluate our model against a Bayesian nonparametric regression approach using simulations and a case study.

email: nj1154@gmail.com

17d. R-Squared and Goodness of Fit in the Linear Mixed Model: A Cautionary Tale

Boyi Guo*, University of Alabama at Birmingham
Byron C. Jaeger, University of Alabama at Birmingham

The R2 statistic is a well -known, frequently applied metric to measure the overall goodness of fit for a linear model. Due to the lack of a unifying definition of an R2 statistic for the linear mixed model, a number of competing methodologies have been introduced. While each of the proposed methods has merit, analysts are forced to face with a difficult decision: which R2 statistic should be applied to measure the overall goodness of fit for linear mixed models? In this article, we answer this question by quantifying how accurately each competing R2 statistic can measure goodness of fit in a broad range of simulated scenarios. Our investigation applies a simple premise: if an R2 statistic accurately measures goodness of fit, it will be maximized by the models that most closely mimic the underlying data generating process. Hence, we evaluate each computing R2 statistic based on the performance of the model that it selects.

email: boyiguo1@uab.edu

17e. On the Optimality of Group Testing Estimation

Sarah Church*, Radford University
Md S. Warasi, Radford University

This paper considers the problem of estimating the prevalence of a disease (e.g., HIV) by using hierarchical group testing. Under this framework, testing for pools, composed of individual biospecimens, is performed at the initial stage, and retesting is performed subsequently for case identification. Statistical models for group testing data are complicated by two factors: (a) presence of testing error and (b) dependence among test responses. Consequently, estimation and performing optimality for the disease prevalence are challenging. This article presents a new approach to studying optimality properties of a group testing estimator, including bias and efficiency. We quantify bias by comparing a true model with a model that is misspecified due to testing error, and measure the loss (or gain) in efficiency by comparing two true models when a testing error is either present or absent. Doing this enables us to develop optimal pooling designs, which can be implemented in public health surveillance applications for an accurate and efficient estimate of the disease prevalence. We illustrate our results by using chlamydia and gonorrhea data from Nebraska Public Health Laboratory.

email: schurch10@radford.edu

ABSTRACTS & POSTER PRESENTATIONS

17f. Bayesian Wavelet-Packet Historical Functional Linear Models

Mark J. Meyer*, Georgetown University
Elizabeth J. Malloy, American University
Brent A. Coull, Harvard T. H. Chan School of Public Health

Historical Functional Linear Models (HFLM) quantify associations between a functional predictor and functional outcome where the predictor is an exposure variable that occurs before, or at least concurrently with, the outcome. Current research on the HFLM is largely limited to frequentist estimation techniques that employ spline-based basis representations. In this work, we propose a novel use of the discrete wavelet-packet transformation, which has not previously been used in functional models, to estimate historical relationships in a fully Bayesian model. Since inference has not been an emphasis of the existing work on HFLMs, we also adapt two Bayesian inference procedures to the historical functional setting. We investigate the operating characteristics of our wavelet-packet HFLM, as well as the two inference procedures, in simulation and use the model to analyze data on the impact of lagged exposure to particulate matter finer than 2.5μg on heart rate variability in a cohort of journeyman boilermakers over the course of a day's shift.

email: mark.john.meyer@gmail.com

17g. EMBVs: An EM-Bayesian Approach for Analyzing High-Dimensional Clustered Mixed Outcomes

Yunusa Olufadi*, University of Memphis
E. Olusegun George, University of Memphis

The recent surge in the increase for the integration of multiple data sets leads to a requirement for new methods applicable to mixed outcomes of different kinds where the straightforward application of existing methods is not necessarily the best approach. This presentation is motivated by the statistical and computational challenges that arise from analyzing clustered multiple mixed outcomes with high-dimensional covariates. Such data are becoming increasingly common in toxico(epi)genomics, developmental neurotoxicity, and developmental toxicity (DT) studies. Aside from risk assessments, investigators are now interested in identifying biomarkers of DT or detect new biomarkers of DT. We propose an EM Bayesian variable selection procedure to guide both the estimation and efficient extraction of potential biomarkers. Thresholding and regularization plots informed both model evaluation and variable selection. Synthetic and real data are used to demonstrate the performance and application of this novel procedure.

email: yolufadi@gmail.com

17h. Generalized Additive Dynamic Effect Change Models: An Interpretable Extension of GAM

Yuan Yang*, University of Michigan
Jian Kang, University of Michigan
Yi Li, University of Michigan

Generalized additive models (GAMs) have been widely used for modeling nonlinear effects of predictors on a variety of outcomes. However, the explanation of covariates' effects in GAMs is intriguing and statistical inference on effects is challenging. Extending GAMs, we propose a new class of models that can directly characterize the dynamic effect change of each predictor. Our model, which incorporates derivatives of nonlinear effects as functional parameters of interest, is termed the generalized additive dynamic effect change model. We develop an efficient statistical procedure for inferring functional parameters embedded in the reproducing Hilbert kernel space. As opposed to GAMs, our derivative-based model renders a straightforward interpretation of model parameters. We establish large sample properties for the proposed method and show its superior performance compared to GAM in various simulation scenarios. We apply our method to construct an individualized risk prediction model for opioid use, which provides a better understanding of dynamic effect changes of potential risk factors.

email: yuanyang@umich.edu

17i. A Functional Generalized Linear Mixed Model for Estimating Dose Response in Longitudinal Studies

Madeleine E. St. Ville*, Clemson University
Andrew W. Bergen, Oregon Research Institute
Carolyn M. Ervin, BioRealm
Christopher McMahan, Clemson University
James W. Baurley, BioRealm
Joe Bible, Clemson University

We propose a functional generalized linear mixed model for estimating dose response in longitudinal studies. Our methodology is intended to accommodate data where subjects are subjected to adaptive dosing regimens (exposure by design) as well as addressing variation in dosage due to study compliance (random exposure). Our proposed methodology is motivated by and applied to a meta-analysis of six clinical trials designed to measure the effects of buprenorphine on opiate use cessation.

email: mstvill@g.clemson.edu

ABSTRACTS & POSTER PRESENTATIONS

18. MODERN FUNCTIONAL DATA ANALYSIS

Minimax Powerful Functional Analysis of Covariance Tests for Longitudinal Genome-Wide Association Studies

Yehua Li*, University of California, Riverside

We model the Alzheimer's Disease (AD) related phenotype response variables observed on irregular time points in longitudinal Genome-Wide Association Studies (GWAS) as sparse functional data and propose nonparametric test procedures to detect functional genotype effects, while controlling the confounding effects of environmental covariates. Existing nonparametric tests do not take into account within-subject correlations, suffer from low statistical power, and fail to reach the genome-wide significance level. We propose a new class of functional analysis of covariance (fANCOVA) tests based on a seemingly unrelated (SU) kernel smoother, which can incorporate the correlations. We show that the proposed SU-fANCOVA test combined with a uniformly consistent nonparametric covariance function estimator enjoys the Wilks phenomenon and is minimax most powerful. In an application to the Alzheimer's Disease Neuroimaging Initiative data, the proposed test leads to discovery of new genes that may be related to AD.

email: yehuali@ucr.edu

Bayesian Function-on-Scalars Regression for High-Dimensional Data

Daniel R. Kowal*, Rice University
Daniel C. Bourgeois, Rice University

We develop a fully Bayesian framework for function-on-scalars regression with many predictors. The functional data response is modeled nonparametrically using unknown basis functions, which produces a flexible and data-adaptive functional basis. For variable selection in functional regression, we propose a decision theoretic posterior summarization technique, which identifies a subset of covariates that retains nearly the predictive accuracy of the full model. Our approach is broadly applicable for Bayesian functional regression models, and unlike existing methods provides joint rather than marginal selection of important predictor variables. Computationally scalable posterior inference is achieved using a Gibbs sampler with linear time complexity in the number of predictors. The resulting algorithm is empirically faster than existing frequentist and Bayesian techniques, and provides estimation of model parameters, prediction and imputation of functional trajectories, and uncertainty quantification. The methodology is applied to actigraphy data to investigate the association between intraday physical activity and responses to a sleep questionnaire.

email: daniel.kowal@rice.edu

Modern Functional Data Analysis for Biosciences

Ana-Maria Staicu*, North Carolina State University
Alex Long, North Carolina State University
Meredith King, Northrop Grumman

In many applications, functional data are recorded at various spatial locations, and it is reasonable to assume that the curves are spatially correlated. When the domain for the spatial indices is fixed, it is well known that the common functional principal component analysis methods produce inconsistent estimators. In this paper, we propose a functional principal component estimation methodology that mitigates challenges associated with estimation in a fixed or bounded spatial domain. We investigate the properties of the proposed estimators theoretically as well as numerically, in finite samples. We further illustrate our approach by analyzing spatially indexed fine particulate nitrate profiles monitored across the United States.

email: astaicu@ncsu.edu

Mean and Covariance Estimation for Functional Snippets

Jane-Ling Wang*, University of California, Davis
Zhenhua Lin, National University of Singapore

Estimation of mean and covariance functions is fundamental for functional data analysis. While this topic has been studied extensively in the literature, a key assumption is that there are enough data in the domain of interest to estimate both the mean and covariance functions. However, in many longitudinal studies subjects enter the study at different times and are followed for a brief period of time which is much shorter than the length of the study. This results in functional snippets, which are short segments of functions possibly observed irregularly on an individual-specific subinterval that is much shorter than the entire study interval. Estimation of the covariance function for functional snippets is challenging since information for the far off-diagonal regions of the covariance structure is completely missing. In this talk, we discuss two approaches to tackle this challenge and provide theoretical and numerical support.

email: janelwang@ucdavis.edu

ABSTRACTS & POSTER PRESENTATIONS

19. DISTRIBUTED AND PRIVACY-PRESERVING METHODS FOR ELECTRONIC HEALTH RECORDS DATA

Communication Efficient Federated Learning from Multiple EHRs Databases

Changgee Chang*, University of Pennsylvania
Zhiqi Bu, University of Pennsylvania
Qi Long, University of Pennsylvania

Electronic health records (EHRs) offer great promises of advancing precision medicine and, at the same time, present significant analytical challenges. Particularly, it is often the case that patient-level data in EHRs cannot be shared across institutions (data sources) due to government regulations and/or institutional policies. As a result, there is a growing interest in federated learning from multiple EHRs databases that does not require sharing patient-level data. To tackle such challenges, we propose a new communication efficient algorithm that aggregates the local parameter estimates. Our approach requires only a single one-way communication from all sites to the central site, where the analysis is taking place, and is capable of utilizing the local data in the central site to improve estimation and inference. We investigate the operating characteristics of the proposed method and evaluate its performance in both simulations and real data analysis in comparison with several recently developed methods.

email: changgee@pennmedicine.upenn.edu

Adaptive Noise Augmentation for Privacy-Preserving Empirical Risk Minimization

Fang Liu*, University of Notre Dame
Yinan Li, University of Notre Dame

We propose adaptive Noise Augmentation for Privacy-Preserving (NAPP) empirical risk minimization (ERM) problems. With appropriately designed augmented data, NAPP-ERM iteratively utilizes non-private ERM solvers to promote regularization and sparsity in model estimation and simultaneously achieve differentially privacy. NAPP leads to strong convexity of the noise-perturbed ERM and eliminates the need for an additional l2 penalty term just for the purposes of achieving strong convexity. As a result, there are decreases in both the excess risk bound and the sample complexity. With the efficiency of NAPP in introducing strong convexity to ERM, achieving regularization and differential privacy simultaneously, we can retrieve part of the privacy budget originally reserved for achieving DP alone, which can be recycled to improve the accuracy of the ERM solutions or be saved to enjoy a higher level of differential privacy guarantee on the ERM learner. We illustrate through simulated and real-life data the improvement of NAPP in accuracy in differentially privately learning generalized linear models against existing privacy-preserving ERM learners.

email: fang.liu.131@nd.edu

Generating Poisson-Distributed Differentially Private Synthetic Data

Harrison Quick*, Drexel University

The dissemination of synthetic data can be an effective means of making information from sensitive data publicly available while reducing the risk of disclosure associated with releasing the sensitive data directly. While mechanisms exist for synthesizing data that satisfy formal privacy guarantees, the utility of the synthetic data is often an afterthought. More recently, the use of methods from the disease mapping literature has been proposed to generate spatially-referenced synthetic data with high utility, albeit without formal privacy guarantees. The objective for this paper is to help bridge the gap between the disease mapping and the formal privacy literatures. In particular, we generalize an existing approach for generating formally private synthetic data to the case of Poisson-distributed count data in a way that accommodates heterogeneity in population sizes and allows for the infusion of prior information. We demonstrate via simulation study that the proposed approach for generating differentially private synthetic data outperforms a popular technique when the counts correspond to events arising from subgroups with unequal population sizes or unequal event rates.

email: hsq23@drexel.edu

---

dblink: Distributed End-to-End Bayesian Entity Resolution

Rebecca Steorts*, Duke University
Neil Marchant, University of Melbourne
Ben Rubinstein, University of Melbourne
Andee Kaplan, Colorado State University
Daniel Elazar, Australian Bureau of Statistics

Entity resolution (ER) (record linkage or de-duplication) is the process of merging together noisy databases, often in the absence of a unique identifier. A major advancement in ER methodology has been the application of Bayesian generative models. Such models provide a natural framework for clustering records to unobserved (latent) entities, while providing exact uncertainty quantification and tight performance bounds. Despite these advancements, existing models do not both scale to realistically-sized databases and incorporate probabilistic blocking in an end-to-end approach. In this paper, we propose “distributed blink” or dblink--the first scalable and distributed end-to-end Bayesian model for ER, which propagates uncertainty in blocking, matching, and merging. We conduct experiments on real and synthetic data which show that dblink can achieve significant efficiency gains--in excess of 200 times--when compared to existing methodology.

email: beka@stat.duke.edu

ABSTRACTS & POSTER PRESENTATIONS

20. INNOVATIVE STATISTICAL METHODS IN ENVIRONMENTAL MIXTURE ANALYSIS

Group Inverse-Gamma Gamma Shrinkage for Estimation and Selection in Multipollutant Models

Jonathan Boss*, University of Michigan
Jyotishka Datta, University of Arkansas
Sehee Kim, University of Michigan
Bhramar Mukherjee, University of Michigan

One of the principal methodological challenges in multipollutant modeling is regression coefficient estimation and variable selection in the presence of high intra-exposure class correlations. Adapting intuition from other global-local shrinkage prior approaches, we propose the group inverse-gamma gamma (GIGG) prior where the groups are pre-determined by known exposure classes. The GIGG prior has two group-specific hyperparameters which control the strength of overall shrinkage and the degree of within-group co-selection. Therefore, the GIGG prior flexibly adjusts shrinkage for each exposure class based on the within-group correlation structure and the within-group signal distribution. For fixed hyperparameter values, the full conditional distributions can be derived in closed form, resulting in a computationally efficient posterior simulation algorithm. We illustrate the performance of the GIGG prior through extensive simulation studies and demonstrate the proposed method by conducting an exposome-wide association study for cardiometabolic outcomes with data from the National Health and Nutrition Examination Survey.

email: bossjona@umich.edu

Bayesian Copula Regression for Inference on Dose-Response Curves

Federico H. Ferrari*, Duke University
Stephanie M. Engel, University of North Carolina, Chapel Hill
David B. Dunson, Duke University
Amy H. Herring, Duke University

Chemicals often co-occur in the environment or in synthetic mixtures, and as a result, exposure levels can be highly correlated. Hence, assessing joint effects is of critical public health concern. We propose a latent copula factor regression that provides interpretability via grouping of variables, flexibility for dose response curves, scalability for new data challenges, and uncertainty quantification. Interpretable grouping of variables is attained using a copula factor model for the exposures. This approach treats the marginal densities of the exposures as nuisance parameters and models the covariance among them, effectively accommodating the limit of detection and missing data. The latent factor specification provides dimensionality reduction in modeling of the covariance structure in the predictors and characterizing the impact of correlated groups of predictors on the response. Then, we regress the outcome on the latent variables with a flexible model specification allowing estimation of nonlinear dose-response curves. We propose a Bayesian approach to inference and provide a fast MCMC algorithm. We evaluate the performance using a simulation study and data examples.

email: amy.herring@duke.edu

Do Males Matter? A Couple-Based Statistical Model for Association Between Environmental Exposures to Pollutants and Infertility

Zhen Chen*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Motivated by the Longitudinal Investigation of Fertility and the Environment (LIFE) Study that investigated the association between exposures to environmental pollutants and human reproductive outcomes, we propose a joint latent risk class modeling framework with an interaction between female and male partners of a couple. This formulation introduces a dependence structure between chemical patterns within a couple and between the chemical patterns and the risk of infertility. The interaction enables interplay between partners' chemical patterns on the risk of infertility in a parsimonious way. We developed Markov chain Monte Carlo algorithms to obtain posterior estimates of model parameters and conducted simulations to examine the performance of the estimation approach. We found that in addition to the effect of PCB exposures of females, the male partners' PCB exposures play an important role in determining risk of infertility. Further, this risk is sub-additive in the sense that there is likely a ceiling effect which limits the probability of infertility when both partners of the couple are at high risk.

e-mail: chenzhe@mail.nih.gov

Accommodating Assay Limit-of-Detection in Environmental Mixture Analysis

Jason P. Fine*, University of North Carolina, Chapel Hill
Ling-Wan Chen, National Institute of Environmental Health Sciences, National Institutes of Health
Shanshan Zhao, National Institute of Environmental Health Sciences, National Institutes of Health

Humans are exposed to a multitude of environmental toxicants daily, and there is a great interest in developing statistical methods for assessing the effects of environmental mixtures on various health outcomes. One difficulty is that multiple chemicals in the mixture can be subject to left-censoring due to varying limits of detection (LOD). Conventional approaches either ignore these measures, dichotomize them at the limits, or replace them with arbitrary values such as LOD/2. Methods have been proposed to handle a single biomarker with limit of detection in such setting, by joint modeling the left-censored biomarker measure with an AFT model and the disease outcome with a generalized linear model. We extend this method to handle multiple correlated biomarkers subject to LOD, through a newly proposed nonparametric estimator of the multivariate survival function and innovative computational approaches. We apply the proposed method to the LIFECODES birth cohort to elucidate the relationship between maternal urinary trace metals and oxidative stress markers.

e-mail: shanshan.zhao@nih.gov

ABSTRACTS & POSTER PRESENTATIONS

21. MENTORING THROUGHOUT A LIFETIME: CONSIDERATIONS FOR MENTORS AND MENTEES AT ALL CAREER STAGES

Panel Discussion:

Leslie McClure, Drexel University
Brian Millen, Eli Lilly and Company
Dionne Price, U.S. Food and Drug Administration
Manisha Desai, Stanford University

Mentoring is recognized to be essential to the growth of junior statisticians. However, mentoring is also important for mid-career and senior statisticians, for continued professional growth and service as advisors. Often, junior statisticians struggle with how to put good mentoring relationships into place, and how to make the most of their mentoring sessions. Midcareer professionals struggle with the dual role of being mentored and being a mentor, especially without mentoring training. Senior researchers may struggle to find mentors as they move into senior and leadership roles. The Regional Advisory Board (RAB) will host a session with advice for mentors and mentees at different career stages. Issues regarding seeking out peer and senior mentors, becoming a mentor, and making the most of a mentoring relationship will be addressed. This session consists of a panel of experts and leaders from academia, government, industry, and consulting who are known leaders. The session will include prepared questions to elicit a concise variety of perspectives and opportunities for questions and answers from the audience.

email: naomi.brownstein@moffitt.org

22. INNOVATIVE STATISTICAL APPROACHES FOR HIGH-DIMENSIONAL OMIC AND MICROBIOMIC DATA

Advances and Challenges in Single Cell RNA-Seq Analysis

Susmita Datta*, University of Florida
Michael Sekula, University of Louisville
Jeremy Gaskins, University of Louisville

While the bulk RNA-sequencing (RNA-Seq) data with average (over millions of cells) transcriptomic abundance measurements have been valuable in countless studies, they often conceal cell-specific heterogeneity in expression signals that may be paramount to new biological findings. Fortunately, with single cell RNA-sequencing (scRNA-Seq) data from individual cells are now accessible, providing opportunities to investigate functional states of cells and identify rare cell populations. However, there are challenges in analyzing such data with multimodality, sparsity and heterogeneity. We will describe ways of modeling such data, finding differentially expressed genes and possible ways of constructing gene-gene interaction network. We will compare the performance of our modeling and differential analysis with respect to some other existing methods.

email: susmita.datta@ufl.edu

Predicting DNA Methylation from Genetic Data Lacking Racial Diversity Using Shared Classified Random Effects

J. Sunil Rao*, University of Miami
Hang Zhang, University of Miami
Melinda Aldrich, Vanderbilt University Medical Center

Public genomic repositories are notoriously lacking in racially and ethnically diverse samples which limits the reach of exploration. Our particular focus here is to provide a model-based framework for accurately predicting DNA methylation from genetic data using racially sparse public repository data. We develop a new prediction approach which uses shared random effects from a nested error mixed effects regression model. The sharing of random effects allows borrowing of strength across racial groups greatly improving predictive accuracy. Additionally, we show how to further borrow strength by combining data from different cancers in TCGA even though the focus of our predictions is DNA methylation in cervical cancer. We compare our methodology against other popular approaches including the elastic net shrinkage estimator and random forest prediction. Results are very encouraging with the shared classified random effects approach uniformly producing more accurate predictions - overall and for each racial group.

email: jrao@miami.edu

Sparse Generalized Dirichlet Distributions for Microbiome Compositional Data

Jyotishka Datta*, University of Arkansas
David B. Dunson, Duke University

Global-local shrinkage priors have been established as the current state-of-the art Bayesian tool for sparse signal detection leading to a huge literature proposing elaborate shrinkage priors for real-valued parameters. However, there has been limited consideration of discrete data structures including sparse compositional data, routinely occurring in microbiomics. I will discuss two methodological challenges. First, the Dirichlet is highly inflexible as a shrinkage prior for high-dimensional probabilities for its inability to adapt to an arbitrary level of sparsity. We address this gap by proposing the Sparse Generalized Dirichlet distribution, specially designed to enable scaling to data with many categories. A related problem is associating the compositional response data with environmental or clinical predictors. I will develop Bayesian variable selection strategies using global-local shrinkage priors for detecting significant associations between available covariates and taxonomic abundance tables. I will provide some theoretical support for the proposed methods and show improved performance in several simulation settings and application to microbiome data.

email: jyotishka.datta@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Nonparametric Differential Analysis for Dependent Multigroup Data with Application to DNA Methylation Analyses

Subharup Guha*, University of Florida
Chiyu Gu, Monsanto Company
Veerabhadran Baladandayuthapani, University of Michigan

Cancer omics datasets involve widely varying sizes and scales, measurement variables, and correlation structures. An overarching goal is the development of general statistical techniques that can cleanly sift the signal from the noise in identifying genomic signatures of the disease across a set of experimental or biological conditions. We propose BayesDiff, a nonparametric Bayesian approach based on a novel class of first order mixture models, called the Sticky Poisson-Dirichlet process or multicuisine restaurant franchise. The BayesDiff methodology utilizes information from all the measurements and adaptively accommodates any serial dependence in the data, accounting for the inter-probe distances, to perform simultaneous inferences on the variables. In simulation studies, we demonstrate the effectiveness of the BayesDiff procedure relative to existing techniques for differential DNA methylation. In a DNA methylation gastrointestinal cancer (GI) dataset, we detect the genomic signature for four types of cancer. The results support and complement known features of DNA methylation as well as gene associations with GI cancer.

email: s.guha@ufl.edu

23. BAYESIAN NONPARAMETRICS FOR CAUSAL INFERENCE AND MISSING DATA

Bayesian Nonparametric Models to Address Positivity Assumption Violations in Causal Inference

Jason Roy*, Rutgers University

Postivity assumption violations occur when the conditional probability of treatment is 0 for some values over the set of confounders. Due to extrapolation, caution is needed when trying to inference population causal effects in those settings. We develop a Bayesian nonparametric approach that is a model-based approach for ensuring that regions of the covariate space where positivity violations are questionable contribute noisier counterfactual predictions when estimating causal effects. The methods are assessed via a series of simulation studies.

email: jason.roy@rutgers.edu

Sensitivity Analysis using Bayesian Additive Regression Trees

Nicole Bohme Carnegie*, Montana State University
Vincent Dorie, Columbia University
Masataka Harada, Fukuoka University
Jennifer Hill, New York University

Bayesian Additive Regression Trees (BART) are an increasingly popular non-parametric modeling tool used in causal inference applications to reduce model misspecification. In this talk, we introduce a two-parameter, semi-parametric sensitivity analysis approach using BART to simultaneously address model misspecification and sensitivity to unmeasured confounding, based on our earlier work using simulated potential confounders. Using simulation studies, we show that our approach is competitive with existing approaches when the underlying response surface is linear and additive, and outperforms existing approaches when there is nonlinear confounding. The method is demonstrated on data from the third National Health and Nutrition Examination Survey, examining the effect of anti-hypertensive medication as used in general clinical practice.

email: nicole.carnegie@montana.edu

Variable Selection in Bayesian Nonparametric Models for High-Dimensional Confounding

Michael J. Daniels*, University of Florida
Kumaresh Dhara, University of Florida
Jason Roy, Rutgers University

Enriched Dirichlet processes (EDPs) provide very flexible conditional distributions that are powerful for modeling the distribution of outcomes given confounders for causal inference. However in most applications, many of the potential confounders are not true confounders. Further complicating any sort of variable selection is that such flexible models, which detect non-additivity and non-linearity in default ways, do not correspond to a single parameter (e.g., being zero) for each potential confounder. As such, we propose a backward selection algorithm that identifies covariates potentially unrelated to the outcome and then use a Bayesian model selection statistic to decide whether to remove the variable. The procedure is explored via simulations and applied to a real dataset.

email: mdaniels@stat.ufl.edu

Accelerated Bayesian G-Computation Algorithms

Antonio R. Linero*, University of Texas, Austin

In causal inference problems, estimands are frequently expressed not as closed form functions of parameters. For moderately complicated problems, these quantities must be evaluated numerically, which is often done by Monte Carlo integration. For simulation-based computation methods, such as the bootstrap or Markov chain Monte Carlo, this dramatically increases the computational burden, as a naive implementation of Monte Carlo integration will need to be performed at every iteration of the original sampling algorithm. We show how to dramatically accelerate computations, making the total computational burden of Monte Carlo integration the same (or of smaller) order than the original sampling algorithm. We refer to our class of algorithms as accelerated G-computation (AGC) algorithms. We describe relationships between AGC algorithms and the warp-speed double bootstrap and certain variants of multiple imputation. The algorithms we develop are general, and can be applied to longitudinal studies with time varying treatments, confounders, and mediators. We illustrate our algorithms on problems in mediation analysis and longitudinal clinical trials.

email: antonio.linero@austin.utexas.edu

ABSTRACTS & POSTER PRESENTATIONS

24. VARIABLE SELECTION: HOW TO CHOOSE?

Sparse Nonparametric Regression with Regularized Tensor Product Kernel

Hang Yu*, University of North Carolina, Chapel Hill
Yuanjia Wang, Columbia University
Donglin Zeng, University of North Carolina, Chapel Hill

Feature selection continues to be an important and challenging problem in the modern era of large scale data. Most of existing methods for feature selection are based on either parametric or semiparametric models, so the performance can severely suffer from model misspecification when high-order nonlinear interactions among features are present. Existing approaches for nonparametric feature selection are limited numbers, computationally intensive and may not converge. In this paper, we propose a novel and computationally efficient approach for nonparametric feature selection based on a tensor-product kernel function over the feature space. The importance of each feature is governed by a parameter in the kernel function which can be efficiently computed iteratively from a modified alternating direction method of multipliers (ADMM) algorithm. We prove the oracle selection property of our method. Finally, we demonstrate superior performance of our method compared to existing methods via simulations and application to the prediction of Alzheimer's disease.

email: hangyu@live.unc.edu

Pursuing Sources of Heterogeneity in Mixture Regression

Yan Li*, University of Connecticut
Chun Yu, Jiangxi University of Finance and Economics
Yize Zhao, Yale University
Weixin Yao, University of California, Riverside
Robert H. Aseltine, University of Connecticut
Kun Chen, University of Connecticut

Mixture regression models have been widely used for modeling mixed regression relationships arising from heterogeneous population. When there are many candidate predictors, it is not only of interest to identify the predictors that are jointly associated with the outcome, but also to find out the sources of heterogeneity, i.e., to identify the ones that have different effects among the clusters and thus are true contributors to the formation of the clusters. To achieve the two objectives simultaneously, we propose a regularized finite mixture effects regression, in which the effect of each predictor is decomposed to a common effect and a set of cluster-specific effect that are constrained to sum up to zero. A sparse estimation of these effects then leads to the identification of two types of variables simultaneously. A generalized EM algorithm is developed, in which a Bregman coordinate descent algorithm is proposed for constrained estimation in the M step. Theoretically, our approach can achieve both estimation and selection consistency. Simulations demonstrate the effectiveness of our method in various scenarios. The new method is illustrated by a public health study.

email: yan.4.li@uconn.edu

An Investigation of Fully Relaxed Lasso and Second-Generation P-Values for High-Dimensional Feature Selection

Yi Zuo*, Vanderbilt University School of Medicine
Jeffrey D. Blume, Vanderbilt University School of Medicine

In high-dimensional settings where inference is desirable, regularization can be used to reduce the feature space. Fully relaxed LASSO retains much of the desirable prediction performance from regularization while yielding a model with interpretable coefficients. Second-generation p-values (SGPV) were proposed in large-scale multiple testing where an interval null hypothesis can be constructed and where it is desirable to indicate when the data support only null, only alternative hypotheses or inconclusive. We explored the degree to which SPGVs can be used in the fully relaxed LASSO: use the SGPVs for feature selection and refit the model on features that survive. Simulations of linear regressions with various data structures and effect sizes were used. Statistical properties such as Type I Error rate, power, false discovery rate (FDR) and coverage rate of outcome associated predictors, were compared. The differences between feature selection with and without SGPVs depended on the ratio of observations to features, effect sizes and the correlation among features. Overall, feature selection with the SGPVs appears to be a viable option and have advantages in certain settings.

email: yi.zuo@vanderbilt.edu

Adaptive Lasso for the Cox Regression with Interval Censored and Possibly Left Truncated Data

Chenxi Li*, Michigan State University
Daewoo Pak, University of Texas MD Anderson Cancer Center
David Todem, Michigan State University

We propose a penalized variable selection method for the Cox proportional hazards model with interval censored data. It conducts a penalized nonparametric maximum likelihood estimation with an adaptive lasso penalty, which can be implemented through a penalized EM algorithm. The method is proven to enjoy the desirable oracle property. We also extend the method to left truncated and interval censored data. Our simulation studies show that the method possesses the oracle property in samples of modest sizes and outperforms available existing approaches in many of the operating characteristics. An application to a dental caries data set illustrates the method's utility.

email: cli@epi.msu.edu

ABSTRACTS & POSTER PRESENTATIONS

Variable Selection for Model-Based Clustering of Functional Data

Tanzy Love*, University of Rochester
Kyra Singh, Google
Eric Hernady, University of Rochester
Jacob Finkelstein, University of Rochester
Jacqueline Williams, University of Rochester

In studying the health effects of radiation, clustering techniques to identify subpopulations with densely sampled functional data are important for detecting late effects of radiation. However, extraneous variables can mask the true group structure. Utilizing a variable selection technique is particularly important in model-based clustering where there is little or no a priori knowledge of the structure or number of groups. We propose a greedy search algorithm to integrate variable selection into the clustering procedure, as in "Variable Selection for Model-Based Clustering" (Raftery and Dean 2006) for functional data. At each step, two models are compared using AICc becasue the number of basis coefficients can be much larger than the sample size. Another difficulty in implementing this approach is the lack of software available for constructing multivariate fully functional linear models of functional data represented by splines. We avoid this obstacle by creating a full model using a series of univariate partial regressions with the "fda" package in R. Our new method successfully finds the most important variables for clustering.

email: tanzy_love@urmc.rochester.edu

Inconsistency in Multiple Regression Model Specifications

Changyong Feng*, University of Rochester
Bokai Wang, University of Rochester
Hongyue Wang, University of Rochester
Xin M. Tu, University of California, San Diego

Model selction is usualy implemented in regression analysis with many cocariates. However, each time when we use different subset of covariates in the model, the form of regression fucntion and the distrbution of the residual part should change accordingly. If the same regression function is assumed for all subsets of covariates, the model specifications will be inconsistent, which makes the estiamtes of parameters difficult to interpret. In this project we discuss how this concistency can easly happen in regression analysis even under the most favorable conditions.

email: changyong_feng@urmc.rochester.edu

C2pLasso: The Categorical-Continuous Pliable Lasso to Identify Brain Regions Affecting Motor Impairment in Huntington Disease

Rakheon Kim*, Texas A&M University
Samuel Mueller, University of Sidney
Tanya Pamela Garcia, Texas A&M University

Developing prediction models with informative predictors is a challenge when the relationship between the predictors and the response depends on other structured modifying variables such as categorical variables. We formalize this problem as the varying-coefficient model selection and propose a novel variable selection method to account for both continuous and categorical modifying variables. Our method is shown to better screen irrelevant variables over the existing method that ignores the structure of modifying variables. Also, our method is designed not only for categorical modifying variables but also for other group-structured modifying variables. Last, our method screens irrelevant variables better than the existing method even for continuous modifying variables. With all these features, our method provides us with a prediction model with higher specificity, lower false discovery rate and lower mean squared error than existing methods. The proposed methodology is applied to data from a Huntington disease study. The result identifies brain regions associated with motor impairment accounting for differentiated relationship by disease severity.

email: rkim@stat.tamu.edu

---

25. FUNCTIONAL DATA ANALYSIS

Covariate-Adjusted Hybrid Principal Components Analysis for EEG Data

Aaron Wolfe Scheffler*, University of California, San Francisco
Abigail Dickinson, University of California, Los Angeles
Shafali Jeste, University of California, Los Angeles
Damla Senturk, University of California, Los Angeles

Electroencephalography (EEG) studies produce region-referenced functional data from electrical signals recorded across the scalp. The data capture neural dynamics that can reveal neurodevelopmental (ND) differences between diagnostic groups, e.g. typically developing (TD) and autism spectrum disorder (ASD) children. Valid inference requires characterization of the complex dependency structure which exhibits covariate-dependent heteroscedasticity (CAD), e.g. differential variation over age. The peak alpha frequency (PAF), the center of a peak in the alpha spectral density, is a biomarker linked to ND that increase in frequency as children age. To model the alpha spectrum across the scalp and chronologically across development, we propose a covariate-adjusted hybrid principal components analysis (CA-HPCA) for EEG data, which utilizes vector and functional PCA while simultaneously adjusting for CAD. CA-HPCA assumes the covariance process is weakly separable conditional on observed covariates, allowing for covariate-adjustments on the marginal covariances rather than the full covariance leading to efficient estimation and insights into ND differences between TD and ASD children.

email: aaron.scheffler@ucsf.edu

ABSTRACTS & POSTER PRESENTATIONS

Evidence-Based Second-Generation P-values on Functional Magnetic Resonance Imaging Data

Ya-Chen Lin*, Vanderbilt University
Valerie F. Welty, Vanderbilt University
Jeffrey D. Blume, Vanderbilt University Medical Center
Kimberly M. Albert, Vanderbilt University Medical Center
Brian D. Boyd, Vanderbilt University Medical Center
Warren D. Taylor, Vanderbilt University Medical Center
Hakmook Kang, Vanderbilt University Medical Center

The large number of multiple tests in imaging analysis can result in large Type II Error rates that preclude important findings when stringent Type I Error controls are used. Commonly used multiple comparison corrections include family-wise error rate control and false discovery rate control, both of which depend on p-values. However, stringent reliance on p-values has been criticized recently in the statistics and neuroscience community. A newly proposed method, the Second-generation p-value (SGPV), overcomes interpretability issues with traditional p-values and has good performance characteristics. Further, by specifying a clinically meaningful region beforehand, the Type I Error rate is naturally controlled and encompasses a proper scientific correction for multiple comparison. For functional imaging analysis, we construct the null interval with the data observed in functionally null region, the cerebrospinal fluid (CSF). In this study, we evaluate the usage of SGPVs in group inference compared to other traditional multiple correction methods. An R shiny app is developed for general usage and easy visualization.

email: ya-chen.lin.1@vanderbilt.edu

Modeling Non-Linear Time Varying Dependence with Application to fMRI Data

Ivor Cribben*, Alberta School of Business

We develop tools for characterizing non-linear time varying dependence between spontaneous brain signals, based on parametric copula models. Specifically, we introduce new methods that detect change points in the R-Vine copula constructed over a multivariate time series, where the number and location of the change points are unknown. We consider various segmentation methods such as binary segmentation and wild binary segmentation. After detecting the change points, we estimate the brain network between each pair of change points. We apply our new methodology to a task-based functional magnetic resonance imaging (fMRI) data set and a resting state fMRI data set. Finally, the time varying models are further used to classify patients and controls and we show that classification results based on the time varying models outperform those on static network models.

email: cribben@ualberta.ca

Average Treatment Effect Estimation with Functional Confounders

Xiaoke Zhang*, The George Washington University
Rui Miao, The George Washington University

Functional regression is an essential tool in functional data analysis. Although causation is of primary interest in many scientific studies, a predominant majority of functional regression approaches can only reveal correlation rather than causation. To fill this gap, this paper studies propensity-score-based estimation for the average treatment effect at the presence of functional confounders. This paper investigates various propensity score models and outcome models via a simulation study and a brain imaging data.

email: xkzhang@gwu.edu

Model-based Statistical Depth with Applications to Functional Data

Weilong Zhao*, Florida State University
Zishen Xu, Florida State University
Yun Yang, University of Illinois at Urbana-Champaign
Wei Wu, Florida State University

Statistical depth, a commonly used analytic tool in non-parametric statistics, has been extensively studied for functional observations over the past few decades. Although various forms of depth were introduced, they are mainly procedure-based whose definitions are independent of the generative model for observations. To address this problem, we introduce a generative model-based approach to define statistical depth for both multivariate and functional data. The proposed model-based depth framework permits simple computation via Monte Carlo sampling and improves the depth estimation accuracy. Specifically, we view functional data as realizations from a second-order stochastic process and define their depths through the eigensystem of the covariance operator. These new definitions are given through a proper metric related to the reproducing kernel Hilbert space of the covariance operator. We propose efficient algorithms to compute the proposed depths and establish estimation consistency. Via real data, we demonstrate that the proposed functional depths reveal important statistical information such as those captured by the quantiles and detect outliers.

e-mail: wz15b@my.fsu.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Inference for Brain Activity from Multi-Resolution Functional Magnetic Resonance Imaging

Andrew Whiteman*, University of Michigan
Jian Kang, University of Michigan
Timothy Johnson, University of Michigan

As a presurgical tool, neurosurgeons often use functional magnetic resonance imaging (fMRI) to map out functionally relevant brain regions. This application requires a high degree of spatial accuracy, but the signal-to-noise ratio (SNR) of fMRI decreases as spatial resolution increases. In practice, both high and standard resolution fMRI may be used, and it is of interest to make more accurate inference on brain activity by combining data with different resolutions. To this end, we develop a new Bayesian model to leverage both better spatial precision in high-resolution fMRI and higher SNR in standard-resolution fMRI. We assume the observed fMRI data measure the mean intensity of brain activity at different resolutions with different levels of noise. We assign a Gaussian process prior to the mean intensity function and develop scalable posterior computation algorithms using a low rank approximation method. We show in simulation our method makes inference on the mean intensity more accurately than full rank models using either high or standard resolution fMRI data alone. We also illustrate our method in analysis of real, de-identified fMRI data from presurgical patients.

e-mail: awhitem@umich.edu

26. PENALIZED AND OTHER REGRESSION MODELS WITH APPLICATIONS

On More Efficient Logistic Regression Analysis via Extreme Ranking

Hani Samawi*, Georgia Southern University

Logistic regression models for dichotomous dependent variables is one of the generalized linear models. They have been frequently applied in several fields. In this chapter, we present more powerful logistic regression models analysis performance when a modified extreme ranked set sampling or moving extreme ranked set sampling are used and further improving the performance when a modified Double Extreme Ranked Set Sampling is used. We propose that ranking could be performed based on an available and easy to rank auxiliary variable which is associated with the response variable. Using the theoretical approach and through simulations, we showed the superiority of the performance of the logistics regression analysis when modified samples are used compared with using the simple random sample. We illustrated the procedure developed using a real dataset.

e-mail: samawi.hani2@gmail.com

Penalized Models for Analysis of Multiple Mediators

Daniel J. Schaid*, Mayo Clinic
Jason P. Sinnwell, Mayo Clinic

Mediation analysis attempts to determine whether the relationship between an independent variable (e.g., exposure) and an outcome variable can be explained, at least partially, by an intermediate variable, called a mediator. Most methods for mediation analysis focus on one mediator at a time, although multiple mediators can be simultaneously analyzed by structural equation models. This has the advantage of accounting for correlations among the mediators. We extend the use of structural equation models for analysis of multiple mediators by creating a penalized model such that the penalty considers the natural groupings of parameters that determine mediation, as well as encourage sparseness of the model parameters. This provides a way to evaluate many mediators and simultaneously select the most impactful mediators, a feature of modern penalized models. Simulations are used to evaluate the benefits and limitations of our approach, and application to a study of genes that mediate response to rubella vaccination illustrate the practical utility of our new methods and software, called regmed.

e-mail: schaid@mayo.edu

Fitting Equality-Constrained, L1-Penalized Models with Inexact ADMM to Find Gene Pairs

Lam Tran*, University of Michigan
Lan Luo, University of Michigan
Hui Jiang, University of Michigan

The usage of gene pairs as a prognostic for cancer patient outcomes has become increasingly popular, but selecting these pairs is computationally expensive. Taking advantage of the lasso's capability for variable selection, gene pair selection can be formulated as an equality-constrained, L1-penalized problem, but conventional algorithms used to solve the lasso are inefficient under the equality constraint. We present ECLasso, an R package that applies an efficient, inexact alternating direction method of multipliers (ADMM) algorithm to t such constrained and L1-penalized Cox proportional hazard models, regressing patient survival times on gene expressions. For other outcomes, the package additionally supports linear, logistic, and Poisson regression models.

e-mail: lamtran@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

A Comparative Analysis of Penalized Linear Mixed Models in Structured Genetic Data

Anna Reisetter*, University of Iowa
Patrick Breheny, University of Iowa

Many genetic studies that aim to identify genetic variants associated with complex phenotypes are subject to confounding due to unobserved factors. This poses a challenge to the detection of multivariate associations of interest and is known to induce spurious associations when left unaccounted for. Linear mixed models (LMMs) are an attractive method to correct for unobserved confounding. These methods simultaneously correct for relatedness and population structure by modeling it as a random effect with a covariance structure estimated from observed genetic data. However, population structure itself does not confound the phenotype-genotype relationship, rather differential environmental exposures among subgroups do. Population structure may or may not serve as a good proxy for such differences. Given these subtle distinctions in confounding sources, the ability of LMMs to accurately estimate fixed, sparse genetic effects has not been well studied. Considering this, we evaluate the performance of penalized LMMs in terms of MSE and false positive rates in the presence of confounding due to varying levels of environmental heterogeneity and relatedness.

e-mail: anna-reisetter@uiowa.edu

A Two-Stage Kernel Machine Regression Model for Integrative Analysis of Alpha Diversity

Runzhe Li*, Johns Hopkins Bloomberg School of Public Health
Ni Zhao, Johns Hopkins Bloomberg School of Public Health

There is increasing evidence that human microbiota play a crucial role in multiple diseases. Alpha diversity captures the richness and evenness of the microbial community. However, analysis results from individual study are always inconsistent due to technical variabilities, it is thus necessary to develop integrative analysis of multiple microbiome datasets to address the biases from sequencing protocols. We propose a two-stage kernel machine regression model to associate alpha diversity with the phenotype of interests. In the first stage, we model the relationship between the alpha diversity and the phenotype via a linear mixed model; in the second stage, we further incorporate the study-specific characteristics through a nonparametric function to allow for the between-study heterogeneities. A joint hypothesis testing is then performed by combining the mean and variance component score statistics. Preliminary simulation studies are conducted to demonstrate the capability of our method to achieve a high power as well as controlling the Type I error. We also apply our method to HIV microbiome dataset to investigate the association between alpha diversity and HIV status.

e-mail: lrz14.thu@gmail.com

Penalized Semiparametric Additive Modeling for Group Testing Data

Karl B. Gregory*, University of South Carolina
Dewei Wang, University of South Carolina
Chris S. McMahan, Clemson University

We consider fitting a sparsity-penalized semiparametric regression model to binary response data, for example to model the disease status of subjects, when instead of observing the true value of the response, we observe the outcomes of several error-prone assays on groups of subjects or individuals, resulting in potentially false diagnoses. We assume an additive structure for the nonparametric components of our semiparametric model, such that each covariate may influence the probabilities of the response outcomes according to a function of unspecified form. We perform variable selection by imposing a sparsity-smoothness penalty on the nonparametric components using the group Lasso. Our methodologies can accommodate any group testing scheme and are an extension of recent advances in regularized linear regression with group testing data to nonparametric regression, which is important for discovering nonlinear relationships between disease outcomes and continuous-valued covariates.

e-mail: gregorkb@stat.sc.edu

Penalized Likelihood Logistic Regression with Rare Events-An Application to the Regeneration Dynamics of Pine Species in Oak-Pine Forest Types

Dilli Bhatta*, University of South Carolina Upstate

Logistic regression is one of the most popular ways to model binary data. The parameter of logistic regression are obtained using standard maximum likelihood estimation (MLE) method. However, with rare events (i.e. the proportion of 1's is smaller than five percent), MLE estimates could be biased and can sharply underestimate the probability of rare events. To overcome this problem, Penalized maximum likelihood estimation (PMLE) proposed by Firth (1993) is useful. In this paper we investigate the vegetation structure and seedling dynamics of shortleaf, slash, and longleaf pine in oak-pine forest types of the southeastern United States. We used logistic regression with penalized maximum likelihood estimation because of the rarity of the regeneration from these pines. We also determined the factors that influence the regeneration dynamics of these species.

e-mail: dbhatta@uscupstate.edu

ABSTRACTS & POSTER PRESENTATIONS

27. METHODS FOR NEUROIMAGING DATA: GET THE PICTURE?

Letting the LaxKAT Out of the Bag: Packaging, Simulation, and Neuroimaging Data Analysis for a Powerful Kernel Test

Jeremy S. Rubin*, University of Maryland, Baltimore County
Simon Vandekar, Vanderbilt University
Lior Rennert, Clemson University
Mackenzie Edmonson, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania

Biomedical research areas including genomics and neuroimaging often have a number of independent variables that is much greater than the sample size. The sequence kernel association test (KAT) and sum of scores tests can offer improved power in this feature setting; however, power is significantly reduced in the presence of a large number of unassociated independent variables. We propose the Linear Maximal KAT (LaxKAT), which maximizes the KAT test statistic over a subspace of linear kernels to increase power. A permutation testing scheme is used to estimate the null distribution of the LaxKAT statistic and perform hypothesis testing. Calculation of the LaxKAT was implemented using a combination of the R and C++ programming languages. We find that this test has power and controls the type I error for different sample sizes and signal distributions. It is expected that the LaxKAT will have competitive power relative to other high-dimensional testing procedures when applied to detect predictors of memory impairment in cortical thickness measurements from the Alzheimer's Disease Neuroimaging Initiative study (ADNI).

email: jrub1@umbc.edu

Comparison of Two Ways of Incorporating the External Information via Linear Mixed Model Design with Application in Brain Imaging

Maria Paleczny*, Institute of Mathematics of the Jagiellonian University

The significant difficulty in the brain imaging research is the problem of incorporating information from different sources. To better understand the nature of the issue, and to keep mutual information, it is important to analyze all the data together. A novel method is presented in the paper to address those challenges. It incorporates the external information about the connections between brain regions into the linear mixed model. The other regularization technique focused on presented problem is also described, as it is the starting point for new method. Both approaches are compared via simulations and their similarity is described by the theorem.

email: maria.paleczny@gmail.com

Interpretable Classification Methods for Brain-Computer Interface P300 Speller

Tianwen Ma*, University of Michigan
Jane E. Huggins, University of Michigan
Jian Kang, University of Michigan

A brain-computer interface (BCI) is a device allowing human brains to communicate with computers directly. This technology is used for disabled people to control computers. The fundamental statistical problem in BCI is classification. In the electroencephalography (EEG) BCI system, the P300-event related potential (ERP) is an elicited response by alternating regular and target stimuli, which is a reliable signal for BCI classification. Many ML methods were applied to construct classifiers, but few provided insights of underlying mechanism. In this work, we propose a new statistical method to model the conditional distributions of EEG signals with P300-ERPs by BCI system design, from which the predictive probability of signals can be derived. Our method can detect time periods when EEG signals have strong predictive power, providing a potential better understanding of neural activities from different brain regions in response to computer-interaction stimuli. Simulation studies and real data application demonstrate the advantage of the proposed method compared with existing alternatives.

email: mtianwen@umich.edu

Copula Random Field with Application to Massive Neuroimaging Data Analysis

Jie He*, University of Michigan
Jian Kang, University of Michigan
Peter X.-K Song, University of Michigan

Motivated by the needs of analyzing large-scale and complex imaging data, we introduce a new spatial random field model: the copula random field that describes the dependence among arbitrarily many spatially distributed random variables independently of their marginal distributions. In particular, we propose a copula based image-on-scalar regression model to estimate spatially varying effects of scalar covariates taking into account for complex spatial dependence in imaging data. The proposed model provides a general framework for the analysis of different types of imaging outcomes. For model inference, we adopt the sieve composite likelihood approach and establish the asymptotical theory. We develop a computationally feasible algorithm which performs maximizations by parts in the sieve composite likelihood function. The computation is scalable to high resolution images. We evaluate the performance of the proposed model and the estimation method via extensive simulation studies. We demonstrate the usefulness of the proposed method by analyzing the fALFF data in the Autism Brain Imaging Data Exchange (ABIDE) study.

email: jiehe@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Neural Networks Guided Independent Component Analysis with Application to Neuroimaging

Daiwei Zhang*, University of Michigan
Ying Guo, Emory University
Jian Kang, University of Michigan

There is an emerging interest in utilizing neural networks (NNs) to predict clinical outcomes from neuroimaging data. Typical neuroimaging studies cannot obtain the sample size required for training deep NNs. While independent component analysis (ICA) can be used to extract imaging features for prediction, standard ICA methods decompose imaging data without incorporating the clinical outcomes, which may not be optimal for a prediction model. In this work, we propose a joint Bayesian ICA-NN model that simultaneously extracts latent imaging features and predicts clinical outcomes. The simple but intepretable NN architecture of our model can separate the linear from the nonlinear effects of latent imaging features. Horseshoe priors are assigned to the NN weights to impose sparsity. In addition, we develop an efficient posterior computation algorithm for model inference and prediction. Through extensive simulation studies, we demonstrate that ICA-NN outperforms deep learning, random forest, support vector machine, and other prediction methods using imaging data with small sample sizes. We also illustrate ICA-NN via an analysis of the fMRI data in the Human Connectome Project.

email: daiweiz@umich.edu

Removal of Scanner Effects in Covariance of Neuroimaging Measures

Andrew Chen*, University of Pennsylvania
Haochang Shou, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania

To acquire larger samples for answering complex questions in neuroscience, researchers have increasingly turned to multi-site neuroimaging studies. However, these studies are marred by the existence of scanner effects in the raw images and derived measurements. These effects have been shown to hinder comparison between sites, mask biologically meaningful associations, and even introduce unwanted associations. Previous methods have focused on harmonizing the mean and variance of measurements across sites in order to remove these effects, but none have addressed how covariance between measurements can vary across scanners. Using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, we show that considerable differences in covariance exist across sites and that current harmonizations do not address this issue. We then propose a new method that is able to harmonize covariance across sites and demonstrate that it performs better on both removal of scanner effects and detection of clinically relevant associations.

email: andrewac@pennmedicine.upenn.edu

Classifying Brain Edema with Low-Resolution MRI

Danni Tu*, University of Pennsylvania
Dylan Small, University of Pennsylvania
Manu S. Goyal, Washington University School of Medicine, St. Louis
Theodore Satterthwaite, University of Pennsylvania
Kelly Clark, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania

For high-resolution, high-quality magnetic resonance images (MRI), state-of-the-art approaches to extract biomarkers from imaging data often work well. However, in low resource settings, the only MRI machines available may have low resolution. In the case of cerebral malaria (a complication of infection by the malaria parasite) in Malawi, the severity of cerebral edema is currently scored based on a low resolution 0.35 tesla MRI. Because manual scoring is time-consuming and there are currently few neuroradiologists in sub-Saharan Africa, it is of interest to automate the scoring. Automation, however, is complicated by the fact that many standard imaging pipelines perform inadequately on low-quality, noisy data. We propose a method to first process this low-quality imaging data in a way that borrows strength from high-quality brain atlases from other studies. We then assess edema using volume- and intensity-based measures. Finally, we develop a classification method to identify severe cases of cerebral edema.

email: danni.tu@pennmedicine.upenn.edu

28. CAUSAL EFFECT ESTIMATION

Assessing Exposure Effects on Gene Expression

Sarah A. Reifeis*, University of North Carolina, Chapel Hill
Michael G. Hudgens, University of North Carolina, Chapel Hill
Karen L. Mohlke, University of North Carolina, Chapel Hill
Michael I. Love, University of North Carolina, Chapel Hill

In observational genomics datasets, there is often confounding of the effect of an exposure on gene expression. To adjust for confounding when estimating the exposure effect, a common approach involves including potential confounders as covariates with the exposure in a regression model of gene expression. However, when the exposure and confounders interact to influence gene expression, the fitted regression model does not necessarily estimate the overall effect of the exposure. Using inverse probability weighting (IPW) or the parametric g-formula in these instances is straightforward to apply and yields consistent effect estimates. IPW can readily be integrated into a genomics data analysis pipeline with upstream data processing and normalization, while the g-formula can be implemented by making simple alterations to the regression model. The regression, IPW, and g-formula approaches to exposure effect estimation are compared herein using simulations; advantages and disadvantages of each approach are explored. The methods are applied to a case study estimating the effect of current smoking on gene expression in adipose tissue.

email: sreifeis@live.unc.edu

ABSTRACTS & POSTER PRESENTATIONS

Sensitivity of Clinical Trial Estimands under Imperfect Compliance

Heng Chen*, Southern Methodist University
Daniel F. Heitjan, Southern Methodist University and University of Texas, Southwestern

The selection of an estimand is challenging in clinical trials with noncompliance. The intention-to-treat (ITT) analysis is a pragmatic approach, but the ITT estimand cannot measure the causal effect of the actual treatment received and is sensitive to the level of compliance. An alternative is the complier average causal effect (CACE), which refers to the average effect of treatment received in the latent subset of subjects who would comply with either treatment. Under the Rubin Causal Model (RCM), five assumptions are sufficient to identify CACE, permitting its consistent estimation from trial data. We observe that CACE can also vary with the fraction of compliance when the compliance class is regarded as a random quantity. We propose a "sixth assumption" which guarantees the robustness of CACE to the compliance fraction. We conduct a simulation study to illustrate the phenomenon and to assess the robustness of ITT and CACE to different proportions of compliers when the "sixth assumption" is violated. We observe that only CACE can be robust to varying levels of compliance, and only when the assumption is satisfied.

email: hengc@smu.edu

Borrowing from Supplemental Sources to Estimate Causal Effects from a Primary Data Source

Jeffrey A. Boatman*, University of Minnesota
David M. Vock, University of Minnesota
Joseph S. Koopmeiners, University of Minnesota

The increasing multiplicity of data sources offers exciting possibilities in estimating the effects of a treatment or exposure, particularly if observational and experimental sources can be used simultaneously. Borrowing between sources can potentially result in more efficient estimators, but it must be done in a principled manner to mitigate increased bias and Type I error. Furthermore, when the effect of treatment is confounded, as in observational sources or in clinical trials with noncompliance, causal estimators are needed to simultaneously adjust for confounding and to estimate effects across data sources. We consider the problem of estimating causal effects from a primary source and borrowing from any number of observational sources. We propose using linear regression and regression-tree estimators that borrow by assuming exchangeability of the regression coefficients or parameters between data sources. Borrowing is accomplished with multisource exchangeability models and Bayesian model averaging. We introduce the estimators and apply them to recently completed clinical trials investigating the effects of very low nicotine content cigarettes on smoking behavior.

email: boat0036@umn.edu

Estimating Causal Treatment Effects: A Bayesian Inference Approach Adopting Principal Stratification with Strata Predictive Covariates

Duncan C. Rotich*, University of Kansas Medical Center
Bin Dong, Janssen Research & Development
Jeffrey A. Thompson, University of Kansas Medical Center

Treatment effect in randomized clinical trials is often evaluated after post-randomization intercurrent events such as treatment discontinuation. Without appropriate adjustment for intercurrent events, the estimand of treatment effect is likely to be subject to bias and therefore misleading since it no longer reflects treatment causal effect. Recent revisions of the ICH E9/R1 guidelines on estimands also emphasizes the importance of this adjustment to ensure statistical validity and clinical meaningfulness of estimated treatment effect. Generally, there are two ways to evaluate causal effect: potential outcomes and principal stratification frameworks. We adopt the principal stratification framework where we first identify the latent strata membership based on observed baseline characteristics and then evaluate treatment effect within the stratum. Since the true causal effect of a treatment is not known in a real setting, we assessed the performance of our approach using simulations by comparing it with standard methods using ATE without intercurrent event adjustment. The results showed a reduction in treatment effect bias of up to 89% as compared to standard approaches.

email: duncancheru@gmail.com

Estimating Causal Effects in the Presence of Positivity Violations

Yaqian Zhu*, University of Pennsylvania
Nandita Mitra, University of Pennsylvania
Jason Roy, Rutgers University

In observational studies, whether a subject is given treatment or control depends on various factors. The lack of variability in treatment assignment for certain groups threatens the identifiability of causal effects as it violates the positivity assumption, which requires the conditional probability of treatment to be bounded from 0 and 1. These violations may increase the bias and variance of estimates. Popular methods for dealing with random violations due to finite sample non-overlap include trimming and weighting; an alternative involves model extrapolation. While these methods have been implemented in practice, there has been limited work examining the impact that uncertainty about design and analysis decisions regarding overlap has on inference. We provide a comprehensive review and evaluation of methods for estimating causal effects in the presence of positivity violations. Through simulation studies, we compare how defining overlap based on propensity scores or covariate distances and including instruments affect estimates of ATE in terms of bias, variance, and coverage. We also give insights into which methods may be more appropriate for different settings.

email: yazhu@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Estimating Causal Effect of Multiple Treatments with Censored Data in Observational Studies

Youfei Yu*, University of Michigan
Min Zhang, University of Michigan
Bhramar Mukherjee, University of Michigan

In observational studies comparing two or more treatments, estimating causal treatment effect is prone to selection bias. In addition to confounding effect, another potential source of selection bias is non-randomly censored observations. Methods based on propensity score (PS) are widely used to adjust for confounders in the binary treatment case. However, the applications of PS methods in multiple treatment studies remain limited, and the relative performance of these methods remain unclear for comparing multiple treatments, especially when the outcome of interest is binary. To correct the bias due to censoring, we combine the existing PS methods with weighting by inverse probability of not being censored (IPCW) given baseline covariates for each treatment. We conduct simulation studies with multiple treatments to compare different propensity score methods combined with IPCW. We apply these methods to estimate the effect of four common treatments for castration-resistant advanced stage prostate cancer, using claims data from a large national private health insurance network with the outcome being admission to the emergency room within a short time window of treatment initiation.

email: youfeiyu@umich.edu

29. NEW PERSPECTIVES ON DATA INTEGRATION IN GENOME-WIDE ASSOCIATION STUDIES

TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits

Jingjing Yang*, Emory University School of Medicine
Sini Nagpal, Georgia Institute of Technology
Xiaoran Meng, Emory University School of Public Health
Shizhen Tang, Emory University School of Public Health
Gregory C. Gibson, Georgia Institute of Technology
David A. Bennett, Rush University Medical Center
Philip L. De Jager, Columbia University
Aliza P. Wingo, Atlanta VA Medical Center
Thomas S. Wingo, Emory University School of Medicine
Michael P. Epstein, Emory University School of Medicine

Transcriptome-wide association studies (TWAS) have enhanced the discovery of genetic risk loci for complex traits. The existing tools like PrediXcan and FUSION use parametric imputation models that have limitations for modeling the complex genetic architecture of transcriptomic data. To improve on this, we employ a nonparametric Bayesian method that was originally proposed for genetic prediction of complex traits, assuming a data-driven nonparametric prior for cis-eQTL effect sizes. Our simulation studies showed that the nonparametric Bayesian model improved both imputation for transcriptomic data and the TWAS power over PrediXcan when 1% cis-SNPs coregulate gene expression and gene expression heritability 0.2. In real applications, the nonparametric Bayesian method fitted transcriptomic imputation models for 57.8% more genes over PrediXcan with improved TWAS power. We implement both parametric PrediXcan and nonparametric Bayesian methods in a convenient software tool TIGAR, which imputes transcriptomic data and performs subsequent TWAS using individual-level or summary-level GWAS data by both burden and sequence kernel association tests (SKAT).

email: jingjing.yang@emory.edu

Integrating Gene Expression Regulatory Variation Across Populations and Tissues to Understand Complex Traits

Heather E. Wheeler*, Loyola University Chicago

Our broad goal is to better understand how genetic variation leads to phenotypic variation for complex traits including disease susceptibility and drug response. We develop systems approaches, including PrediXcan, for complex trait association studies by building computational models that leverage and integrate similarity in genetic, transcriptomic or other omics-level data. Our current focus is understanding the degree of transferability of genetic association results and implicated genes across diverse populations and tissues. We recently investigated the underlying genetic architecture of gene expression by optimizing gene expression prediction within and across diverse populations and tissues through statistical machine learning. Sparse models perform better than polygenic models for gene expression prediction from genotypes. Prediction models that integrate expression from multiple tissues identify more complex trait associations than single tissue models. We further show that a training set with ancestry similar to the test set results in better gene expression predictions, demonstrating the need to incorporate diverse populations in genomic studies.

email: hwheeler1@luc.edu

Transcriptome-Wide Transmission Disequilibrium Analysis Identifies Novel Risk Genes for Autism Spectrum Disorder

Qiongshi Lu*, University of Wisconsin, Madison
Kunling Huang, University of Wisconsin, Madison
Yuchang Wu, University of Wisconsin, Madison

We introduce a novel statistical framework to quantify the transmission disequilibrium of genetically regulated transcriptomic activities from parents to offspring. We generate three pseudo-sibling controls for each proband based on phased parental genotypes, impute gene expression levels in multiple brain tissues, and use conditional logistic regression to identify over- or under-transmission of imputed gene expression. We applied our method to conduct a transcriptome-wide association study (TWAS) in 7,805 ASD trios and replicated findings using data from the iPSYCH project (N=35,740). Meta-analysis identified 31 transcriptome-wide significant associations. Among the identified genes, transcription factor POU3F2 is a master regulator of a brain gene expression module associated with psychiatric disorders. The identified genes showed minimal overlap with loss-of-function intolerant genes or known ASD genes enriched for pathogenic mutations. However, target genes regulated by POU3F2 are enriched for known risk genes for ASD. These results provide fundamental new insights into the genetic basis of ASD.

email: qiongshi.lu@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Model Checking and More Powerful Inference in Transcriptome-Wide Association Studies

Wei Pan*, University of Minnesota

Transcriptome-wide association studies (TWAS, or PrediXcan) have been increasingly used to identify causal genes by integrating GWAS with eQTL data. The basic methodology underlying TWAS (and Mendelian randomization, MR) is the (two-sample) two-stage least squares (2SLS) instrumental variables regression for causal inference, which imposes strong assumptions on the SNPs to be valid instrumental variables (IVs). These assumptions are most likely to be violated in practice, e.g. due to widespread horizontal pleiotropy of the SNPs. We first consider some simple and powerful methods to detect invalid IVs/SNPs, then propose more robust and more powerful methods than existing ones for causal inference in the presence of invalid IVs.

email: panxx014@umn.edu

30. ADVANCES IN CAUSAL INFERENCE AND JOINT MODELING WITH SURVIVAL AND COMPLEX LONGITUDINAL DATA

Causal Proportional Hazards Estimation with a Binary Instrumental Variable

Limin Peng*, Emory University
Behzad Kianian, Emory University
Jung In Kim, University of North Carolina, Chapel Hill
Jason Fine, University of North Carolina, Chapel Hill

Instrumental variables (IV) are a useful tool for estimating causal effects in the presence of unmeasured confounding. In this work, we develop a simple causal hazard ratio estimator in a proportional hazards model with right censored data. The method exploits a special characterization of IV which enables the use of an intuitive inverse weighting scheme that is generally applicable to more complex survival settings with left truncation, competing risks, or recurrent events. We establish the asymptotic properties of the estimators, and provide plug-in variance estimators. The proposed method can be implemented in standard software. The finite sample performance of the proposed method was evaluated through extensive simulation studies. We also illustrate the new method via an application to a data set from the Prostate, Lung, Colorectal and Ovarian cancer screening trial.

email: lpeng@sph.emory.edu

Joint Modeling of Zero-Inflated Longitudinal Microbiome and Time-to-Event Data

Huilin Li*, New York University
Jiyuan Hu, New York University
Chan Wang, New York University
Martin Blaser, Rutgers University

Recently more and more longitudinal microbiome studies are conducted to identify candidate microbes as biomarkers for the disease prognosis. We propose a novel joint modeling framework JointMM for longitudinal microbiome and time-to-event data to investigate the effect of dynamic changes of microbiome abundance profile on disease onset. JointMM comprises of two sub-models, i.e., the zero-inflated scaled-Beta mixed-effects regression sub-model aimed at depicting the temporal structure of microbial abundances among subjects; and the survival sub-model to characterize the occurrence of disease and its relationship with microbiome abundances changes. JointMM is specifically designed to handle the zero-inflated and highly skewed longitudinal microbiome abundance data and exhibits better interpretability that JointMM can examine whether the temporal microbial presence/absence pattern and/or the abundance dynamics would alter the time to disease onset. Comprehensive simulations and real data analyses demonstrated the statistical efficiency of JointMM compared with competing methods.

email: huilin.li@nyumc.org

Causal Comparative Effectiveness Analysis of Dynamic Continuous-Time Treatment Initiation Rules with Sparsely Measured Outcomes and Death

Liangyuan Hu*, Icahn School of Medicine at Mount Sinai
Joseph W. Hogan, Brown University

We leverage a large observational data and compare, in terms of mortality and CD4 cell count, the dynamic treatment initiation rules for HIV-infected adolescents. Our approaches extend the marginal structural model for estimating outcome distributions under dynamic treatment regimes (DTR), developed in Robins et al. (2008), to allow the causal comparisons of both specific regimes and regimes along a continuum. Furthermore, we propose strategies to address three challenges posed by the complex dataset: continuous-time measurement of the treatment initiation process; sparse measurement of longitudinal outcomes of interest, leading to incomplete data; and censoring due to dropout and death. We derive a weighting strategy for continuous time treatment initiation; use imputation to deal with missingness caused by sparse measurements and dropout; and dene a composite outcome that incorporates both death and CD4 count as a basis for comparing treatment regimes. Our analysis suggests that immediate ART initiation leads to lower mortality and higher median values of the composite outcome, relative to other initiation rules.

email: liangyuan.hu@mountsinai.org

ABSTRACTS & POSTER PRESENTATIONS

31. OPPORTUNITIES AND CHALLENGES IN THE ANALYSIS AND INTEGRATION OF LARGE-SCALE BIOBANK DATA

Empowering GWAS Analysis with Missing Data Using Surrogate Phenotypes in Biobanks

Xihong Lin*, Harvard University
Zachary McCaw, Google

Incomplete observation of the target outcome is common in GWAS in biobanks. Jointly modeling this target outcome with a correlated surrogate can help to improve power. We develop Surrogate Phenotype Regression Analysis (SPRAY) for leveraging information from a single surrogate to improve inference on the target outcome by incorporating information from multiple surrogates through the concept of Synthetic Surrogate Analysis (SSA). In SSA, the candidate surrogates are combined into a single summary measure, the synthetic surrogate, which is jointly analyzed with the target outcome. The synthetic surrogate constitutes a prediction of the target outcome as a function of the candidate surrogates, or using an unsupervised reduction of the candidate surrogates, such as the leading principal component. Moreover, we introduce a computationally efficient least squares algorithm for performing SPRAY when missingness is confined to the target outcome only. We perform GWAS of lung function traits in the UK Biobank (UKB) using both supervised and unsupervised SSA.

email: xlin@hsph.harvard.edu

Fast and Efficient Generalized Estimating Equations for Fitting Non-Linear Model to Biobank Scale Data

Nilanjan Chatterjee*, Johns Hopkins University
Diptavo Dutta, Johns Hopkins University

We consider a generalized estimating equation approach for association analysis for binary and time-to-even outcomes in UK Biobank and other similar studies accounting for relatedness of individuals. We propose to estimate association parameters under a marginal mean model, while efficiently weighting individuals based on a derived variance-covariance matries of the traits unders a liability threshold conidtional model. The resulting method has similar computational complexitiy as linear mixed models and yet can produce valid inference on parameters on various popular non-linear models such as the logistic regression or the Cox proportional hazrd model. We will evaluate the performance of the methods through simulation studies and various applications in the UK Biobank dataset.

email: nilanjan@jhu.edu

Modeling Functional Enrichment Improves Polygenic Prediction Accuracy in UK Biobank and 23andMe Data Sets

Carla Marquez-Luna*, Icahn School of Medicine at Mount Sinai
Steven Gazal, Harvard T.H. Chan School of Public Health
Po-Ru Loh, Brigham and Women's Hospital and
Harvard Medical School
Samuel S. Kim, Massachusetts Institute of Technology
Nicholas Furlotte, 23andMe Inc.
Adam Auton, 23andMe Inc.
Alkes L. Price, Harvard T.H. Chan School of Public Health

Genetic variants in functional regions of the genome are enriched for complex trait heritability. We introduce a new method for polygenic prediction, LDpred-funct, that leverages trait-specific functional enrichments to increase prediction accuracy. We fit priors from the baseline-LD model, which includes coding, conserved, regulatory and LD-related annotations. We analytically estimate posterior mean causal effect sizes and then use cross-validation to regularize these estimates, improving prediction accuracy for sparse architectures. We applied LDpred-funct to predict 21 highly heritable traits in the UK Biobank. We used association statistics from British-ancestry samples as training data (avg N=365K) and samples of other European ancestries as validation data (avg N=22K), to minimize confounding. LDpred-funct attained a +9% relative improvement in average prediction accuracy (avg R2=0.145) compared to LDpred (the best method that does not incorporate functional information), consistent with simulations. Our results show that modeling functional enrichment improves polygenic prediction accuracy, consistent with the functional architecture of complex traits.

email: cmarquezluna@alumni.harvard.edu

Handling Sampling and Selection Bias in Association Studies Embedded in Electronic Health Records

Bhramar Mukherjee*, University of Michigan
Lauren J. Beesley, University of Michigan

In this talk we will discuss statistical challenges and opportunities with joint analysis of electronic health records and genomic data through "Genome and Phenome-Wide Association Studies (GWAS and PheWAS)". We posit a modeling framework that helps us to understand the effect of both selection bias and outcome misclassification in assessing genetic associations across the medical phenome. We will propose various inferential strategies that handle both sources of bias to yield improved inference. We will use data from the UK Biobank and the Michigan Genomics Initiative, a longitudinal biorepository at Michigan Medicine, launched in 2012 to illustrate the analytic framework. The examples illustrate that understanding sampling design and selection bias matters for big data, and are at the heart of doing good science with data. This is joint work with Lauren Beesley at the University of Michigan.

email: bhramar@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

32. COMPOSITIONAL NATURE OF MICROBIOME DATA: CHALLENGES AND NEW METHODS

Association Testing for Longitudinal Multiomics Data

Anna M. Plantinga*, Williams College

The human microbiota plays an important role in health and disease, but the mechanism of association is often not known. Recent studies combining microbiome data with other omics data sources (host gene expression, metabolomics, etc.) attempt to provide more clarity. These studies are increasingly often longitudinal in nature, in an attempt to reduce unmeasured confounding, but few methods are available for testing the association between the microbiome and other structured high-dimensional data types across time. We propose a combined kernel RV approach that tests for longitudinal association between the microbiota and genomic features, such as host genetic variants or gene expression. We also assess the importance of accounting for data compositionality in this analysis setting. The method is tested on both simulated and real microbiome data.

email: anna.plantinga@gmail.com

Scalable Inference for Count Compositional Microbiome Data

Justin D. Silverman*, Duke University

Due to the measurement process, microbiome data contains information regarding only the relative abundances of taxa. Commonly such relative data is analyzed using tools from compositional data analysis (CoDA). Yet the CoDA approach fails to account for other features of the data such as count variability and technical variation. In this talk, I will introduce an alternative formulation of sequence count data as count-compositional and will introduce tools in line with this formulation. Based on the compound multinomial logistic-normal distribution I will introduce a class of Bayesian models for the analysis of sequence count data. While such models are typically difficult to fit, I will introduce the collapse-uncollapse sampler as a means of efficiently inferring these models.

email: Justin.Silverman@duke.edu

Robust and Powerful Differential Composition Tests on Clustered Microbiome Data

Zhengzheng Tang*, University of Wisconsin, Madison
Guanhua Chen, University of Wisconsin, Madison

Clustered microbiome data have become prevalent in recent years from designs such as longitudinal studies, family studies, and matched case-control studies. The within-cluster dependence compounds the challenge of the microbiome data analysis. Methods that properly accommodate intra-cluster correlation and features of the microbiome data are needed. We develop robust and powerful differential composition tests for clustered microbiome data. The methods do not rely on any distributional assumptions on the microbial compositions, which provides flexibility to model various correlation structures among taxa and among samples within a cluster. By leveraging the adjusted sandwich covariance estimate, the methods properly accommodate sample dependence within a cluster. Different types of confounding variables can be easily adjusted for in the methods. We perform extensive simulation studies under commonly-adopted clustered data designs to evaluate the methods. The usefulness of the proposed methods is further demonstrated with a real dataset from a longitudinal microbiome study on pregnant women.

email: ztang2@wisc.edu

33. STATISTICAL MODELING IN ALZHEIMER'S DISEASE

Bent Lines and Quantiles in Longitudinal Modeling of Alzheimer's Progression

Rick Chappell*, University of Wisconsin, Madison

Alzheimer's disease (AD) is often investigated using times to clinical symptoms such as mild or full cognitive impairment. These are expressed either as binary states or through proxies such cognitive test scores. Questions about such outcomes may be more usefully answered with inference on quantiles than with expectations. For example, although mean cognitive scores may decline somewhat with age, 20th percentiles of scores show sharper declines which correlate with clinical AD diagnosis. In addition, such declines may show sudden downturns or bends. The purpose of this talk is to summarize existing results and discuss current research.

email: chappell@stat.wisc.edu

Partly Conditional Modeling for Ordinal Outcomes with Application to Alzheimer's Disease Progression

Dandan Liu*, Vanderbilt University
Jacquelyn Neal, Vanderbilt University

Alzheimer's disease progression could be measured using transitions between multiple clinical/pathological disease states and is often modeled using panel data. Existing methods modeling disease progression usually rely on the Markov assumption where disease progression in the future is independent of previous disease history given the current state information. Our preliminary work suggests this assumption may be violated for the AD disease progression. In addition, some risk factors for AD and their effects might change over time making it even more challenging to predict disease progression. Methods that could relax Markov assumption and incorporate time-dependent risk factors and time-varying effects while providing risk prediction is warranted. In this work, we extended partly conditional model that was developed for binary and survival outcomes to longitudianl ordinal outcome. The proposed method was compared with existing methods using simulation studies and was applied to National Alzheimer's Coordinating Center UDS data to model clinical disease progression of AD.

email: dandan.liu@vumc.org

ABSTRACTS & POSTER PRESENTATIONS

Leveraging Disease Progression Modeling to Improve Clinical Trial Design in Alzheimer's Disease

Barbara Wendelberger*, Berry Consultants
Melanie Quintana, Berry Consultants
Scott Berry, Berry Consultants

Designing well-powered clinical trials in progressive diseases is contingent upon understanding the rate of progression of key clinical endpoints, as well as their heterogeneity. This is epitomized in the field of Alzheimer's disease. In this presentation, we describe how disease progression modeling can provide realistic estimates of disease state, quantify potential disease modifications, and lead to better informed and more powerful clinical trials. We present a concrete example of the benefits of using disease progression modeling for clinical trial design in autosomal dominant Alzheimer's disease and discuss extending this approach to sporadic Alzheimer's disease.

email: barbara@berryconsultants.net

Integrative Modeling and Dynamic Prediction of Alzheimer's Disease

Sheng Luo*, Duke University
Kan Li, Merck & Co., Inc.

This paper is motivated by combining serial neurocognitive assessments and other clinical variables for monitoring the progression of Alzheimer's disease (AD). We propose a novel framework for the use of multiple longitudinal neurocognitive markers to predict the progression of AD. The conventional joint modeling is not applicable when there is a large number of longitudinal outcomes. We introduce various approaches based on the functional principal component analysis for dimension reduction and feature extraction from multiple longitudinal outcomes. We use these features to extrapolate the health outcome trajectories and use scores on these features as predictors in a survival model to conduct predictions over time. We propose a personalized dynamic prediction framework that can be updated as new observations collected to reflect the patient's latest prognosis, and thus intervention could be initiated in a timely manner. Simulation studies and application to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset demonstrate the robustness of the method for the prediction of future health outcomes and risks of target events.

email: sheng.luo@duke.edu

---

34. RECENT ADVANCES IN BAYESIAN METHODS FOR SPATIAL-TEMPORAL PROCESSES

Multivariate Disease Mapping using Directed Acyclic Graph Autoregressive Models

Abhi Datta*, Johns Hopkins University

Hierarchical models for regionally aggregated disease incidence data commonly involve region specific latent random effects that are modeled jointly as having a multivariate Gaussian distribution. The covariance or precision matrix incorporates the spatial dependence between the regions. We propose a new scalable parametric model for the precision matrix --- directed acyclic graph autoregressive model (DAGAR) which establishes a link between the parameters used and the variance and covariances of the random effects. We extend DAGAR to multivariate settings for joint analysis of multiple disease data. The extension leverages the sparse Cholesky factors of DAGAR precision matrices and ensures that computational burden remains manageable, allows disease-specific spatial correlation parameters while retaining parameter interpretability of the univariate DAGAR model. We present simulation studies and an application.

email: abhidatta@jhu.edu

Modeling Heroin-Related EMS Calls in Space and Time

Zehang Richard Li*, Yale School of Public Health
Forrest Crawford, Yale School of Public Health
Gregg Gonsalves, Yale School of Public Health

Opioid use and overdose have become an important public health issues in the United States. However, understanding the spatial and temporal dynamics of opioid overdose incidents and effects of public health interventions and policy changes can be challenging. Effects may be heterogeneous across space and time, and may exhibit spillover into regions in which the intervention did not take place. Using a publicly available dataset consisting of the time, location, and nature of heroin-related emergency calls in the city of Cincinnati, Ohio, we discuss considerations in mapping the risk of overdose in small areas over time, and models to characterize the dynamics of overdose incidents. We will also outline a framework for estimating causal impacts of public health interventions from surveillance data under spatialtemporal confounding.

email: lizehang@gmail.com

Bayesian Spatial Prediction of Collective Efficacy Across an Urban Environment

Catherine Calder*, University of Texas, Austin

In sociology, 'collective efficacy' refers to the level of social cohesion among neighbors and their willingness to intervene on behalf of the common good and has been used to explain various neighborhood-level social processes (e.g., crime). Traditionally, measures of the collective efficacy of a neighborhood are derived from sample surveys of neighborhood residents, which are analyzed using traditional multilevel (spatial) statistical models. In this talk, we introduce a novel approach to estimating collective efficacy across an urban environment using unique data collected as part of the Adolescent Health and Development in Context (AHDC) Study in Columbus, OH, USA. We introduce a Bayesian hierarchical model that accommodates the cross-classified nature of the AHDC reports of collective efficacy at routine activity locations of study participants.

email: calder@austin.utexas.edu

ABSTRACTS & POSTER PRESENTATIONS

Estimating Subnational Variation in Health Indicators in a Low- and Medium-Income Countries Setting

Jon Wakefield*, University of Washington

In countries without vital registration systems, subnational estimates of helath indicators are based on data from surveys, health facilities, sample registration systems and censuses. However, these data sources each have their own idiosyncrasies and modeling potential biases is essential though requires great care. Examples that will be presented include estimating HIV prevalence and child mortality. The presentation of results will also be discussed. In particular, we will discuss the granularity at which results should be given, and how uncertainty can be conveyed.

email: jonno@uw.edu

35. SPEED POSTERS: EHR DATA, EPIDEMIOLOGY, PERSONALIZED MEDICINE, CLINICAL TRIALS

35a. Extending Difference-in-Difference Methods to Test the Impact of State-Level Marijuana Laws on Substance Use Using Published Prevalence Estimates

Christine M. Mauro*, Columbia University Mailman School of Public Health
Melanie M. Wall, Columbia University Mailman School of Public Health

Colorado and Washington legalized the use of recreational marijuana in 2012; since then eight more states and DC have passed laws. Evaluating the impact of these laws on substance use and substance use disorders is of critical public health importance. Available data are two-year aggregated prevalence and standard error estimates of past-month marijuana, alcohol, and tobacco use from the National Survey on Drug Use and Health, 2006-2007 through 2016-17. We first extend difference-in-difference methods (DID) to assess the impact of these laws on a state-by-state basis using simulated datasets constructed from the published prevalence and standard error estimates. Next, we consider estimating the average effect of law passage from all states with recreational laws using linear mixed-effects models that account for historical trends in use over time borrowing information from states without laws. Lastly, we present a method for visually assessing the parallel-trends assumption when time of exposure varies across units, as is the case in this scenario.

email: cmm2212@cumc.columbia.edu

35b. Methods of Analysis when an Outcome Variable is a Prediction with Berkson Error

Pamela A. Shaw*, University of Pennsylvania
Paul Gustafson, University of British Columbia
Daniela Sotres-Alvarez, University of North Carolina, Chapel Hill
Victor Kipnis, National Cancer Institute, National Institutes of Health
Laurence Freedman, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center

For many epidemiologic settings, the outcome of interest can only be imprecisely measured and the gold-standard measure is too expensive to obtain on all subjects. To address measurement error, sometimes the analyst will adjust the outcome, say through a calibration or prediction equation, and use the resulting predicted value in the analysis in place of an observed value. Errors in continuous outcomes are often treated as ignorable; however, these predicted values have Berkson error, which can bias regression coefficients if analyses are not appropriately adjusted. Buonaccorsi (1991,1996) developed an adjustment procedure to eliminate this bias for non-differential Berkson error in the outcome variable for the linear model. We present a method to check this non-differentially condition when only error-prone observations of the outcome are available. Using simulation studies, the performance of the proposed method is studied and compared to the naive analysis that ignores error. The sensitivity of the method to mild departures from non-differentiality will also be studied. Methods are further illustrated with a real data example within the Hispanic Community Health Study.

email: shawp@pennmedicine.upenn.edu

35c. Confidence Intervals for the Youden Index and Its Optimal Cut-Off Point in the Presence of Covariates

Xinjie Hu*, Georgia State University
Gengsheng Qin, Georgia State University
Chenxue Li, Georgia State University
Jinyuan Chen, Lanzhou University

In medical diagnostic studies, the Youden index is a summary measure widely used in the evaluation of the diagnostic accuracy of a medical test. When covariates are not considered, the diagnostic accuracy of the test can be biased or misleading. By incorporating information from covariates using induced linear regression models, we propose generalized confidence intervals for the covariate-adjusted Youden index and its optimal cut-off point. Furthermore, under heteroscedastic regression models, we propose various confidence intervals for the covariate-adjusted Youden index. Extensive simulation studies are conducted to evaluate the finite sample performance of the proposed intervals, the bias corrected and accelerated (BCa) confidence intervals, and the bootstrap-based confidence intervals for the Youden index in the presence of covariates. To illustrate the application of our recommended methods, we apply the methods to a dataset on postprandial blood glucose measurements.

email: anna.xjh@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

35d. Critical Window Variable Selection for Pollution Mixtures

Joshua L. Warren*, Yale University

Understanding the impact of environmental exposure during different stages of pregnancy on the risk of adverse birth outcomes is vital for protection of the fetus. Statistical models to estimate critical windows of susceptibility have been developed and widely applied. Recently, critical window variable selection (CWVS) was developed and shown to outperform competing techniques in terms of correctly identifying critical windows and accurately estimating risk parameters. However, CWVS is limited in that it can accommodate a single exposure. We extend CWVS to the pollution mixtures setting (CWVSmix) to investigate critical windows of vulnerability with respect to simultaneous exposure to multiple pollutants. CWVSmix provides a framework for assessing the relative contribution of individual pollutants to a potentially harmful mixture and the ability to quantify the impact that each mixture profile has on risk across the exposure period, while still allowing for a direct critical window definition. We compare CWVSmix to competing methods and explore associations between simultaneous exposure to multiple ambient air pollutants and adverse pregnancy outcomes.

WITHDRAWN

email: joshua.warren@yale.edu

35e. Learning Individualized Treatment Rules for Multiple-Domain Latent Outcomes

Yuan Chen*, Columbia University
Donglin Zeng, University of North Carolina, Chapel Hill
Yuanjia Wang, Columbia University

For many mental disorders, latent mental status from multiple-domain psychological or clinical symptoms may perform as a better characterization of the underlying disorder status than a simple summary score of the symptoms, and they may also serve as more reliable and representative features to differentiate treatment responses. Therefore, in order to address the complexity and heterogeneity of treatment responses of mental disorders, we provide a new paradigm for learning optimal individualized treatment rules (ITRs) by modeling patients' latent mental status. We first learn the multi-domain latent states at baseline from the observed symptoms under a machine learning model, through which patients' heterogeneous symptoms are represented using an economical number of latent variables. We optimize a value function defined by the latent states after treatment by exploiting a transformation of the observed symptoms without modeling the relationship between the latent mental states before and after treatment. The optimal treatment rules are derived using a weighted large margin classifier. We derive the convergence rate of the proposed estimator under the latent models.

email: yc3281@cumc.columbia.edu

35f. Semi-Parametric Efficient Prediction of Binary Outcomes when Some Predictors are Incomplete via Post-Stratification

Yaqi Cao*, University of Pennsylvania
Sebastien Haneuse, Harvard T.H. Chan School of Public Health
Yingye Zheng, Fred Hutchinson Cancer Research Center
Jinbo Chen, University of Pennsylvania

When a logistic regression model is fit to two-phase data where some covariates are only available for a subgroup of subjects, a wide range of methods are available for the inference of association parameters. Substantially less research, however, has been devoted to ensuring optimal use of data when risk prediction is of interest. We extend existing maximum likelihood (ML) method to estimate the risk distribution and measures of predictive performance. We then propose a novel post-stratification semi-parametric ML analysis strategy in which subjects are cross-classified by the binary outcome and discretized predictions from a working model that utilizes fully-observed covariates. We subsequently propose a new two-phase sampling scheme for risk prediction based on the post-sampling strata as defined above. We show that the standard and proposed ML methods yield consistent estimates, and the proposed sampling strategy improves efficiency for estimating measures of predictive accuracy relative to a standard two-phase design. Finally, we develop a model for predicting hospital readmissions using data from the Pennsylvania Health Care Cost Containment Council.

email: Yaqi.Cao@pennmedicine.upenn.edu

35g. Optimal Sampling Plans for Functional Linear Regression Models

Hyungmin Rha*, Arizona State University
Ming-Hung Kao, Arizona State University
Rong Pan, Arizona State University

This research is concerned with the best sampling schedule on the predictor time axis to precisely recover the trajectory of a predictor function and to predict a scalar/functional response through functional linear regression models. Three optimal designs are considered, namely, the schedule that maximizes the precision of recovering predictor function, the schedule that is the best for predicting response (function), and the schedule that optimizes a user-defined mixture of the relative efficiencies of the two objectives. A search algorithm that can efficiently generate nearly optimal designs is proposed and it is demonstrated that the proposed approach outperforms the previously proposed methods.

email: hrha1@asu.edu

ABSTRACTS & POSTER PRESENTATIONS

35h. Optimal Experimental Design for Big Data: Applications in Brain Imaging

Eric W. Bridgeford*, Johns Hopkins University
Shangsi Wang, Johns Hopkins University
Zeyi Wang, Johns Hopkins University
Brian Caffo, Johns Hopkins University
Joshua Vogelstein, Johns Hopkins University

Reference datasets-benchmark datasets that serve to answer many disparate questions for different individuals-are becoming ubiquitous across many fields. How can one optimally design these reference datasets and the pipelines used to process them to yield derivative data that are simultaneously useful for different tasks? The key insight is that each measurement of the same item should be more similar to other measurements of that item, as compared to measurements of any other item. We formalize the notion of discriminability, and introduce both a non-parametric and parametric statistic to quantify the discriminability of potentially multivariate or non-Euclidean datasets. We show that optimizing decisions with respect to discriminability yields improved performance on subsequent inference tasks. We apply this strategy to 24 disparate neuroimaging datasets, each with up to hundreds of individuals that were imaged multiple times. We show that by optimizing pipelines with respect to discriminability, we improve performance on multiple subsequent inference tasks, even though discriminability does not consider the tasks whatsoever.

email: ericwb95@gmail.com

35i. New Statistical Learning for Evaluating Nested Dynamic Treatment Regimes with Test-and-Treat Observational Data

Ming Tang*, University of Michigan
Lu Wang, University of Michigan
Jeremy M.G. Taylor, University of Michigan

Dynamic treatment regimes (DTRs) include a sequence of treatment decision rules, in which treatment is adapted over time in response to the changes in an individual's disease progression. In practice, nested test-and-treat strategies are common to improve cost-effectiveness. For example, patients at risk of prostate cancer need costly and invasive biopsy to confirm the diagnosis and help determine the treatment. A decision about treatment happens after the biopsy, and is thus nested within the decision of whether to do the test. However, current existing statistical methods are not able to accommodate such a naturally embedded property of the treatment decision within the test decision. Therefore, we developed new statistical learning methods to evaluate DTR within such a nested multi-stage dynamic decision framework. Robust semi-parametric estimation is combined with a modified tree-based reinforcement learning method to deal with the counterfactual optimization. The simulation studies showed robust performance. We further applied our method to evaluate the necessity of prostate biopsy and identify the optimal test-and-treatment regimes for prostate cancer patients.

email: mingtang@umich.edu

35j. A Sequential Strategy for Determining Confidence in Individual Treatment Decisions in Personalized Medicine

Nina Orwitz*, New York University
Eva Petkova, New York University
Thaddeus Tarpey, New York University

New and evolving medical technologies are motivating researchers to develop treatment decision rules (TDRs) that incorporate complex, expensive data types such as genetic testing and imaging (e.g. EEGs). In clinical practice, we aim for these TDRs to be valuable for physicians and patients, such that we reduce unnecessary or costly testing while preserving the treatment decision. For a given TDR, the decision for a patient with certain baseline measures will lie some distance from the optimal decision boundary separating treatment classes. There is greater uncertainty, or less confidence, for patients with decisions near the boundary, while those far from the boundary are associated with more confidence. We propose a novel measure of confidence in individual decisions in which the TDR is sequentially updated with more data and reassessed until high confidence is achieved. Our approach estimates the probability of a specific treatment assignment for a patient using current measures and prior treatment decisions made with less data. We investigate the confidence measure through extensive simulation studies and provide recommendations for practical use.

email: nina.orwitz@nyulangone.org

35k. Hidden Analyses: A Systematic Framework of Data Analyses Prior to Statistical Modeling and Recommendations for More Transparent Reporting

Marianne Huebner*, Michigan State University
Werner Vach, University Hospital Basel, Switzerland
Saskia le Cessie, Leiden University Medical Center, Netherlands
Carsten Schmidt, University Medicine of Greifswald, Germany
Lara Lusa, University of Primorksa, Slovenia

In the data pipeline from the data collection process to the planned statistical analyses, initial data analyses (IDA) typically take place during and after the data collection and do not touch the research questions. A systematic process for IDA and clear reporting of the findings helps to understand the potential shortcomings of a dataset, such as missing values, subgroups with small sample sizes, issues in the collection process, and to evaluate the impact of these shortcomings on the results. The IDA process is divided into six steps: Metadata setup, Data cleaning, Data screening, Initial reporting, Updating the analysis plan, and Reporting IDA in publications. A clear reporting of findings is relevant when making data sets available and can provide valuable insights into the suitability of a dataset for research studies. In a review of papers from five highly ranked medical journals we found that reporting of IDA is sparse and statements on IDA are located throughout the papers, illustrating a lack of systematic reporting of IDA. In this contribution we discuss the IDA framework and present thoughts on improving the currently poor practice of reporting.

email: huebner@msu.edu

ABSTRACTS & POSTER PRESENTATIONS

35l. A Bayesian Adaptive Design for Early Phase Biomarker Discovery Study

Yi Yao*, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center
Liang Li, University of Texas MD Anderson Cancer Center

Stored biospecimens are scarce resources and are expensive to acquire. In the initial phase of biomarker discovery studies, there is often insufficient preceding information about the biomarker under investigation. Thus, it is desirable to terminate the study early when poor performance is evident in order to conserve samples for future studies. To this end, this article provides a first investigation of using Bayesian adaptive study design in biomarker discovery study. The proposed design allows multiple interim analysis in which the probability of the posterior samples of AUC greater than a prespecified threshold is calculated to support termination/continuation decisions. Our methodology can accommodate both normally and non-normally distributed data in the case and control groups. The performance of the proposed design was evaluated across various simulation scenarios as well as in a diagnostic biomarker study for early detection of chronic pancreatitis. Our study shows that the proposed adaptive design could substantially reduce sample size compared to a fixed sample design while still preserving desired statistical properties.

email: Yi.Yao@uth.tmc.edu

35m. Association Between Tooth Loss and Cancer Mortality: NHANES 1999-2015

Xiaobin Zhou*, Agnes Scott College
Kelli O'Connell, Memorial Sloan Kettering Cancer Center
Mengmeng Du, Memorial Sloan Kettering Cancer Center

Few studies have investigated the relationship between oral health and cancer mortality. We examined the association of tooth loss with cancer mortality in a nationally representative US population. We analyzed data from 30,432 participants aged 20 or older in the National Health and Nutrition Examination Survey enrolled between 1999-2012. We used Cox proportional hazards models to estimate multivariable-adjusted hazard ratios (HRs) and 95% CIs. We explored the potential mediation role of inflammation, comorbid conditions, and diet. During the follow-up (median: 8.8 years), 994 cancer-specific deaths were documented. After adjusting for potential confounding factors, compared with 0-4 missing teeth, participants with 5-8 missing teeth had a suggested increased risk (HR= 1.27, 95% CI=0.95, 1.70) and with 9+ missing teeth had an increased risk (HR= 1.41, 95% CI=1.09-1.83) of cancer-specific mortality (P-trend<0.01). No evidence that CRP levels, comorbid conditions, or diet mediated this association. Excessive tooth loss was associated with increased cancer mortality. These results provide evidence for the importance of maintaining oral health in reducing cancer mortality.

email: xzhou@agnesscott.edu

36. ADAPTIVE DESIGNS FOR CLINICAL TRIALS

Keyboard Design for Phase I Drug-Combination Trials

Haitao Pan*, St. Jude Children's Research Hospital
Ruitao Lin, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center

The keyboard design is a novel model-assisted dose-finding method to find the maximum tolerated dose (MTD) for single-agent trials. The keyboard design is easy to implement and has superior performance. In this talk, we extend the keyboard design to dual-agent dose-finding trials. The proposed keyboard combination trial design maintains the simplicity of the original single-agent keyboard design, and its dose escalation and de-escalation rules can be pre-tabulated before conducting the trial. We show that the keyboard combination design has desirable theoretical properties, including the optimality of its decision rules, coherence in dose transition, and convergence to the target dose. Extensive simulations are conducted to evaluate the performance of the proposed keyboard combination design using a novel, random two-dimensional dose--toxicity scenario generating algorithm. The simulation results confirm the desirable and competitive operating characteristics of the keyboard design. An R Shiny application is developed to facilitate implementing the keyboard combination design in practice.

email: haitao.pan@stjude.org

Interim Adaptive Decision-Making for Small n Sequential Multiple Assignment Randomized Trial

Yan-Cheng Chao*, University of Michigan
Thomas M. Braun, University of Michigan
Roy N. Tamura, University of South Florida
Kelley M. Kidwell, University of Michigan

A small n, sequential, multiple assignment, randomized trial (snSMART) is a small sample multi-stage design where subjects may be re-randomized to treatment based on intermediate endpoints. This design is motivated by ARAMIS (NCT02939573), an ongoing snSMART focusing on the evaluation of three drugs for isolated skin vasculitis. By formulating an interim decision rule for removing one of the treatments, we use a Bayesian model and the resulting posterior distributions to provide sufficient evidence that one treatment is inferior to the other treatments before enrolling more subjects. By doing so, we can remove the worst performing treatment at an interim analysis and prevent the subsequent subjects from receiving the removed treatment. Based on simulation results, we have evidence that the treatment response rates can still be unbiasedly and efficiently estimated in our new design, especially for the treatments with higher response rates. In addition, by adjusting the decision rule criteria for the posterior probabilities, we can control the probability of incorrectly removing an effective treatment.

email: ycchao@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Adaptive Enrichment Trial Design for Continuous Predictive Biomarkers with Possibly Non-Linear or Non-Monotone Effects

Yusha Liu*, Rice University
Lindsay Ann Renfro, University of Southern California

Biomarker-driven adaptive designs, which initially enroll an unselected population and then adapt to enroll only marker-positive patients based on interim analysis results, are increasingly popular. While many predictive biomarkers are naturally continuous, existing trial designs often require the driving biomarker of interest to be dichotomous in nature, ignoring the possibility that the continuous biomarker may have a truly non-linear or non-monotone prognostic relationship with outcome or predictive relationship with treatment effect. We propose an adaptive enrichment design for continuous biomarkers that takes uncertainty regarding the shape and strength of these two relationships into account, where following one or more interim analyses, decisions to continue enrollment in all patients, restrict enrollment to a subgroup, or terminate for efficacy or futility are considered. Using simulations and patient-level data from an actual clinical trial, we derive the operating characteristics of our design framework and evaluate its performance compared to a traditional adaptive enrichment approach that forces marker dichotomization.

email: yl95@rice.edu

Robust Blocked Response-Adaptive Randomization Designs

Thevaa Chandereng*, University of Wisconsin, Madison
Rick Chappell, University of Wisconsin, Madison

Response-adaptive randomization (RAR) has been proposed for ethical reasons, where the randomization ratio is tilted successively to favor the better performing treatment. However, the substantial disagreement regarding bias due to time-trends in adaptive randomization is not fully recognized. The type-I error is inflated in the traditional Bayesian RAR approaches when a time-trend is present. In our approach, patients are assigned in blocks and the randomization ratio is recomputed for blocks rather than traditional adaptive randomization where it is done per patient. We further investigate the design with a range of scenarios for both frequentist and Bayesian designs. We compare our method with equal randomization and with different numbers of blocks including the traditional RAR design where randomization ratio is altered patient by patient basis. Small blocks should be avoided due to the possibility of not acquiring any information from the μi. On the other hand, RAR with large blocks has a good balance between efficiency and treating more subjects to the better-performing treatment, while retaining blocked RAR's unique unbiasedness.

email: chandereng@wisc.edu

Streamlined Hyperparameter Tuning in Mobile Health

Marianne Menictas*, Harvard University

Mobile health technologies are increasingly being employed to deliver interventions to users in their natural environments. With the advent of sophisticated sensing devices and phone-based EMA, it is becoming possible to deliver interventions at moments when they can most readily influence a person's behavior. Thus, our goal is to learn the optimal time and intervention for a given user and context. A significant challenge to learning is that there are often only a few opportunities per day to provide treatment. Additionally, when there is limited time to engage users, a slow learning rate can raise the risk that users will abandon the intervention. To prevent disengagement and accelerate learning, information may be pooled across users and time in a dynamic manner, combining a bandit algorithm with a Bayesian random effects model for the reward function. As information accumulates, however, tuning user and time specific hyperparameters becomes computationally intractable. We focus on solving this computational bottleneck.

email: mmenictas@gmail.com

A Two-Stage Sequential Design for Selecting the Best Treatments

Mingyue Wang*, Syracuse University
Pinyuen Chen, Syracuse University

Thall, Simon, and Ellenberg (1988) proposed a fixed-sample-size two-stage selection and testing design for selecting the best of several treatments with comparisons to a control based on binary outcomes. Having as a reference Thall et al.'s design, we propose a sequential design for the generalized goal of selecting the t best of k ($1 < t < k$) experimental treatments. The experimental treatments in each stage are simultaneously compared against a control. The proposed design allows early decision making in both stages, subject to constraints of specified type I error and overall power. Numerical results of optimal sample sizes and design parameters are presented to demonstrate the efficient and ethical advantages of the proposed design.

email: mwang55@syr.edu

ABSTRACTS & POSTER PRESENTATIONS

Adaptive Monitoring: Optimal Burn-in to Control False Discoveries Allowing Unlimited Monitoring

Jonathan J. Chipman*, Huntsman Cancer Institute, University of Utah
Jeffrey D. Blume, Vanderbilt University
Robert A. Greevy, Jr., Vanderbilt University

Adaptive monitoring allows a study to stop collecting data once drawing a clear scientific conclusion. To control false discoveries, frequentist inference uses alpha-spending functions (Pocock; OBrien and Fleming; Lan DeMets) and provide classical p-values. The Second-Generation p-Value focuses on an interval null hypothesis formed on scientific context. It is the overlap between an estimated interval (Confidence, Credible, or Support) and an interval hypothesis. No overlap (SGPV=0) is evidence against the interval null; complete overlap (SGPV=1) is evidence for the interval null. With unlimited data, the interval null hypothesis will either be supported or rejected. In this talk we show how to adaptively monitoring using SGPVs, until ruling out clinically trivial or clinically desirable effects. For this design, we present a wait-time for applying fully sequential stopping rules that assures a nominal False Discovery Rate. This work establishes an easy to use adaptive monitoring scheme, which does not require simulations to show a bounded error rate, and can open up adaptive monitoring to broad range of scientific researchers.

email: jonathan.chipman@hci.utah.edu

37. BAYESIAN SEMIPARAMETRIC AND NONPARAMETRIC METHODS

Heterogeneity Pursuit for Spatial Point Process with Applications: A Bayesian Semiparametric Recourse

Jieying Jiao*, University of Connecticut
Guanyu Hu, University of Connecticut
Jun Yan, University of Connecticut

Spatial point pattern data are routinely encountered in various fields such as seismology, ecology, environmental science, and epidemiology. Building a flexible regression model for spatial point process is an important task in order to reveal data's spatial pattern and relationships with various factors. We propose a Bayesian semiparametric regression model for spatial Poisson point process data based on powered Chinese resturant process. Further, we allow variable selection through the spike-slab prior. An efficient Markov chain Monte Carlo (MCMC) algorithm is developed for the proposed methods, followed with an extensive simulation studies to evaluate the empirical performance. The proposed methods are further applied to the analysis of the Forest of Barro Colorado Island (BCI) data.

email: jieying.jiao@uconn.edu

A Bayesian Finite Mixture Model-Based Clustering Method with Variable Selection for Identifying Disease Phenotypes

Shu Wang*, University of Florida

Identification of distinct clinical phenotypes of a disease may allow clinicians to prescribe more precise therapy and improve patient care. The nature of the heterogeneous syndrome of a disease sometimes can be determined by studying clinical data in large databases through clustering. Challenges exist when using existing unsupervised clustering methods to uncover hidden patient subgroups. First, most clustering methods were designed for either continuous or categorical variables. Second, not all methods are able to conduct variable selection. Third, currently there is no clustering method that handles censored variables. To address these challenges, we propose a Bayesian finite mixture model (FMM) to simultaneously conduct variable selection, account for censored biomarkers, and achieve clustering for variables of mixed types. Simulations under various scenarios were conducted to compare the performance of this method with existing methods. We applied proposed Bayesian FMM to identify sepsis phenotypes from EHR and investigated the association between identified phenotypes and various clinical adverse outcomes.

email: swang0221@ufl.edu

A Bayesian Nonparametric Model for Zero-Inflated Outcomes: Prediction, Clustering, and Causal Estimation

Arman Oganisian*, University of Pennsylvania
Nandita Mitra, University of Pennsylvania
Jason A. Roy, Rutgers University

Common tasks such as clustering, prediction, and causal estimation are often complicated in the presence of skewed, zero-inflated, and multi-modal outcomes. Motivated by medical cost outcomes in particular, we present a Bayesian nonparametric solution using Dirichlet Processes. Predictions from this model not only capture skewness and structural zeros, but also induce a clustering of subjects into groups with similar joint data distributions. Furthermore, the model can be used to compute robust causal estimates via a nonparametric Standardization procedure. Full posterior inference allows for probabilistically valid uncertainty estimates as well as posterior predictive checks of the Positivity assumption. We apply our method to analyze inpatient medical costs among endometrial cancer patients in the SEER-Medicare database.

email: aoganisi@upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Longitudinal Structural Topic Models for Estimating Latent Health Trajectories using Administrative Claims Data

Mengbing Li*, University of Michigan
Zhenke Wu, University of Michigan

Administrative claims data present a unique opportunity for longitudinal assessment of patients' health. We adapt topic models, widely used for text mining, to analyzing such data. In this work, we estimate an unobserved patient-specific trajectory that characterizes her progression of multiple latent biological aberrations, each of which is an unobserved topic that yields distinct content distributions of the diagnosis codes. We propose a novel extension of the structural topic model (Roberts et al. 2016) that builds in important features of claims data: repeated multivariate diagnosis codes, and time-varying covariates for topic prevalences and content distribution. Our model specifies the topic prevalences by logistic mixed models and the content distributions by regularized logistic models. We derive a scalable variational EM inference algorithm. We apply the model to data from 15k cancer-associated thrombosis patients extracted from OptumInsight claims database. By aggregating monthly diagnosis codes (ICD-9) over multiple months as correlated documents from a patient, we quantify the latent disease progression and the effects of baseline and time-varying covariates.

email: mengbing@umich.edu

Novel Semiparametric Bayesian Methods for the Competing Risks Data with Length-Biased Sampling

Tong Wang*, Texas A&M University

Analysis of registry data, like the SEER data, using a competing risk model is challenging due to two main reasons, 1) the sheer volume of the data and 2) length-bias sampling (also known as the selection bias). The sheer size of the SEER data is computationally challenging. Ignoring the selection bias issue may lead to a distorted inference about the underlying population. To handle these issues, we develop a flexible semiparametric Bayesian method for analyzing such data. The novelty of our approach lies in the robust modelling of data and innovative variational method of handling computational challenges.

email: tong@stat.tamu.edu

A Bayesian Nonparametric Approach for Estimating Causal Effects for Longitudinal Data

Kumaresh Dhara*, University of Florida
Michael J. Daniels, University of Florida

We propose a Bayesian nonparametric approach to estimate causal effects in the longitudinal treatment setting. We use Enriched Dirichlet process mixtures to model the longitudinal treatments, time-varying confounders, and the response and use G-computation for compute the causal effects. This flexible modeling avoids typical parametric G-computation approaches. In addition, most of the existing literature requires separate sets of models to estimate different causal effects (e.g., mean causal effect vs quantile causal effects). Contrary to such procedures, in this Bayesian nonparametric approach, we estimate the distribution of the potential outcomes, allowing us to calculate any causal effect. The proposed method can address ignorable missing covariates automatically through data augmentation; this method does not depend on a separate imputation model. We provide simulation studies and applications to real-world data to demonstrate the efficacy of the proposed method.

email: k.dhara@ufl.edu

---

38. STATISTICAL METHODS IN CANCER RESEARCH

Identifying Gene-Environment Interactions Using Integrative Multidimensional Omics Data for Cancer Outcomes

Yaqing Xu*, Yale University
Mengyun Wu, Shanghai University of Finance and Economics
Shuangge Ma, Yale University

Analyzing gene-environment (G-E) interactions can improve the understanding of disease mechanisms. Recent studies have found combining multidimensional omics measurements can provide more comprehensive findings compared to using single-type data. Few statistical approaches incorporate multidimensional omics data to identify G-E interactions. We proposed a three-step procedure for selecting hierarchical G-E interactions with main effects using integrative omics data. Our proposed method properly addresses the relationship between gene expressions and regulators by multivariate regression and biclustering technique, and establishes integrative omics covariates to facilitate interaction analysis. This process is biologically sensible and straightforward to apply. The proposed joint model respects the hierarchical structure of main effects and interactions, and is able to accommodate different types of response including survival outcomes. Extensive simulation shows our proposed approach outperforms multiple alternatives in marker identification. In the analysis of The Cancer Genome Atlas data, interesting findings with superior stability and prediction are made.

email: yaqing.xu@yale.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Modeling of Metagenomic Sequencing Data for Discovering Microbial Biomarkers in Colorectal Cancer Detection

Shuang Jiang*, Southern Methodist University
Qiwei Li, University of Texas, Dallas
Andrew Y. Koh, University of Texas Southwestern Medical Center
Guanghua Xiao, University of Texas Southwestern Medical Center
Xiaowei Zhan, University of Texas Southwestern Medical Center

Colorectal cancer (CRC) is a major cause of mortality globally. Optical colonoscopy, though currently the most effective CRC screening test, is costly and invasive. Recent CRC studies demonstrate significant association between tumorigenesis and abnormalities in microbial community, shedding light on utilizing microbial features as non-invasive CRC biomarkers. Here, we propose a Bayesian hierarchical model to identify a set of differentially abundant taxa that potentially serve as microbial biomarkers. The first level is a count generative model that links the raw counts to the normalized abundances and a sample-specific normalized factor. We use Dirichlet process as a flexible nonparametric mixing distribution to estimate those normalized factors. The second level is a Gaussian mixture model with a feature selection scheme to identify those discriminating taxa between CRC and non-CRC groups. A comprehensive comparative analysis based on synthetic data is conducted. A case study shows that a diagnostic model based onï»¿ the taxonomic features identified by our model significantly improves the prediction performance in an independent cohort.

email: shuangj@smu.edu

Propensity Score Methods in the Presence of Missing Covariates

Kay See Tan*, Memorial Sloan Kettering Cancer Center

Propensity score methods such as pairwise matching and inverse-probability treatment weighting (IPTW) have been proposed as a means to recover covariate balance between treatment groups of interest in observational studies. When estimating propensity scores, missing covariate data is a major issue that is commonly overlooked, leading to suboptimal matching or improper IPTW. Multiple imputation (MI) is a natural procedure to handle missing data in this context. However, open questions still exist regarding the implementation of MI for propensity score analysis. A series of simulation study will (1) investigate two opposing proposed methods to combine the MI and propensity score analysis steps, and (2) address variance estimation of the IPTW estimators after MI. This study will focus on estimands that target average treatment effect for the treated (ATT), such as the matching-weights approach, a propensity-score weighting analogue to pairwise 1:1 matching (Li and Greene, 2011). These concepts are then illustrated in a retrospective study that examines the impact of single- vs multiple-unit blood transfusions on survival after lung cancer surgery.

email: tank@mskcc.org

Pathway-Structured Predictive Modeling for Multi-Level Drug Response in Multiple Myeloma

Xinyan Zhang*, Georgia Southern University
Bingzong Li, Soochow University
Wenzhuo Zhuang, Soochow University
Nengjun Yi, University of Alabama at Birmingham

Multiple myeloma (MM) is composed of distinct subtypes with various response rates to certain treatments. Drug responses in MM are usually recorded as a multi-level ordinal outcome. One of the goals of drug response studies is to predict drug response for patients based on their clinical and molecular features. However, gene-based models may provide limited predictive accuracy. In that case, methods for predicting multi-level ordinal drug responses by incorporating biological pathways are desired but have not been developed yet. We propose a pathway-structured method for predicting multi-level ordinal responses using a two-stage approach. Our two-stage approach first obtains the pathway score for each pathway by fitting all predictors within each pathway using the hierarchical ordinal logistic approach, and then combines the pathway scores as new predictors to build a predictive model. We applied the proposed method to two publicly available datasets. Our results show that our approach not only significantly improved the predictive performance compared with the corresponding gene-based model but also allowed us to identify biologically relevant pathway.

email: xzhang@georgiasouthern.edu

Integrative Network Based Analysis of Metabolomic and Transcriptomic Data for Understanding Biological Mechanism of Lung Cancer

Christopher M. Wilson*, Moffitt Cancer Center
Brooke L. Fridley, Moffitt Cancer Center
Doug W. Cress, Moffitt Cancer Center
Farnoosh Abbas Aghabazadeh, Princess Margaret Cancer Centre

Lung cancer is the leading cause of cancer deaths globally. Pre-clinical cancer studies are essential for gauging the basic biology of lung cancer to develop targeted therapies and often assay multiple high throughput data sources for a limited sample size. The integration of these data sources can aid in gaining insight into the global effects of treatment. To understand the biology of lung cancer, we have developed a new integrative framework that determines modules that are differentially expressed in different phenotypes. The approach leverages sparse Gaussian graphical models, which compute conditional correlation thus eliminating indirect correlation. We propose an integrative step-wise analysis that is conducted by determining which transcriptomics and metabolomics modules are differentially expressed and then determining which of the transcriptomic and metabolomic modules are related. We compare our approach to weighted gene co-expression network analysis by comparing the properties of the resulting network and using biologically related features.

email: christopher.wilson@moffitt.org

ABSTRACTS & POSTER PRESENTATIONS

A General Framework for Multi-Gene, Multi-Cancer Mendelian Risk Prediction Models

Jane W. Liang*, Harvard T.H. Chan School of Public Health
Gregory Idos, University of Southern California Norris Comprehensive Cancer Center
Christine Hong, University of Southern California Norris Comprehensive Cancer Center
Stephen B. Gruber, University of Southern California Norris Comprehensive Cancer Center
Giovanni Parmigiani, Dana-Farber Cancer Institute
Danielle Braun, Dana-Farber Cancer Institute

Identifying individuals who are at increased risk for having a cancer susceptibility mutation is important for risk reduction and clinical management. Mendelian models use principles of Mendelian genetics, Bayesian probability theory, and mutation-specific knowledge to estimate the probability of being a carrier. Though widely used for research and clinical decision-making, existing models are generally limited to a specific mutation and syndrome. However, there is growing evidence that syndromes once thought to be distinct are determined by mutations that increase the risk of multiple cancers. Recent advancements in sequencing technology have lead to the availability of data from multi-gene panel testing, supporting the need and possibility of multi-gene, multi-syndrome risk prediction models. We present a flexible, computationally efficient model that incorporates an arbitrary number of genes and cancers. This framework addresses the limitations of single-syndrome models in its ability to provide posterior carrier probabilities for multi-syndrome associations for a large number of genes. With simulations and panel testing data, we validate and demonstrate its usage.

email: jwliang@g.harvard.edu

The Impact of Design Misspecification in Oncology Trials with Survival Endpoint

Tyler Zemla*, Mayo Clinic
Jennifer Le-Rademacher, Mayo Clinic

The primary endpoint of a clinical trial should be chosen to reflect the effectiveness of a new treatment. When the design assumptions deviate from the true survival patterns, the trial may be underpowered or overpowered or the statistical test may not be appropriate. Misspecification can lead to regulatory approval of ineffective treatments or premature termination of effective treatments which can have detrimental impact on patients. We conducted a simulation to evaluate the impact of deviations from design assumptions on the trial conclusions on each of three statistical designs (the log-rank test, the restricted mean survival time test, and the difference in survival probability at a fixed time point). We explored the impact of deviations from: a) the assumed baseline survival distribution, b) the expected treatment effect, c) the proportional hazards assumption on the power. The results of our simulation study help us to provide recommendations for selection of an appropriate endpoint or model when designing oncology trials so that it can correctly capture the effect of treatment on cancer patients.

email: zemla.tyler@mayo.edu

---

39. NETWORK ANALYSIS: CONNECTING THE DOTS

Bayesian Assessment of Homogeneity and Consistency for Network Meta-Analysis

Cheng Zhang*, University of Connecticut
Hao Li, University of Connecticut
Ming-Hui Chen, University of Connecticut
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Arvind K. Shah, Merck & Co., Inc.
Jianxin Lin, Merck & Co., Inc.

One of the long-standing methodological issues in network meta analysis (NMA) is that of assessing consistency in treatment comparisons. However, the consistency assumption may not hold in the presence of heterogeneity. In this paper, we construct general linear hypotheses to investigate consistency under a general fixed effects model. A general algorithm is developed to compute the contrast matrix under homogeneity and consistency assumptions. Under the normal fixed effects model, we show the equivalence of the likelihood ratio test under the proposed linear hypotheses and Bucher's method for testing inconsistency based on comparison of the weighted averages of direct and indirect treatment effects. A novel Plausibility Index (PI) is developed to assess homogeneity and consistency simultaneously. Theoretical properties of the proposed Bayesian methodology are examined in details. We apply the proposed methodology to analyze the network meta data from 29 randomized clinical trials with 11 treatment arms on safety and efficacy evaluation of cholesterol lowering drugs.

email: cheng.2.zhang@uconn.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Community Detection for Multiple Networks

Luoying Yang*, University of Rochester Medical Center
Zhengwu Zhang, University of Rochester Medical Center

Detecting community structure of networks is an interesting topic with a long history. However, most of existing methods are developed for analyzing single network lacking the ability to infer community structures at a group level. Many applications require consideration of multiple networks simultaneously. In this paper, we aim to build a flexible model that can automatically identify the grouping of subjects with similar community structure and estimate the common community structure for multiple networks within one group. We propose a Bayesian framework based on mixture of stochastic block models with Dirichlet-Multinomial prior on the cluster assignments at both the network and node levels. An efficient MCMC algorithm is developed to simultaneously update the clustering configurations of networks and nodes. Simulated studies show the advantages of our model over existing methods, including more accurate inference on the networks with vague block structures. Applications to human brain structural connectomes show that our model can detect differences in brain community structures between subjects using scan-rescan data and between groups with different cognitive measure.

email: luoying_yang@urmc.rochester.edu

Semi-Parametric Bayes Regression with Network Valued Covariates

Xin Ma*, Emory University
Suprateek Kundu, Emory University
Jennifer Stevens, Emory University

Brain network is increasingly recognized as neuroimaging biomarker in mental health and psychiatric studies. Our focus is posttraumatic stress disorder (PTSD), where the brain network interacts with environmental exposures in complex ways to drive the disease progression. Existing linear models characterizing the relation between the clinical phenotype and the entire edge set in the brain network may be overly simplistic, with the inflated number of parameters leading to computational burden and inaccurate estimation. In one of the first such efforts, we develop a novel two stage Bayesian framework to find a node-specific lower dimensional representation of the network using latent scale model, and then use a flexible Gaussian process regression to predict the outcome based on the latent scales and other supplementary covariates. The proposed method relaxes linearity assumption, addresses the curse of dimensionality and is scalable to high dimension while maintaining interpretability at the node level. Extensive simulations and an application to our motivating PTSD data show a distinct advantage of our approach over competing methods in terms of prediction and coverage.

email: xin.ma@emory.edu

Scalable Network Estimation with L0 Penalty

Junghi Kim*, U.S. Food and Drug Administration
Hongtu Zhu, University of North Carolina, Chapel Hill
Xiao Wang, Purdue University
Kim-Anh Do, University of Texas MD Anderson Cancer Center

WITHDRAWN

With the advent of high-throughput sequencing, an efficient computing strategy is required to deal with large genomic data sets. The challenge of estimating a large precision matrix has garnered substantial research attention for its direct application to discriminant analyses and gra[illegible]hods either use a lasso-type penalty that may lead to biased estimators or are computationally intensive, which prevents their application to very large graphs.We propose using an L0 penalty to estimate an ultra-large precision matrix (scalnetL0). We apply scalnetL0 to RNA-seq data from breast cancer patients represented in The Cancer Genome Atlas and find improved accuracy of classifications for survival times. The estimated precision matrix provides information about a large-scale co-expression network in breast cancer. Simulation studies demonstrate that scalnetL0 provides more accurate and efficient estimators, yielding and shorter CPU time and less Frobenius loss on sparse learning for large-scale precision matrix estimation.

email: Junghi.Kim@fda.hhs.gov

Disease Prediction by Integrating Marginally Weak Signals and Local Predictive Gene/Brain Networks

Yanming Li*, University of Michigan

Accurate prediction of lung cancer subtypes or Alzheimer's disease status is critical for early detection and prevention of the diseases. Conventional prediction approaches using ultrahigh-dimensional genomic profiles or brain-wide imaging scans ignore marginally weak signals. Even though marginally weak signals by themselves are not predictive, they could exert strong prediction effects when considered in connection with the marginally strong signals. We propose a classification method which significantly improves the disease prediction accuracy by detecting and integrating the local predictive gene/brain networks. A local predictive gene/brain network contains not only marginally strong signals, but also the marginally weak signals in connection with the strong ones. The detected local predictive networks provide biological insights on how the gene/brain pathways attribute to lung cancer/AD development and progression. We applied the proposed method to the Boston Lung Cancer Study Cohort and the Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets for lung cancer subtype and Alzheimer's disease status prediction.

email: liyanmin@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Scalar-on-Network Regression Via Gradient Boosting

Emily Morris*, University of Michigan
Jian Kang, University of Michigan

Neuroimaging studies have a growing interest in learning the association between the individual brain connectivity networks and their clinical characteristics. It is also of great interest to identify the sub brain networks as biomarkers to predict the clinical symptoms such as disease status, potentially providing insight on neuropathology. This motivates the need for developing a new type of regression model where the response variable is scalar, and predictors are networks that are typically represented as adjacent matrices or weighted adjacent matrices, to which we refer as scalar-on-network regression. In this work, we develop a new gradient boosting method for model fitting with sub-network markers selection. Our approach, as opposed to group lasso or other existing regularization methods, is essentially a gradient descent algorithm incorporating the network topology and thus can automatically detect the sub-network markers. We demonstrate the superiority of our methods via simulation studies and analysis of the rest-state fMRI data in a cognitive developmental cohort study.

email: emorrisl@umich.edu

40. POLICIES AND POLITICS: STATISTICAL ANALYSES OF HEALTH OUTCOMES IN THE REAL WORLD

The Challenges of Electronic Health Record Use to Estimate Individualized Type 2 Diabetes Treatment Strategies

Erica EM Moodie*, McGill University
Gabrielle Simoneau, McGill University

Sequences of treatments that adapt to the patient''s changing condition over time are often needed for the management of chronic diseases. An adaptive treatment strategy (ATS) consists of individualized treatment rules to be applied through the course of a disease that input the patient's characteristics at the time of decision-making and output a recommended treatment. In this talk, I present an application of dynamic weighted survival modeling, a method to estimate an ATS for censored outcomes, to answer an important clinical question about the treatment of type 2 diabetes using data from the Clinical Practice Research Datalink, a large primary care database and show the challenges encountered along the way.

email: erica.moodie@mcgill.ca

Incorporating Statistical Methods to Address Spatial Confounding in Large EHR Data Studies

Jennifer Bobb*, Kaiser Permanente Washington
Andrea Cook, Kaiser Permanente Washington

Many studies using electronic health record (EHR) data lack detailed information on important spatially-varying confounders, such as socioeconomic status. To address this unmeasured spatial confounding, one approach is to fit a spatial model, such as a Gaussian process model or adjusting for spline terms of spatial location. However, such approaches are frequently not applied in longitudinal EHR studies, due to several key challenges: computational issues in fitting spatial models with Big Data, accounting for possibly complex, time-varying spatial correlation structures, and the potential for fitting a spatial model to actually increase (rather than decrease) bias in certain settings. We conduct extensive simulations to investigate alternate approaches to address spatial confounding, and make recommendations for computationally efficient approaches to apply to big data EHR studies. This work is motivated by a study that links data on the built environment with longitudinal EHR data from a large integrated health care system (including 4.2 million weight measures on ~290,000 adult participants) to investigate whether moving to a different environment affects health.

e-mail: jenniferfederbobb@gmail.com

A Spatial Causal Analysis of Wildland Fire-Contributed PM2.5 Using Numerical Model Output

Alexandra E. Larsen*, Duke University School of Medicine
Shu Yang, North Carolina State University
Brian J. Reich, North Carolina State University
Ana Rappold, U.S. Environmental Protection Agency

Wildfire smoke contains hazardous levels of fine particulate matter (PM2.5), which adversely effects health. Estimating fire attributable PM2.5 is key to quantifying the impact on air quality and subsequent health burdens. This challenging since only total PM2.5 is measured at monitoring stations, and both fire-attributable PM2.5 and PM2.5 from all other sources are correlated in space and time. We propose a framework for estimating fire-contributed PM2.5 using a novel causal inference framework and bias-adjusted chemical model representations of PM2.5 under counterfactual scenarios. The chemical model representation of PM2.5 is from Community Multi-Scale Air Quality Modeling System (CMAQ) run with and without fire emissions across the contiguous U.S. for the 2008-2012 wildfire seasons. The CMAQ output is calibrated with observations from monitoring sites for the same spatial domain and time period. We use a Bayesian model that accounts for spatial variation to estimate the effect of wildland fires on PM2.5 and state assumptions under which the estimate has a valid causal interpretation. We also estimate the health burden associated with PM2.5 attributable to wildfire smoke.

e-mail: alexandra.larsen@duke.edu

ABSTRACTS & POSTER PRESENTATIONS

Propensity Score Matching with Time-Varying Covariates: An Application in the Prevention of Recurrent Preterm Birth

Erinn M. Hade*, The Ohio State University
Giovanni Nattino, The Ohio State University
Heather A. Frey, The Ohio State University
Bo Lu, The Ohio State University

In observational studies where a survival outcome is of interest, treatment initiation may be time-dependent, which is likely to be affected by both time-invariant and time-varying covariates. In certain situations, all subjects may be exposed to the treatment sooner or later. In this scenario, the causal effect of interest is the delay in treatment. We propose a propensity score matching strategy to estimate the treatment delay effect. The goal is to balance the covariate distribution between on-time treatment and delayed treatment groups at each time point under risk set matching. We apply this method to data from an EHR based study for the delayed treatment effect of progesterone therapy for patients with recurrent preterm birth.

e-mail: hade.2@osu.edu

A Bayesian Spatio-Temporal Abundance Model for Surveillance of the Opioid Epidemic

David M. Kline*, The Ohio State University
Lance A. Waller, Emory University
Staci A. Hepler, Wake Forest University

The opioid epidemic continues to be a national public health crisis. While the scale of the problem is undeniable, estimates of the local prevalence of opioid misuse are lacking, despite their importance to policymaking and resource allocation. This is due in part to the challenge of directly measuring opioid misuse in the population. Our approach utilizes existing county level surveillance data that quantifies outcomes related to opioid misuse and thus provides indirect information on misuse at the county level. Using a spatio-temporal abundance model framework, we integrate county level rates of opioid overdose deaths and treatment admissions with state level survey data on rates of opioid misuse to estimate county level rates of misuse. We investigate the performance of the proposed model via simulation and apply it to data from the state of Ohio.

email: kline.273@osu.edu

Health Co-Benefits of the Implementation of Global Climate Mitigation Commitments

Gavin Shaddick*, University of Exeter

The Paris Agreement for greenhouse gas emissions (GHG) mitigation, adaptation, and finance aims to reduce the risks and impacts of climate change by restricting the increase in global average temperature to below °C above pre-industrial levels and it is expected that this will lead to a reduction in air pollution. Here, we quantify the potential health co-benefits of climate change mitigation strategies through reductions in ambient air pollution (PM2.5). This process comprises of four stages: (i) estimating GHG and air pollutant emissions; (ii) estimating global concentrations of PM2.5 based on emissions; (iii) producing country-level distributions of population exposures; (iv) performing burden of disease calculations. This process is performed for a range different climate and socioeconomic scenarios to allow the potential co-benefits to be compared. Significant reductions in the expected number of deaths from causes associated with exposure to PM2.5 were observed in 2050 under a scenario with a long-term temperature stabilisation target when compared to a reference scenario, equating to ca. 4% of the total deaths associated with air pollution.

email: g.shaddick@exeter.ac.uk

---

41. STATISTICAL CONSIDERATIONS FOR OPTIMAL TREATMENT

Optimal Treatment Regime Estimation using Pseudo Observation with Censored Data

Taehwa Choi*, Korea University
Sangbum Choi, Korea University

Optimal treatment regime (OTR), recommending personalized medicine based on patients' status, is important to maximize treatment effects. There are several studies related with OTR for complete dataset. However, when observed data is censored, estimation procedure becomes unstable if we use methods based on complete data. In this study, we propose general estimation framework of OTR in survival and competing risks data for single stage and magnify to multiple stage so called dynamic treatment regime. To deal with censored data, we use pseudo observation based on jack-knife method for three well-known measures in survival analysis: t-year survival time, restricted mean survival time and t-year cumulative incidence function. By replacing survival time with pseudo observation, censored data can be dealt with complete dataset. Also, we provide simulation study to verify proposed method is effective way to estimate OTR of censored data. Finally, we apply our method to AIDS Clinical Trial Group (ACTG) study 175 data.

email: taehwa_choi@korea.ac.kr

ABSTRACTS & POSTER PRESENTATIONS

Boosting Algorithms for Estimating Optimal Individualized Treatment Rules

Duzhe Wang*, University of Wisconsin, Madison
Haoda Fu, Eli Lilly and Company
Po-Ling Loh, University of Wisconsin, Madison

We present nonparametric algorithms for estimating optimal individualized treatment rules. The proposed algorithms are based on the XGBoost algorithm, which is known as one of the most powerful algorithms in the machine learning literature. Our main idea is to model the conditional mean of clinical outcome or the decision rule via additive regression trees, and use the boosting technique to estimate each single tree iteratively. Our approaches overcome the challenge of correct model specification which is required in current parametric methods. The major contribution of our proposed algorithms is providing the efficient and accurate estimation of the highly nonlinear and complex optimal individualized treatment rules which arise in practice. Finally, we illustrate the superior performance of our algorithms by extensive simulation studies and conclude with an application to the real data from a diabetes Phase III trial.

email: dwang282@wisc.edu

Capturing Heterogeneity in Repeated Measures Data by Fusion Penalty

Lili Liu*, Shandong University and Washington University in St. Louis
Lei Liu, Washington University in St. Louis

In this paper we are interested in capturing heterogeneity in clustered or longitudinal data. Traditionally such heterogeneity is modeled by either fixed effects or random effects. In fixed effects models, the number of degree of freedom for the heterogeneity equals the number of clusters/subjects minus 1, which could result in less efficiency. In random effects models, the heterogeneity across different clusters/subjects is described by e.g., a random intercept with 1 parameter (for the variance of random intercept), which could lead to oversimplification and biases (shrinkage estimates). Our "fusion effects" model stands in between these two approaches: we assume that there are un-known number of different levels of heterogeneity, and use the fusion penalty approach for estimation and inference. We evaluate and compare the performance of our method to the fixed and random effects models by simulation studies. We apply our method to the Ocular Hypertension Treatment Study (OHTS) to capture the heterogeneity in the progression rate toward primary open-angle glaucoma of left and right eyes of different subjects.

email: lulinsdu@163.com

Optimal Individualized Decision Rules Using Instrumental Variable Methods

Hongxiang Qiu*, University of Washington
Marco Carone, University of Washington
Ekaterina Sadikova, Harvard Medical School
Maria Petukhova, Harvard Medical School
Ronald C. Kessler, Harvard Medical School
Alex Luedtke, University of Washington

There is an extensive literature on the estimation and evaluation of optimal individualized treatment rules in settings where all confounders of the effect of treatment on outcome are observed. We study the development of individualized decision rules in settings where some confounders may not have been measured but a valid binary instrument is available for a binary treatment. We first consider individualized rules that recommend treatment based on measured covariates. These rules will be most interesting in settings where it is feasible to intervene on treatment. We then consider a setting where intervening to encourage treatment is feasible but intervening on treatment is not. In both settings, we also handle the case that the treatment is a limited resource so that optimal interventions focus the resources on those individuals who will benefit most from treatment. We evaluate an optimal individualized rule by its average causal effect relative to a prespecified reference rule. We develop methods to estimate optimal individualized rules and construct asymptotically efficient plug-in estimators of the corresponding average causal effect relative to the reference rule.

email: qiuhx@uw.edu

Sample Size and Timepoint Tradeoffs for Comparing Dynamic Treatment Regimens in a Longitudinal SMART

Nicholas J. Seewald*, University of Michigan
Daniel Almirall, University of Michigan

Clinical practice often involves delivering a sequence of treatments which adapts to a patient's changing needs. A dynamic treatment regimen (DTR) is a sequence of pre-specified decision rules which, based on a patient's ongoing data, recommend interventions at multiple stages of treatment. The sequential, multiple-assignment randomized trial (SMART) is a tool which can be used in the development of a high-quality DTR. Often, SMARTs involve longitudinal outcomes collected over the course of the trial. An important consideration in the design of a longitudinal-outcome SMART, as with any trial, is both the sample size and number of measurement occasions. We extend previous work which developed easy-to-use sample size formulae for common SMART designs with three timepoints in which the primary aim is to compare, at end-of-study, two embedded DTRs which recommend different first-stage treatments. We discuss practical and statistical considerations in choosing between adding individuals or measurement occasions, while respecting the unique features of a SMART, including modeling constraints and over/under-representation of sequences of treatment among participants.

email: nseewald@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

42. CAUSAL INFERENCE WITH GENETIC DATA

Estimating Causal Relationship for Complex Traits with Weak and Heterogeneous Genetic Effects

Jingshu Wang*, The University of Chicago
Qingyuan Zhao, University of Cambridge
Jack Bowden, University of Bristol
Gibran Hemani, University of Bristol
George Davey Smith, University of Bristol
Nancy R. Zhang, University of Pennsylvania
Dylan Small, University of Pennsylvania

Genetic association signals tend to be spread across the whole genome for complex traits. The recently proposed "omnigenic" model indicates that, when the risk factor is a complex trait, most genetic variants can weakly affect the risk factor while also easily affecting a common disease not through the risk factor. Existing methods in Mendelian Randomization (MR) are not ideal under such pervasive pleiotropy. We propose a comprehensive framework GRAPPLE (Genome-wide mR Analysis under Pervasive PLEiotropy) for MR, utilizing both strongly and weakly associated genetic variant, and can detect the existence of multiple pleiotropic pathways. We show that GRAPPLE is a comprehensive tool that can detect and adjust for pleiotropy, simultaneously estimate the causal effect of multiple risk factors and can determine the causal relationship direction using three-sample GWAS summary statistics datasets. With GRAPPLE, we conduct a screening for the lipid traits (HDL-C, LDL-C and triglycerides) with around 30 common diseases to understand their roles as risk factors and detect potential pleiotropic pathways.

email: jingshuw@uchicago.edu

Distinguishing Genetic Correlation from Causation in GWAS

Luke J. O'Connor*, Broad Institute
Alkes L. Price, Harvard T.H. Chan School of Public Health

Genome-wide association studies (GWAS) have revealed that many diseases and traits are have a shared genetic basis: many SNPs affect both traits, with correlated effect sizes, usually suggesting shared biology. In some cases, these correlations arise from causal relationships. We developed the latent causal variable (LCV) model to quantify causal relationships between genetically correlated traits using GWAS summary association statistics. We quantify what part of the genetic component of trait 1 is also causal for trait 2 using mixed fourth moments of genetic effect sizes for each trait. We validate this approach in extensive simulations. Across 52 traits (average N=331k), we identified 30 putative genetically causal relationships, many novel. Extending our method to multiple traits, we are able to learn the latent variables underlying pleiotropic relationships. These results demonstrate that it is possible to distinguish between genetic correlation and causation using genetic association data.

email: loconnor@broadinstitute.org

Robust Methods with Two-Sample Summary Data Mendelian Randomization

Hyunseung Kang*, University of Wisconsin, Madison

Mendelian randomization (MR) is a popular method in epidemiology to estimate the causal effect of an exposure on an outcome using genetic variants as instrumental variables (IV). Often, two-sample summary data is used in MR where in one sample, summary statistics about the marginal correlations between the IVs and the exposure are available and in another sample, summary statistics about the marginal correlations between the IVs and the outcome are available. Unfortunately, many methods in MR are biased under weak or invalid instruments, where the correlation between the IVs and exposure is small or the instruments have a a direct effect on the outcome. In this work, we propose estimators and tests for the exposure effect that (i) are robust to weak or invalid instruments and (ii) work with two-sample summary-level data. We also conduct simulation studies and conclude with a data exampe. This is joint work with Sheng Wang (UW-Madison) and Ting Ye (University of Pennsylvania).

email: hyunseung@stat.wisc.edu

---

43. RECENT ADVANCES IN STATISTICAL METHODS FOR SINGLE-CELL OMICS ANALYSIS

Fast and Accurate Alignment of Single-Cell RNA-seq Samples Using Kernel Density Matching

Mengjie Chen*, The University of Chicago
Yang Li, The University of Chicago
Qi Zhan, The University of Chicago

With technologies improved dramatically over recent years, single cell RNA-seq (scRNA-seq) has been transformative in studies of gene regulation, cellular differentiation, and cellular diversity. As the number of scRNA-seq datasets increases, a major challenge will be the standardization of measurements from multiple different scRNA-seq experiments enabling integrative and comparative analyses. However, scRNA-seq data can be confounded by severe batch effects and technical artifact. In addition, scRNA-seq experiments generally capture multiple cell-types with only partial overlaps across experiments making comparison and integration particularly challenging. To overcome these problems, we have developed a method, dmatch, which can both remove unwanted technical variation and assign the same cell(s) from one scRNA-seq dataset to their corresponding cell(s) in another dataset. By design, our approach can overcome compositional heterogeneity and partial overlap of cell types in scRNA-seq data. We further show that our method can align scRNA-seq data accurately across tissues biopsies.

email: mengjiechen@uchicago.edu

ABSTRACTS & POSTER PRESENTATIONS

Novel Computational Methods for Analyzing Single Cell Multi-Omics Data

Wei Chen*, University of Pittsburgh

Recent droplet-based single cell transcriptome sequencing (scRNA-seq) technology, largely represented by the 10X Genomics Chromium system, is able to measure the gene expression of tens of thousands of single cells from multiple individuals simultaneously. More recently, coupled with scRNA-seq technology, Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) and cell hashing have allowed us to simultaneously measure cell surface proteins, transcriptome profiling, and sample origin within the same cell. In this talk, I will present a probabilistic model for identifying sample origins for each cell and estimating multiplet rate based on cell hashing data. Then, I will present a model-based framework for jointly clustering single-cell multi-omics data with the aim to improve clustering accuracy and identify source-specific and shared clusters. In both simulation studies and real data analysis, our proposed methods can accurately identify cell origin and achieve satisfactory clustering performance. Our tools will facilitate the study design of single cell experiments and advance our fundamental knowledge of gene-protein relations.

email: weichen.mich@gmail.com

DNA Copy Number Profiling: From Bulk to Single-Cell Sequencing

Yuchao Jiang*, University of North Carolina, Chapel Hill

Copy number variation (CNV) is an important type of genetic variation that has been associated with diseases. High-throughput DNA sequencing enables detection of CNVs on the genome-wide scale with fine resolution, but suffers from many sources of biases and artifacts that lead to false discoveries and low sensitivity. In this talk, I will present statistical methods for DNA copy number profiling using both bulk-tissue and single-cell DNA sequencing. First I will present CODEX2, a statistical framework for CNV profiling by bulk DNA sequencing, which is sensitive for variants with both common and rare population frequencies and is applicable to study designs with and without negative control samples. In the second part of the talk, I will present SCOPE, a normalization and copy number estimation method for scDNA-seq data, which is sparse, noisy, and highly variable even within a homogeneous cell population. I will show that SCOPE more accurately estimates cancer subclonal copy number aberrations via benchmark studies and will further demonstrate on scDNA-seq data from the recently released Single Cell CNV Solution by the 10X Genomics.

email: yuchaoj@email.unc.edu

Statistical Analysis of Spatial Expression Pattern for Spatially Resolved Transcriptomic Studies

Xiang Zhou*, University of Michigan
Shiquan Sun, University of Michigan
Jiaqiang Zhu, University of Michigan

Recent development of various spatially resolved transcriptomic techniques has enabled gene expression profiling on complex tissues with spatial localization information. Detecting spatially expressed genes in these studies requires the development of statistical methods that can properly model spatial count data, provide effective type I error control, have sufficient statistical power, and are computationally efficient. Here, we developed such a method, SPARK. SPARK directly models count data generated from various spatial resolved transcriptomic techniques through generalized linear spatial models. With a new efficient penalized quasi-likelihood based algorithm, SPARK is scalable to data sets with tens of thousands of genes measured on tens of thousands of samples. Importantly, SPARK relies on newly developed statistical formulas for hypothesis testing, producing well-calibrated p-values and yielding high statistical power. We illustrate the benefits of SPARK through extensive simulations and in-depth analysis of four published spatially resolved transcriptomic data sets.

email: xzhousph@umich.edu

44. RECENT ADVANCES IN MICROBIOME DATA ANALYSIS

Incorporating Auxiliary Information to Improve Microbiome-Based Prediction Models

Michael C. Wu*, Fred Hutchinson Cancer Research Center

The microbiome, or the collection of microbes that inhabit the human body, plays a vital role in many areas of health and disease. The low cost of sequencing combined with easy collection of samples has led to considerable interest in the development of prediction and classification models which could be useful for diagnosis, prognosis, or risk assessment. However, the high-dimensionality and compositional nature of the data combined with inherent biological structure and relationships among the features pose serious challenges. These are exacerbated by the limited availability of samples. To mitigate some of these issues, we propose a strategy for construction of prediction models that harness auxiliary information to improve prediction accuracy. Auxiliary information can include structure such as phylogenetic and functional relationships among the microbes as well as other types of -omics data that may have been collected. We use simulations and applications to real microbiome experiments to show that the proposed methods can improve prediction accuracy.

email: mcwu@fredhutch.org

ABSTRACTS & POSTER PRESENTATIONS

ESTIMATION and INFERENCE WITH non-Random Missing Data and Latent Factors

Christopher McKennan*, The University of Pittsburgh

Metabolomics is the systematic study of tissue- or body fluid-specific small molecule metabolites, and has the potential to lead to new insights into the origin of human disease, as well as better understand host-microbe interactions. Recent technological advances have made it possible to collect high throughput metabolomic data, which are fraught with both non-ignorable missing observations and latent factors that influence a metabolite's measured concentration. However, current methods to analyze these data can only account for the missing data or latent factors, but not both. We therefore developed MetabMiss, a statistically rigorous method to account for both non-random missing data and latent confounding factors in high throughput metabolomics data. Our methodology does not require the practitioner specify a likelihood for the missing data, and makes investigating the relationship between the metabolome and tens, or even hundreds, of phenotypes computationally tractable. We demonstrate the fidelity of MetabMiss's estimates using both simulated and real metabolomics data, and prove their asymptotic correctness when the sample size and number of metabolites grows to infinity.

email: to come

Statistical Methods for Tree Structured Microbiome Data

Hongyu Zhao*, Yale University
Tao Wang, Shanghai Jiao Tong University
Yaru Song, Shanghai Jiao Tong University
Can Yang, Hong Kong University of Science and Technology

Recent advances in DNA sequencing technology have enabled rapid advances in our understanding of the contribution of the human microbiome to many aspects of normal human physiology and disease. A major goal of human microbiome studies is the identification of important groups of microbes that are predictive of host phenotypes, and the prediction of host phenotypes. However, the large number of bacterial taxa and the compositional nature of the data make this goal difficult to achieve using traditional approaches. Furthermore, the microbiome data are structured in the sense that bacterial taxa are not independent of one another and are related evolutionarily by a phylogenetic tree. In this presentation, we will describe some methods recently developed by us to incorporate the tree information as well as the compositional nature of the data for both selecting taxa that are associated with patient phenotypes and predicting host traits. The usefulness of our methods will be demonstrated through both simulations and real data applications. This is joint work with Tao Wang, Can Yang, and Yaru Song.

email: hongyu.zhao@yale.edu

High-Dimensional Log-Error-in-Variable Regression with Applications to Microbial Compositional Data Analysis

Anru Zhang*, University of Wisconsin, Madison
Pixu Shi, University of Wisconsin, Madison
Yuchen Zhou, University of Wisconsin, Madison

In microbiome and genomic studies, the regression of compositional data has been a crucial tool for identifying microbial taxa or genes that are associated with clinical phenotypes. To account for the variation in sequencing depth, the classic log-contrast model is often used where read counts are normalized into compositions. However, zero read counts and the randomness in covariates remain critical issues. In this article, we introduce a surprisingly simple, interpretable, and efficient method for the estimation of compositional data regression through the lens of a novel high-dimensional log-error-in-variable regression model. The proposed method provides both corrections on sequencing data with possible overdispersion and simultaneously avoids any subjective imputation of zero read counts. We provide theoretical justifications with matching upper and lower bounds for the estimation error. The merit of the procedure is illustrated through real data analysis and simulation studies.

email: anruzhang@stat.wisc.edu

---

45. NOVEL METHODS TO EVALUATE SURROGATE ENDPOINTS

Using a Surrogate Marker for Early Testing of a Treatment Effect

Layla Parast*, RAND
Tianxi Cai, Harvard University
Lu Tian, Stanford University

The development of methods to identify, validate and use surrogate markers to test for a treatment effect has been an area of intense research interest given the potential for surrogate markers to reduce the required cost and follow-up of future studies. Several quantities and procedures have been proposed to assess the utility of a surrogate marker. However, few methods have been proposed to address how to use the surrogate marker information to test for a treatment effect at an earlier time point, especially in settings where the primary outcome and the surrogate marker are subject to censoring. We propose a novel test statistic to test for a treatment effect using surrogate marker information measured prior to the end of the study in a survival setting. We propose a robust nonparametric estimation procedure and propose inference procedures. In addition, we evaluate the power for the design of a future study based on surrogate marker information. We illustrate the proposed procedure and relative power of the proposed test compared to a test performed at the end of the study using simulation studies and an application to data from the Diabetes Prevention Program.

email: parast@rand.org

ABSTRACTS & POSTER PRESENTATIONS

Mediation Analysis with Illness-Death Model for Right-Censored Surrogate and Clinical Outcomes

Isabelle Weir*, Harvard T.H. Chan School of Public Health
Jennifer Rider, Boston University
Ludovic Trinquart, Boston University

We introduce a counterfactual-based mediation analysis for surrogate outcome evaluation when both the surrogate and clinical endpoint are time-to-event outcomes subject to right-censoring. We use a multistate model for risk prediction to account for both direct transitions towards the clinical endpoint and transitions through the surrogate endpoint. We use the counterfactual framework to define the natural direct and indirect effects with a causal interpretation. Based on these measures, we define the proportion of the treatment effect on the clinical endpoint mediated by the surrogate endpoint. We define ratios and differences in both the cumulative risk and restricted mean survival time. We illustrate our approach using 18-year follow-up data from the SPCG-4 randomized controlled trial of radical prostatectomy for prostate cancer. We assess time to metastasis as a potential surrogate outcome for all-cause mortality.

email: iweir@sdac.harvard.edu

Incorporating Patient Subgroups During Surrogate Endpoint Validation

Emily Roberts*, University of Michigan
Michael Elliott, University of Michigan
Jeremy MG Taylor, University of Michigan

We consider the previously proposed principal surrogacy framework focused on the causal effect predictiveness (CEP) surface to model the joint distribution of four normally-distributed potential outcomes. In this work, we incorporate covariates in the statistical validation process of a surrogate endpoint using Bayesian methods. The framework can be extended to the non-normal setting by using a Gaussian copula model. By allowing the mean structure of potential outcomes to depend on covariates, we assess for which individuals the surrogate may be valid, i.e., where small causal effects on a surrogate are associated with small causal effects on the outcome, and where large positive/negative causal effects on the surrogate are associated with large positive/negative causal effects on the outcome. While adjusting for baseline patient characteristics, we explore to what extent patient subgroups affect the causal effect predictiveness surface. We assess the plausibility of conditional independence assumptions and reasonable prior distributions on correlation parameters subject to corresponding constraints on the covariance terms.

email: ekrobe@umich.edu

Assessing a Surrogate Predictive Value: A Causal Inference Approach

Ariel Alonso Abad*, University of Leuven
Wim Van der Elst, Janssen Pharmaceutica
Geert Molenberghs University of Leuven

Several methods have been developed for the evaluation of surrogate endpoints within the causal-inference and meta-analytic paradigms. In both paradigms effort has been made to assess the capacity of the surrogate to predict the causal treatment effect on the true endpoint. In the present work, the so-called surrogate predictive function (SPF) is introduced for that purpose, using potential outcomes. The relationship between the SPF and the individual causal association (ICA), a metric of surrogacy recently proposed in the literature, is described. It is shown that the SPF, in conjunction with the ICA, can offer an appealing quantification of the surrogate predictive value. The identifiability issues are tackled using a two-step procedure. In the first step, the region of the parametric space of the distribution of the potential outcomes, compatible with the data at hand, is geometrically characterized. Further, in a second step, a Monte Carlo approach is used to study the behavior of the SPF on the region. The method is illustrated using data from a clinical trial involving schizophrenic patients.

email: ariel.alonsoabad@kuleuven.be

46. RECENT ADVANCES IN THE UNCERTAINTY ESTIMATION AND PROPERTIES OF BAYESIAN ADDITIVE REGRESSION TREES

Heteroscedastic BART via Multiplicative Regression Trees

Matthew T. Pratola*, The Ohio State University
Hugh A. Chipman, Acadia University
Edward I. George, University of Pennsylvania
Robert E. McCulloch, Arizona State University

BART has become increasingly popular as a flexible and scalable nonparametric regression approach for modern applied statistics problems. For the practitioner dealing with large and complex nonlinear response surfaces, its advantages include a matrix-free formulation and the lack of a requirement to prespecify a confining regression basis. Although flexible in fitting the mean, BART has been limited by its reliance on a constant variance error model. Alleviating this limitation, we propose HBART, a nonparametric heteroscedastic elaboration of BART. In BART, the mean function is modeled with a sum of trees, each of which determines an additive contribution to the mean. In HBART, the variance function is further modeled with a product of trees, each of which determines a multiplicative contribution to the variance. Like the mean model, this flexible, multidimensional variance model is entirely nonparametric with no need for the prespecification of a confining basis. Moreover, with this enhancement, HBART can provide insights into the potential relationships of the predictors with both the mean and the variance.

email: mpratola@stat.osu.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Nonparametric Modeling with Tree Ensembles for Predicting Patient Outcomes

Robert E. McCulloch*, Arizona State University
Rodney Sparapani, Medical College of Wisconsin
Purushottam Laud, Medical College of Wisconsin
Brent Logan, Medical College of Wisconsin

Patients may respond differently to treatments. Flexible and efficient prediction models which can handle potentially complex interactions between patient factors and treatments are needed. Modern Bayesian semiparametric and nonparametric regression models provide an attractive avenue in this regard as these provide accurate predicts as well as natural posterior uncertainty quantification of predictions for patient response. We study an approach using extentions of Bayesian Additive Regression Trees (BART). BART has been shown to perform well in fitting nonparametric regression functions to continuous and binary responses, even with many covariates. However, BART assumes additive IID normal errors for a numeric outcome. In this research we consider extentions of BART which are both heteroskedastic and more fully nonparametric. Heteroskedastic, in that the conditional variability is allowed to depend on predictors, and nonparametric in that the underylying distribution is modeled with Bayesian nonparametrics rather than assumed to be normal.

email: Robert.McCulloch@asu.edu

Bayesian Decision Tree Ensembles in Fully Nonparametric Problems

Yinpu Li*, Florida State University
Antonio Linero, University of Texas, Austin
Junliang Du, Florida State University

In this talk, we introduce several extensions of BART to fully-nonparametric problems which are based on modulating an underlying Poisson process. This allows for simple Gibbs sampling algorithms to be derived for BART models for (i) density regression, (ii) nonparametric survival analysis, and (iii) estimating the intensity function of a spatial point process. Our algorithms are based on two layers of data augmentation, the first of which augments rejected points W from the modulated process and the second of which is an Albert-Chib type data augmentation step generating Z's which are treated as the response in a standard BART model. Taking advantage of the strong theoretical properties of certain BART priors, we are able to establish posterior concentration at near-minimax optimal rates for these problems, adaptively over a large class of function spaces. We illustrate our methodology on simulated and benchmark datasets.

email: yinpuli@icloud.com

On Theory for BART

Veronika Rockova*, The University of Chicago
Enakshi Saha, The University of Chicago

Ensemble learning is a statistical paradigm built on the premise that many weak learners can perform exceptionally well when deployed collectively. The BART method of Chipman et al. (2010) is a prominent example of Bayesian ensemble learning, where each learner is a tree. Due to its impressive performance, BART has received a lot of attention from practitioners. Despite its wide popularity, however, theoretical studies of BART have begun emerging only very recently. Laying the foundations for the theoretical analysis of Bayesian forests, Rockova and van der Pas (2017) showed optimal posterior concentration under conditionally uniform tree priors. These priors deviate from the actual priors implemented in BART. Here, we study the exact BART prior and propose a simple modification so that it also enjoys optimality properties. To this end, we dive into branching process theory. We obtain tail bounds for the distribution of total progeny under heterogeneous Galton-Watson (GW) processes exploiting their connection to random walks. We conclude with a result stating the optimal rate of posterior convergence for BART.

email: Veronika.Rockova@chicagobooth.edu

47. CURRENT DEVELOPMENTs IN ANALYZING EHR AND BIOBANK DATA

Adventures with Large Biomedical Datasets: Diseases, Medical Records, Environment and Genetics

Andrey Rzhetsky*, The University of Chicago

First, I will introduce our recent study analyzing phenotypic data harvested from over 175 million unique patients. I will explain how these non-genetic large-scale data can be used for genetic inferences. We discovered that complex diseases are associated with unique sets of rare Mendelian variants, referred to as the "Mendelian code." The second topic is about using electronic medical record data to 1) estimate the heritability and familial environmental patterns of diseases, and 2) infer the genetic and environmental correlations between disease pairs from a set of complex diseases. The third topic that I hope to cover is analysis of apparent clusters of neurodevelopmental and psychiatric disorders. Growing evidence is beginning to provide insight into how components of air pollution can be toxic to the brain: Our results showed that in the United States and Denmark, exposure to worst air quality was predictive of increase in the rates of a number of psychiatric conditions.

email: arzhetsky@uchicago.edu

ABSTRACTS & POSTER PRESENTATIONS

Association Analysis of Biobank Scale Data Using Minimal Sufficient Statistics

Dajiang Liu*, Penn State College of Medicine

Due to the decreasing cost high throughput sequencing and genotyping, large scale biobank datasets with up to half a million sample sizes have become commonly available. There can also be tens of thousands of traits. Together, biobank scale datasets may contain up to 1016 data entries, about ~10,000 times bigger than a GWAS datasets with 10,000 samples and 1 million genotyped variants. It may take 2 CPU years to complete the standard association analysis for all traits in a biobank scale dataset. The sheer size of biobank scale datasets quickly outdates existing software packages. There is a compelling need to develop more efficient tools that can scale well with these datasets. To address this research need, we develop a novel statistical method and computational implementations that make use of sufficient statistics to maximize dimension reduce, eliminate redundant computation while retaining all necessary information for association analysis. The methods can be hundreds times faster than the fastest available tools such as PLINK2. We expect that the new tool will play an important role in next generation sequencing and electronic medical record based genetic studies.

email: dajiang.liu@psu.edu

Use of Electronic Health Records and a Biobank for Pharmacogenomic Studies: Promises and Challenges

Leena Choi*, Vanderbilt University Medical Center

Electronic health records (EHRs) are an invaluable and rich source of clinical data and increasingly utilized for research. Detailed medication information is maintained for individual patients in EHRs, which provides a great opportunity for diverse medication-related studies including pharmacovigilance, studies of special populations, and (when clinical data are combined with genetic data from a biobank) pharmacogenomics. However, there are many challenges faced when attempting to use clinically-generated data for research, such as extraction of relevant data from EHRs and assurance of data quality. This talk covers our recent efforts in performing a pharmacogenomic study using the data from EHRs and a biobank and discusses challenges and our strategies to overcome some of these challenges.

email: leena.choi@vanderbilt.edu

Assessing the Progress of Alzheimer's Disease Via Electronic Medical Records

Zhijun Yin*, Vanderbilt University Medical Center

Alzheimer's Disease is the 6th leading cause of death in the U.S. and it is estimated that the cost will be $290 billion in 2019. Given its significant impact, it is important to investigate what factors contribute to the onset of Alzheimer's. Studies in this research area often focused on either building effective machine learning models to predict the onset of diagnosis by using images and genes, or identifying the genes that are significantly associated with the disease. Despite notability, these methods are limited in that 1) predicting the diagnosis of the disease focuses on the timepoint of the onset; 2) focusing only on genes may only obtain the prior information when the patient was born. Both methods ignore how the phenotypes along the lifetime contributed to the progress of Alzheimer's. We propose to leverage the electronic medical records (EMRs) to investigate to what extent the phenotypes that showed up five years or even ten years before the onset of Alzheimer's are associated with this disease. If success, our approach might help early screen of patients with Alzheimer's such that the onset of the disease could be postponed.

email: zhijun.yin@vanderbilt.edu

48. SPEED POSTERS: CAUSAL INFERENCE/ LONGITUDINAL DATA/HIGH-DIMENSIONAL DATA/MASSIVE DATA

48a. Bipartite Causal Inference with Interference for Evaluating Air Pollution Regulations

Corwin M. Zigler*, University of Texas, Austin and Dell Medical School

A fundamental feature of evaluating causal health effects of air quality regulations is that air pollution moves through space, rendering health at a particular population location dependent upon actions taken at multiple pollution sources. Motivated by studies of the impacts of power plant regulations in the U.S., we introduce bipartite causal inference with interference, which arises when 1) treatments are defined on units that are distinct from those at which outcomes are measured and 2) there is interference between units in the sense that outcomes for some units depend on the treatments assigned to many other units. Interference in this setting arises due to complex exposure patterns dictated by physical-chemical atmospheric processes, with intervention effects framed as propagating across a bipartite network of power plants and residential zip codes. New causal estimands are introduced, along with an estimation approach based on generalized propensity scores. The new methods are deployed to estimate how emission-reduction technologies implemented at coal-fired power plants causally affect health outcomes among Medicare beneficiaries in the U.S.

email: cory.zigler@austin.utexas.edu

ABSTRACTS & POSTER PRESENTATIONS

48b. Doubly Robust Estimation of Causal Effects with Covariate-Balancing Propensity Score and Machine-Learning-Based Outcome Prediction

Byeong Yeob Choi*, University of Texas Health Science Center at San Antonio

Doubly robust estimation has an advantage in that it produces an unbiased estimator for the average treatment effect unless both propensity score (PS) and outcome models are incorrectly specified. Among PS estimation methods, covariate-balancing propensity score (CBPS) has been noted as a promising method. In this simulation study, we evaluated if combinations of CBPS and various machine learning methods for outcome prediction can enhance the robustness of the doubly robust estimators to model misspecification. We considered four types of outcome prediction methods: least squares, tree-based methods, generalized additive models and shrinkage methods. Simulation scenarios differed by the complexity of treatment and outcome generating models. In simulations, generalized additive models and boosted models yielded more robust doubly robust estimators in terms of reducing bias. Using CBPS instead of maximum likelihood reduced the variance in complex scenarios of PS model when treatment prevalence was low.

email: choib@uthscsa.edu

48c. Percentile-Based Residuals for Model Assessment

Sophie Berube*, Johns Hopkins Bloomberg School of Public Health
Abhirup Datta, Johns Hopkins Bloomberg School of Public Health
Chenguang Wang, Johns Hopkins Bloomberg School of Public Health
Qingfeng Li, Johns Hopkins Bloomberg School of Public Health
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health

Residuals are a key component of diagnosing model fit. The usual practice is to compute standardized residuals using expected values and standard deviations of the observed data, then use these values to detect outliers and assess model fit. Approximate normality of these residuals is key for this process to have good properties, but in many modeling contexts, especially for complex, multi-level models, normality may not hold. In these cases, outlier detection and model diagnostics aren't properly calibrated. Residuals computed from the percentile location of a datum's value in its full predictive distribution lead to well calibrated evaluations of model fit. We generalize an approach described by Dunn and Smyth (1996) and evaluate properties mathematically, via case-studies and by simulation. For the percentile-based residuals, the use of full predictive distributions with the appropriate location, spread and shape is necessary for valid assessments.

email: sberube3@jhmi.edu

48d. Change-Point Detection in Multivariate Time Series

Tong Shen*, University of California, Irvine
Xu Gao, Google
Hernando Ombao, King Abdullah University of Science and Technology
Zhaoxia Yu, University of California, Irvine

The goal of this article is to develop a two-stage approach to identifying change points in the brain activity that is recorded by multiple electroencephalograms (EEG) channels and the change in stock market data due to financial crisis. Our proposed procedure has two stages. In the first stage, we obtain a low-dimensional summary of the high-dimensional time series by spectral principal component analysis (PCA). In the second stage, we use a cumulative sum-type test to the spectral PCA component using a binary segmentation algorithm. Our simulations indicate that our new approach performs significantly better than two competing methods on summarizing high-dimensional time series. We apply this method on epileptic seizure EEG data and stock data and detect some change points corresponding to the onset of different issues.

email: tshen4@uci.edu

48e. Approaches for Modeling Spatially Varying Associations Between Multi-Modal Images

Alessandra M. Valcarcel*, University of Pennsylvania
Simon N. Vandekar, Vanderbilt University
Tinashe Tapera, University of Pennsylvania
Azeez Adebimpe, University of Pennsylvania
David Roalf, University of Pennsylvania
Armin Raznahan, National Institute of Mental Health, National Institutes of Health
Theodore Satterthwaite, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania
Kristin Linn, University of Pennsylvania

Multi-modal magnetic resonance imaging modalities quantify different, yet complimentary, properties of the brain and its activity. When studied jointly, multi-modal imaging data may improve our understanding of the brain. Unfortunately, the vast number of imaging studies evaluate data from each modality separately and do not consider information encoded in the relationships between imaging types. We aim to study the complex relationships between multiple imaging modalities and map how these relationships vary spatially across different anatomical regions of the brain. Given a particular voxel location in the brain, we regress an outcome image modality on the remaining modalities using all voxels in a local neighborhood of the target voxel. In an exploratory analysis, we compare the performance of three estimation frameworks that account for the spatial dependence among voxels in a neighborhood: generalized linear models (GEE), linear mixed effects models with varying random effect structures, and weighted least squares. We apply our framework to a large imaging study of neurodevelopment to study the relationship between local functional connectivity and cerebral blood flow.

email: alval@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

48f. Generalizing Trial Findings using Nested Trial Designs with Sub-Sampling of Non-Randomized Individuals

Sarah E. Robertson*, Brown University
Issa J. Dahabreh, Brown University
Miguel A. Hernan, Harvard University
Ashley L. Buchanan, University of Rhode Island
Jon A. Steingrimsson, Brown University

To generalize inferences from a randomized trial to a target population, investigators can use a nested trial design, embedding the trial within a cohort that is a simple random sample from the target population. In this design, data on routinely collected baseline covariates are available from the entire cohort, and treatment and outcome data need only be collected from the trial. Data on additional baseline covariates are often necessary for valid generalization and are collected in the trial but may be too expensive to collect from all individuals. We describe a novel two-stage nested trial design that improves research economy by collecting additional baseline covariate data only from a sub-sample of non-randomized individuals, using sampling probabilities that may depend on baseline covariates available from all individuals in the cohort. We propose a doubly robust estimator for potential outcome means in the target population under this design, obtain the estimator's large-sample distribution, and examine its finite-sample performance in a simulation study. We illustrate the methods using data from the Coronary Artery Surgery Study.

email: sarah_robertson@brown.edu

48g. Causal Inference with Multiple Mediators in a Survival Context

Hui Zeng*, The Pennsylvania State University
Vernon Michael Chinchilli, The Pennsylvania State University

VanderWeele (2011) derived the natural direct and natural indirect effects in causal mediation analysis with survival data having one mediator. VanderWeele showed an approach for accelerated failure time models in general cases or a proportional hazards model with a rare outcome. We extend the approach to handle more than one mediator for both accelerated failure time models and proportional hazards models, which does not require the assumption for a rare outcome. Since there are multiple mediators, we consider different scenarios for the relationship between these mediators and the interactions of exposure and mediators. We describe statistical inference and explanations for the natural direct and natural indirect effects based on the parameters in the model. We evaluate the performance of the approach via simulations and illustrate the approach by applying it to a multi-center prospective cohort study on acute kidney injury.

email: huz156@psu.edu

48h. Adjusting for Compliance in SMART Designs

William Jeremy Artman*, University of Rochester
Ashkan Ertefaie, University of Rochester
Brent Johnson, University of Rochester

Sequential, multiple assignment, randomized trial (SMART) designs are an important platform for rigorous comparison of sequences of treatments tailored to the individual patient, i.e., dynamic treatment regime (DTR). Identification of optimal DTR promises an alternative to adhoc one-size-fits-all decisions pervasive in patient care. The standard approach to analyzing a SMART is the intention-to-treat (ITT) principal. However, in the presence of non-compliance, ITT analyses may lead to substantially biased estimates of DTR outcomes. Principal stratification is a powerful tool which stratifies patients according to potential compliance profiles and offers an alternative to ITT. An important statistical challenge in adjusting for non-compliance is the number of potential compliances and their correlation structure. We fill the current methodological gap by developing a rigorous principal stratification framework that leverages a flexible Bayesian semiparametric model. We demonstrate the validity of our method through extensive simulation studies. We illustrate its application on an actual SMART.

email: William_Artman@urmc.rochester.edu

48i. Statistical Inference for Cox Proportional Hazards Model with a Diverging Number of Covariates

Lu Xia*, University of Michigan
Bin Nan, University of California, Irvine
Yi Li, University of Michigan

We consider hypothesis testing and confidence intervals in the Cox proportional hazards (Cox PH) model with a diverging number of covariates. We propose to use a quadratic programming procedure to estimate the large inverse of the Fisher information matrix in a de-biasing lasso framework. This quadratic programming procedure is computationally fast and does not require l_0 sparsity of the true inverse matrix, an assumption that generally does not hold in the Cox PH model. An adaptive tuning parameter selection procedure for matrix estimation is proposed. Simulation shows that our proposed method renders reliable confidence interval coverage and improves bias correction. We derive the asymptotic properties for linear combinations of such de-biased coefficient estimates.

email: luxia@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

48j. Bayesian Focal-Area Detection for Multi-Class Dynamic Model with Application to Gas Chromatography

Byung-Jun Kim*, Virginia Tech

Accurately estimating gas chromatograms and identifying chemicals from the mixture of volatile compounds are challenging problems in analytical chemistry. When there are various types of the compounds as mixture, it has difficulty in recognizing the correct patterns of the compounds due to unknown tangled effects of the samples with unknown classes. Motivated by the practical problem of the pattern recognition in gas chromatography, we develop a nonparametric focal-area detection method for multi-class dynamic model using a fully Bayesian hierarchical modeling. Our goal is to estimate unknown functional trends of multilevel compounds in mixture and detect specific focal areas for the efficient identification. Our proposed method can account for the random mixed effect with unknown class on the estimation of the trends using MCMC method. By using two shrinkage priors under Bayesian framework, our method can efficiently estimate an unknown functional trend and have computational efficiency on the estimation of new data based on the significant area detection.

email: bjkim702@vt.edu

48k. The Survivor Separable Effects

Mats Julius Stensrud*, Harvard T. H. Chan School of Public Health
Miguel Hernan, Harvard T. H. Chan School of Public Health
Jessica Julius Young, Harvard Medical School

Many researchers aim to study treatment effects on outcomes that are truncated by death. In these settings, a naive contrast of outcomes in subjects who are alive does not have a causal interpretation, even if the treatment is randomized. Therefore the outcome in the principal stratum of always survivors, i.e. the survivor average causal effect, is often advocated for causal inference. The survivor average causal effect is a well defined causal contrast, but it is often hard to justify that it is relevant to scientists, patients or decision makers, and it cannot be identified without relying on strong untestable assumptions. Here we present a new estimand in truncation by death settings, which allows us to define the survivor separable effects. We describe the causal interpretation of the survivor separable effects, and introduce three different estimators. Finally, we apply this approach to estimate the effect of chemotherapies on quality of life in patients with prostate cancer using data from a randomized clinical trial.

email: mstensrud@hsph.harvard.edu

48l. Adjusted Cox Scores for GWAS and PheWAS Screening in R

Elizabeth A. Sigworth*, Vanderbilt University
Ran Tao, Vanderbilt University
Frank Harrell, Vanderbilt University
Qingxia Chen, Vanderbilt University

Genome-wide and phenome-wide association studies (GWAS and PheWAS) have gained momentum in recent years as methods for studying the relationship between genetic variation and individual health outcomes. These studies are often performed on incredibly high-dimensional data sets, and as such employ logistic models for their computational speed. However, logistic models ignore much of the time-related information that can be extracted from longitudinal health records, in particular time-to-event outcomes. Survival models like the Cox proportional hazards (PH) model can incorporate this information, but are time and resource-intensive for high-dimensional datasets. Most survival packages focus on the Wald test for Cox PH model, making it necessary to fit a separate model for each gene. Meanwhile, the adjusted Cox scores approach (adjusted for a set of covariates common to all models such as demographics and genetic ancestry) can perform genetic screening without needing to fit the full Cox PH model for each gene. As such, we develop a memory-efficient and computationally fast R package to implement this approach, tailored for GWAS and PheWAS studies.

email: elizabeth.a.sigworth@vanderbilt.edu

48m. Microbiome Quantile Regression

Myung Hee Lee*, Weill Cornell Medical College

Cervicovaginal microbiome data are collected from a longitudinal disease cohort study. Of interest is the association between a continuous scale burden of disease and the microbiota composition at baseline. While some bacteria are associated with central outcome, others may be associated with other parts of outcome distributions such as mildly ill patients (lower quantile) or severely ill patients (upper quantile). Identification of bacteria associated with different parts of distribution might provide deeper insights to clinical investigators. We consider regression problem where compositional data are used as covariates. In particular, we will formulate quantile regression by assuming that the distribution of errors follow location and scale mixtures of normal distributions. This approach has been shown to provide solution via a simple EM algorithm and be generalizable for longitudinal quantile regression. We analyze baseline micboriome data and highlights of our findings will be presented.

email: myl2003@med.cornell.edu

ABSTRACTS & POSTER PRESENTATIONS

49. STATISTICAL METHODS FOR OMICS DATA ANALYSIS

Mean-Correlation Relationship Biases Co-Expression Analysis

Yi Wang*, Johns Hopkins Bloomberg School of Public Health
Stephanie C. Hicks, Johns Hopkins Bloomberg School of Public Health
Kasper D. Hansen, Johns Hopkins Bloomberg School of Public Health and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine

“Estimates of correlation between pairs of genes in co-expression analysis are commonly used to construct networks between genes using gene expression data. Here, we show that the distribution of these correlations depends on the expression level of the involved genes, which we refer to this as the mean-correlation relationship in bulk RNA-seq data. This dependence introduces a bias in co-expression analysis whereby highly expressed genes are more likely to be highly correlated. Ignoring this bias can lead to missing potentially biologically relevant pairs of genes that are lowly expressed, such as transcription factors. To address this problem, we introduce spatial quantile normalization, a method for normalizing local distributions in a correlation matrix. We show that spatial quantile normalization removes the mean-correlation relationship and corrects the expression bias in network reconstruction.

email: yiwangthu4@gmail.edu

Efficient Detection and Classification of Epigenomic Changes Under Multiple Conditions

Pedro L. Baldoni*, University of North Carolina, Chapel Hill
Naim U. Rashid, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

In epigenomics, the study of protein-DNA interactions has been of special interest in recent years with the advances and reduced costs of high-throughput assays. These interactions, which may affect DNA transcription, repair, replication, and recombination, have been associated with several complex human disorders. The chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is one of the existing techniques to detect such interactions. Of particular interest is the detection and classification of differential interacting sites between conditions to illuminate treatment responses, for instance. However, data from ChIP-seq assays exhibit a diverse profile of experimental signal and most methods are not optimized for scenarios where broad protein-DNA binding sites are observed. To address these challenges, we present a new flexible and efficient method for the detection and classification of broad differential binding sites across multiple conditions based on a hidden Markov model with finite mixture model as emission distribution. Using data from the ENCODE project, we show that the proposed method outperforms current algorithms in several scenarios.

email: baldoni@email.unc.edu

BREM-SC: A Bayesian Random Effects Mixture Model for Clustering Single Cell Multi-Omics Data

Xinjun Wang*, University of Pittsburgh
Zhe Sun, University of Pittsburgh
Yanfu Zhang, University of Pittsburgh
Heng Huang, University of Pittsburgh
Kong Chen, University of Pittsburgh
Ying Ding, University of Pittsburgh
Wei Chen, University of Pittsburgh

Droplet-based single cell transcriptome sequencing (scRNA-seq) technology is able to measure the gene expression from tens of thousands of single cells simultaneously. More recently, coupled with the cutting-edge Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq), the droplet-based system has allowed for immunophenotyping of single cells based on cell surface expression of specific proteins together with simultaneous transcriptome profiling in the same cell. Despite the rapid advances in technologies, novel statistical methods and computational tools for analyzing multi-modal CITE-Seq data are lacking. In this study, we developed BREM-SC, a novel Bayesian Random Effects Mixture model that jointly clusters paired single cell transcriptomic and proteomic data. Through simulation studies and analysis of public and in-house real data sets, we successfully demonstrated the validity and advantages of this method in fully utilizing both types of data to accurately identify cell clusters. This new method will greatly facilitate researchers to jointly study transcriptome and surface proteins at the single cell level to make new biological discoveries.

e-mail: XIW119@pitt.edu

ABSTRACTS & POSTER PRESENTATIONS

Co-Localization Between Sequence Constraint and Epigenomic Information Improves Interpretation of Whole Genome Sequencing Data

Danqing Xu*, Columbia University
Chen Wang, Columbia University
Krzysztof Kiryluk, Columbia University
Joseph D. Buxbaum, Icahn School of Medicine at Mount Sinai
Iuliana Ionita-Laza, Columbia University

The identification of tissue/cell type specific functional regions in the noncoding human genome is critical for understanding the role of noncoding variation in gene regulation in health and disease. We describe here a co-localization approach and provide putative tissue/cell type specific regulatory regions under sequence constraint within the human lineage for 127 tissues/cell types in the ENCODE/Roadmap Epigenomics Project. We show that the co-localization of sequence constraint and tissue specific regulatory function in regions proximal to transcription start sites correlates with the probability for genes to be intolerant to loss-of-function variation. Noncoding pathogenic variants in ClinVar are more likely to fall in co-localized regions than benign variants. We use the developed co-localization score for brain tissues to score de novo mutations from 1,902 individuals affected with autism spectrum disorder (ASD) and their unaffected siblings in the Simons Simplex Collection. We show that noncoding de novo mutations near genes co-express in midfetal brain with high confidence ASD risk genes, and near FMRP gene targets are more likely to be in co-localized regions.

email: elisexu0308@gmail.com

Covariate Adaptive False Discovery Rate Control with Applications to Epigenome-Wide Association Studies

Jun Chen*, Mayo Clinic
Xianyang Zhang, Texas A&M University

Conventional multiple testing procedures often assume hypotheses are exchangeable. However, in many scientific applications, additional covariates that are informative of the null probability or statistical power of the underlying hypothesis are available. Leveraging such auxiliary covariates could potentially increase the detection power. In this talk, I will introduce a covariate-adaptive false discovery rate control procedure we recently developed to conveniently incorporate covariate information. We show that our method improves over existing procedures by being simultaneously flexible, robust, powerful and computationally efficient. We further benchmark our method on 61 datasets arising from epigenome-wide association studies. We show that the method is overall the most powerful among competing methods, especially when the signal is sparse.

email: chen.jun2@mayo.edu

Estimation of Cell-Type Proportions in Complex Tissue

Gregory J. Hunt*, William & Mary
Johann A. Gagnon-Bartsch, University of Michigan

Human tissue is comprised of a large number of different types of cells. Each cell type is involved in a multitude of important biological processes like signaling, blood flow, immune interactions, etc. Thus an important component in the understanding of such fundamental processes is understanding the heterogeneity among the various types of cells. Additionally, understanding cell-type heterogeneity can help clarify and de-confound other biological endpoints of interest. For this reason, methods to estimate cell type proportions from gene expression data have been extensively studied over the past decade. In the context of high-throughput gene expression data the problem of estimating cell-type proportions is known as cell-type deconvolution. Typically deconvolution methods estimate the proportion of cell types in a mixture by comparing the gene expression in the mixture to gene expressions from reference profiles of the constituent cell types. In this work we explore the literature of deconvolution methods and propose a new method that is both biologically plausible and statistically efficient.

email: hunt.gregory.james@gmail.com

---

50. OBSERVATIONAL AND HISTORICAL DATA ANALYSIS: THE REST IS HISTORY

Identifying the Optimal Timing of Surgery from Observational Data

Xiaofei Chen*, Southern Methodist University and University of Texas Southwestern Medical Center
Daniel F. Heitjan, Southern Methodist University and University of Texas Southwestern Medical Center
Gerald Greil, University of Texas Southwestern Medical Center
Haekyung Jeon-Slaughter, University of Texas Southwestern Medical Center

Infants with hypoplastic left heart syndrome require an initial Norwood operation, followed some months later by a stage 2 palliation (S2P). The timing of S2P is critical for the operation's success and the infant's survival, but the optimal timing, if one exists, is unknown. We attempt to identify the optimal timing of S2P by applying an extension of propensity score analysis to data from the Single Ventricle Reconstruction Trial (SVRT). In the SVRT, the procedure used in the initial Norwood operation was randomized, but the timing of the S2P was chosen idiosyncratically. The trial constitutes a thoroughly documented observational study. We model the time to surgery as a function of confounders using a discrete competing-risk model. We then apply inverse probability weighting to estimate a spline hazard model for predicting survival from the time of S2P. Our analysis suggests that S2P conducted at 6 months after the Norwood gives the patient the best chance of survival.

email: xiaofeic@smu.edu

ABSTRACTS & POSTER PRESENTATIONS

Historical Control Borrowing in Adaptive Designs "To Borrow or Not to Borrow?"

Nusrat Harun*, Cincinnati Children's Hospital Medical Center
Mi-Ok Kim, University of California, San Francisco
Maurizio Macaluso, Cincinnati Children's Hospital Medical Center

Historical control borrowing may increase the power, lead to early termination, and reduce the sample size of clinical trials depending on whether the historical and concurrent controls are commensurate. We simulated data from motivating stroke trials to evaluate the performance of a selective response-adaptive design and evaluated the operating characteristics with commensurate and noncommensurate historical controls when both borrowing and not borrowing. Increased power, greater probability to stop early correctly, and smaller average sample size were observed with borrowing compared to not borrowing irrespective of commensurability. The type I error was inflated while both borrowing and not borrowing using noncommensurate historical controls. Although larger type I error inflation was noted with borrowing, there was greater probability to stop early correctly for efficacy. Higher probability to stopping the trial early correctly at the cost of an inflated type I error with noncommensurate historical controls favor borrowing.

email: nusrat.harun@cchmc.org

Weighted F Test and Weighted Chi-Square Test for Multiple Group Comparisons in Observational Studies

Maiying Kong*, University of Louisville
Xiaofang Yan, University of Louisville
Qi Zheng, University of Louisville

Although F test and chi-square test are commonly used for multiple group comparisons in experimental data, these methods can not be directly used to examine group differences in observational studies. The propensity-score-based inverse probability weighting (IPW) method has become one of the popular methods for estimating average treatment effect (ATE), however, the IPW method has only been applied to compare pairs among multiple treatment groups without control the family-wise error rate (FWER). We propose to examine whether there is an overall significant group difference using a weighted F test for continuous outcome variable and a weighted chi-square test for categorical outcome variable. Only if there is an overall significant group difference, the pairs of interests are further compared. We apply the proposed weighted chi-square test to investigate whether fruit/vegetable intakes are associated with cardiovascular diseases using the 2015 KY BRFSS data set, and we apply the weighted F-test to examine the effect of physical/recreational exercise on weight gain using the NHEFS data set.

email: maiying.kong@louisville.edu

Bayesian Probability of Success of Clinical Trials for the Generalized Linear Model Using Historical Data

Ethan M. Alt*, University of North Carolina, Chapel Hill
Matthew A. Psioda, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Given the cost and duration of phase III and phase IV clinical trials, the development of statistical methods for go/no-go decisions is vital. In this talk, we introduce a Bayesian methodology to compute the probability of success based on the current data of a treatment regimen for the generalized linear model. The method allows for the inclusion of covariates as well as historical data based on the treatment regimen and patient characteristics. The method generalizes the work of Ibrahim et al. and Chuang-Stein and will be indispensable for informing the scientist on the likelihood of success of the trial with respect to hypotheses regarding primary and key secondary endpoints.

email: e.alt@me.com

Borrowing Strength from Auxiliary Variables and Historical Data for Counties with very Small Sample Sizes or No Data

Hui Xie*, Centers for Disease Control and Prevention
Deborah B. Rolka, Centers for Disease Control and Prevention
Lawrence Barker, Centers for Disease Control and Prevention

Bayesian hierarchical regression (BHR) is widely applied in small area estimation. It is used to estimate county-level parameters of interest. It is common to stratify the survey samples by demographic characteristics/socioeconomic measurements. This can result in strata with very small sample sizes, even zero, within counties. In such cases, posterior estimation strongly depends on the prior distribution. To address this, we propose a new approach including four features: clustering counties using auxiliary variables via Gaussian mixture model; borrowing strength from other counties within a cluster, rather than across the whole data set; imputing the likelihood function for empty strata via EM algorithms; and using historical data to construct a power prior. We applied this approach to estimate county-level disability prevalence using data from the BRFSS. We validated our results by comparing them to 1-year estimates from the ACS. In this comparison, our new approach performed better than BHR models that borrow strength across all counties/states or spatial correlations, especially for counties with little to no survey data.

email: hxie@cdc.gov

ABSTRACTS & POSTER PRESENTATIONS

Adaptive Combination of Conditional Treatment Effect Estimators Based on Randomized and Observational Data

David Cheng*, VA Boston Healthcare System
Ross Prentice, University of Washington School of Public Health and Community Medicine
Tianxi Cai, Harvard T.H. Chan School of Public Health

Data from both a randomized trial and an observational study are sometimes simultaneously available for evaluating the effect of an intervention. The randomized data typically allows for reliable estimation of average treatment effects but may be limited in sample size and patient heterogeneity for estimating conditional average treatment effects. Estimates from the observational study can compensate for these limitations, but there may be concerns about whether confounding and treatment effect heterogeneity have been adequately addressed. We propose an approach for combining conditional treatment effect estimators from each source such that it aggressively weights toward the randomized estimator when bias in the observational estimator is detected. When the bias is negligible, the estimators from each source are combined for optimal efficiency. We show the problem can be formulated as a penalized least squares problem and consider its asymptotic properties. We assess its performance in simulations and illustrate the approach by estimating the effects of hormone replacement therapy on the risk of coronary heart disease in data from the Women's Health Initiative.

email: dcheng01@fas.harvard.edu

51. IMMUNOTHERAPY CLINICAL TRIAL DESIGN AND ANALYSIS

Time-to-Event Model-Assisted Designs to Accelerate and Optimize Early-Phase Immunotherapy Trials

Ruitao Lin*, University of Texas MD Anderson Cancer Center

Immunotherapies and molecularly targeted agents have revolutionized cancer treatment. Unlike chemotherapies, these novel agents often take a longer time to show responses. This causes major logistic difficulty for implementing existing adaptive trial designs, which require the observance of the outcome early enough to apply data-adaptive decisions for new patients. In this talk, I will introduce a novel class of Bayesian adaptive designs, known as time-to-event model-assisted designs, to address this practical challenge in phase I dose-finding trials with late-onset toxicity. A unified methodology based on a novel formulation and approximation of the observed data likelihood will be introduced to facilitate seamless, real-time decision making. The dose escalation/de-escalation rules of the proposed designs can be tabulated before the trial begins, which greatly simplifies trial conduct in practice compared to that under existing methods. I will present some theoretical and numerical results to show the desirable properties of the proposed designs. Last, I will introduce user-friendly software for implementing the designs.

email: ruitaolin@gmail.com

Designing Cancer Immunotherapy Trials with Delayed Treatment Effect Using Maximin Efficiency Robust Statistics

Xue Ding*, University of Kentucky
Jianrong Wu, University of Kentucky

The indirect mechanism of action of immunotherapy causes a delayed treatment effect, producing delayed separation of survival curves between the treatment groups, and violates the proportional hazards assumption. Therefore using the log-rank test could result in a severe loss efficiency. Although few statistical methods are available for immunotherapy trial design that incorporate a delayed treatment effect, Ye and Yu proposed use of a maximin efficiency robust test (MERT) for the trial design. However, the weight function of the MERT involves an unknown function which has to be estimated from historical data. Here, for simplicity, we propose use of an approximated maximin test, the V_0 test, which is the sum of the log-rank test for the full data set and the log-rank test for the data beyond the lag time point. The V_0 test fully uses the trial data and is more efficient than the log-rank test when lag exits with relatively little efficiency loss when no lag exists. Simulations are conducted to compare the performance of the V_0 test to the existing tests. A real trial is used to illustrate cancer immunotherapy trial design with delayed treatment effect.

email: xdi226@g.uky.edu

Cancer Immunotherapy Trial Design with Cure Rate and Delayed Treatment Effect

Jing Wei*, University of Kentucky
Jianrong Wu, University of Kentucky

Cancer immunotherapy trials have two special features: a delayed treatment effect and a cure rate. Both features violate the proportional hazards model assumption and ignoring either one of the two features in an immunotherapy trial design will result in substantial loss of statistical power. To properly design immunotherapy trials, we proposed a piecewise proportional hazards cure rate model to incorporate both delayed treatment effect and cure rate into the trial design consideration. A sample size formula is derived for a weighted log-rank test under a fixed alternative hypothesis. The accuracy of sample size calculation using the new formula is assessed and compared with the existing methods via simulation studies. A real immunotherapy trial is used to illustrate the study design along with practical consideration of balance between sample size and follow-up time.

email: jwe239@g.uky.edu

Cancer Immunotherapy Trial Design with Long-Term Survivors

Jianrong Wu*, University of Kentucky
Xue Ding, University of Kentucky

Cancer immunotherapy often reflects the mixture of improvement in short-term risk reduction and long-term survival. In a cancer immunotherapy trial with improvement in both short-term risk reduction and long-term survival, the hazard functions between two groups will ultimately cross over. Thus,

ABSTRACTS & POSTER PRESENTATIONS

standard log-rank test will be inefficient to detect the difference of long-term survival. In this talk, we propose two weighted log-rank tests for the trial designs. Practical consideration of cancer immunotherapy trial designs with long-term survivors are discussed.

email: jianrong.wu@uky.edu

Evaluate the Properties of Cure Model in the Context of Immuno-oncology Trials

Quyen Duong*, Mayo Clinic
Jennifer Le-Rademacher, Mayo Clinic

Immunotherapy has been approved as a treatment for various cancers in recent years. In confirmatory trials in oncology, log-rank test is used as the basis for sample size estimation and treatment effect is often quantified by the hazard ratio. The log-rank test is the most powerful and the hazard ratio is most interpretable under the proportional hazards model. However, results of recent immuno-oncology trials indicate that the effects of immunotherapy are often delayed and the assumption of proportional hazards do not hold in immuno-oncology trials. Alternatively, cure model can evaluate the effect of immunotherapy on survival where patients whose response lasts well beyond treatment duration can be considered being cured of their cancer. Cure model can simultaneously evaluate the effect of the new treatment on the cure rate as well as its effect on survival for uncured patients. Simulation study was conducted on various scenarios to evaluate the effect of design parameters on the trial statistical power using the R package NPHMC. It will be compared to the power of the log-rank test.

email: duong.quyen@mayo.edu

Phase I/II Dose-Finding Interval Design for Immunotherapy

Yeonhee Park*, Medical University of South Carolina

Immunotherapeutics have revolutionized the treatment of metastatic cancers and are expected to play an increasingly prominent role in the treatment of cancer patients. Recent advances in checkpoint inhibition show promising early results in a number of malignancies, and several treatments have been approved for use. However, the immunotherapeutic agents have revealed to have different toxicity profiles and mechanism of antitumor activity from the cytotoxic agents, and many limitations and challenges encountered in the traditional paradigm were recently pointed out for immunotherapy. We propose a utility-based method to determine optimal biological dose of immunotherapeutics by effectively utilizing toxicity, immune response, and tumor response. Moreover, we propose a new algorithm to allocate the dose for next cohort which makes dose transition more appropriate and does not require statistical analysis to make a decision. Simulation studies show that the proposed design has desirable operating characteristics compared to existing dose-finding designs. It also inherits strengths of dose-finding interval designs to have good performance with the simplicity to implement.

email: parkye@musc.edu

52. MACHINE LEARNING AND STATISTICAL RELATIONAL LEARNING

Merging versus Ensembling in Multi-Study Machine Learning: Theoretical Insight from Random Effects

Zoe Guan*, Harvard T.H. Chan School of Public Health, Dana-Farber Cancer Institute
Giovanni Parmigiani, Harvard T.H. Chan School of Public Health, Dana-Farber Cancer Institute
Prasad Patil, Boston University

A critical decision point when training predictors using multiple studies is whether these studies should be combined or treated separately. We compare two multi-study learning approaches in the presence of potential heterogeneity in predictor-outcome relationships across datasets. We consider 1) merging all of the datasets and training a single learner, and 2) cross-study learning, which involves training a separate learner on each dataset and combining the resulting predictions. In a linear regression setting, we show analytically and confirm via simulation that merging yields lower prediction error than cross-study learning when the predictor-outcome relationships are relatively homogeneous across studies. However, as heterogeneity increases, there exists a transition point beyond which cross-study learning outperforms merging. We provide an analytic expression for the transition point that can be used to help guide decisions about whether to merge data from multiple studies.

email: zguan@g.harvard.edu

Informative Dynamic ODE-based-Network Learning (IDOL) from Steady Data

Chixiang Chen*, The Pennsylvania State University
Ming Wang, The Pennsylvania State University
Rongling Wu, The Pennsylvania State University

Dynamic gene regulation networks (GRN) have been widely used as a tool to study and infer the casual relationship of genes that drive genotype to phenotype. However, reconstructing this type of networks critically relies on longitudinal expression data, which may not be available in many studies, such as the GTEx project. In this article, we propose a general framework for recovering dynamic GRN from steady-state data. We incorporate the varying coefficient model with ordinary differential equations to learn informative networks linking to known or unobserved index, which could be time sequence, continuous scaled exposure, disease risk, etc. We derive the asymptotic property of the proposed model and evaluate its statistical performance through computer simulation. To demonstrate its practical application, the new model is used to analyze publicly available GTEx data, leading to the identification of hypertension-related GRN for tissue coronary artery. The biological relevance of this model is also supported by enrichment analysis showing that key genes constituting these GRNs play pivotal roles in shaping, molecular, cellular and developmental processes.

email: chencxxy@psu.edu

ABSTRACTS & POSTER PRESENTATIONS

Examining the Regulatory Use of Machine Learning for Drug Safety Studies

Jae Joon Song*, U.S. Food and Drug Administration
Hana Lee, U.S. Food and Drug Administration
Tae Hyun Jung, U.S. Food and Drug Administration

Propensity score methods have become a popular analytic choice for postmarket drug safety studies utilizing large observational databases. Propensity score, defined as the conditional probability of receiving treatment given covariates, are used to create balanced covariate distributions between non-equivalent drug user groups, to account for confounding bias that may arise from a non-randomized design. While logistic regression models are commonly used to estimate propensity scores in regulatory settings, there is a growing interest in using machine learning algorithms such as random forest, bagging, boosting, or neural network for propensity score estimation. In this study, we assess the utility of machine learning algorithms in propensity score estimation for drug safety studies. We also introduce propensity, an R package to facilitate user-defined simulation studies to evaluate performance of propensity score methods.

email: jaejoon.song@fda.hhs.gov

Mixture Proportion Estimation in Positive-Unlabeled Learning

James Patrick Long*, University of Texas MD Anderson Cancer Center
Zhenfeng Lin, Microsoft

Positive-unlabeled (PU) learning considers two samples, a positive set P with observations from only one class and an unlabeled set U with observations from two classes. The goal is to classify observations in U. Class mixture proportion estimation (MPE) in U is a key step in PU learning. We propose reducing the problem to one dimension via construction of a probabilistic classifier trained on the P and U data sets followed by application of a one-dimensional mixture proportion method from the multiple testing literature to the observation class probabilities. The flexibility of this framework lies in the freedom to choose the classifier and the one-dimensional MPE method. We prove consistency of two mixture proportion estimators using bounds from empirical process theory, develop tuning parameter free implementations, and demonstrate that they have competitive performance on simulated waveform data and a protein signaling problem.

email: jplong@mdanderson.org

Unsupervised Learning of Disease Heterogeneity and Patient Subgroups using Diagnosis Codes in Electronic Medical Records

Yaomin Xu*, Vanderbilt University Medical Center

Unsupervised learning of large-scale EMR data could provide insight into disease heterogeneity and help identify new disease subtypes. The ICD codes are the most commonly used categorization of diseases routinely recorded in EMR for classifying diagnoses and describing patient visits. In this talk, I will present a network-based community detection approach for unsupervised learning of the topological structure of patients based on their shared ICD co-occurrence patterns recorded in EMR. We aimed at building a highly robust approach when applied to real world data. We pursued this by: (1) We estimated a consensus graph based on an ensemble of stochastic block model estimations according to bipartite, patient-ICD relationships; (2) We constructed a hierarchical topological structure of the consensus graph using a top-down recursive partitioning. I will demonstrate a functional interpretation of our approach by applying to a genetic study of a cancer driver mutation and illustrate the findings that recapitulate the existing knowledge as well as those are potentially novel.

email: yaomin.xu@vanderbilt.edu

Deep Learning for Cell Painting Image Analysis

Yuting Xu*, Merck & Co., Inc.
Andy Liaw, Merck & Co., Inc.
Shubing Wang, Merck & Co., Inc.

Cell painting is a high-content image-based assay for morphological profiling, which provides high-throughput multi-channel microscopy images of cells. It is a novel imaging modality that enables the detection of subtle differences in phenotypes. The rich profiles of cell populations have many applications in drug discovery, such as elucidating the compound mechanism of action, characterizing cellular heterogeneity, and identifying disease-associated phenotypes. Developing computational methods to efficiently analyze large scale cell painting image data is in critical need. We proposed several deep learning based models for cell segmentation, feature extraction and classification. The methods are validated and compared to the Cellprofiler software through a large dataset with multiple cell lines. Results suggest that the proposed deep learning image analysis pipeline achieves higher classification accuracy and extracts informative high-level features for characterizing different cell lines.

email: yuting.xu@merck.com

ABSTRACTS & POSTER PRESENTATIONS

Model Building Methods in Machine Learning for Clinical Outcome Prediction

Jarcy Zee*, Arbor Research Collaborative for Health
Qian Liu, Arbor Research Collaborative for Health
Laura H. Mariani, University of Michigan
Abigail R. Smith, Arbor Research Collaborative for Health

Machine learning (ML) methods are useful tools to identify novel biomarkers and predict clinical outcomes, but model building procedures may be underutilized. We used ridge regression and random forest models to predict two clinical time-to-event outcomes using data from the Nephrotic Syndrome Study Network (NEPTUNE), a prospective cohort study of glomerular disease patients. We used a split-sample internal validation approach. Models were developed in the training set (70%) using 56 demographic and clinical characteristics, with and without pre-specification of functional form for continuous variables. Cross-validation was used to tune model parameters, and prediction discrimination in the final models was estimated in the validation set (30%) using integrated area under the curve (iAUC). For ML methods assuming linear associations, like ridge regression, pre-specifying covariate functional forms was important for predictive accuracy and for identifying known risk factors amongst the strongest predictors. A systematic method for identifying non-linear associations is proposed and tested using simulation studies.

email: Jarcy.Zee@arborresearch.org

53. TIME SERIES AND RECURRENT EVENT DATA

Integer-Valued Autoregressive Process with Flexible Marginal and Innovation Distributions

Matheus Bartolo Guerrero*, King Abdullah University of Science and Technology
Wagner Barreto-Souza, Universidade Federal de Minas Gerais
Hernando Ombao, King Abdullah University of Science and Technology

INAR processes are usually defined by specifying the innovations and the operator, implying difficulties to get marginal properties of the process. Moreover, in many practical situations, it is hard to justify the operator in use, which may present modeling limitations, e.g., the usual thinning operator is inadequate to model phenomena where population elements generate many offspring. To overcome drawbacks of the current models, we propose a new approach to build an INAR model. We pre-specify the marginal and innovation distributions; hence, the operator is a consequence rather than an imposition. We explore a new INAR model with both marginal and innovations geometric distributed, a direct alternative to the classical Poisson INAR model. Our process has interesting stochastic properties such as an MA(infty) representation and time-reversibility. Also, we have closed form of transition probabilities h-steps ahead, allowing coherent forecasting. We derive estimators and establish their asymptotic properties. We provide an application to a real dataset of counts of skin lesions in bovines, addressing a comparison of our approach with existing INAR and INGARCH models.

email: matheus.bartologuerrero@kaust.edu.sa

Analysis of N-of-1 Trials Using Bayesian Distributed Lag Model with AR(p) Error

Ziwei Liao*, Columbia University
Ying Kuen Cheung, Columbia University
Ian Kronish, Columbia University
Karina Davidson, Feinstein Institute for Medical Research

A goal of personalized or precision medicine is to identify the best treatment at individual level based on observed personal characteristics. Due to the existence of heterogeneity treatment effect, optimal selection of individual intervention may differ from that shown to be effective on average. N-of-1 trials consider one single subject as the whole trial, in which multiple-period crossover stages are performed within a single individual. Existing statistical analytical methods used in N-of-1 trials include nonparametric test, mixed effect model and autoregressive model. These methods may fail to handle measurements autocorrelation both from consecutive interventions and errors. Distributed lag model is a state-of-the-art regression model that uses lagged predictors. In this article, we propose a Bayesian distributed lag model with autocorrelated errors (BDLAR) that integrate prior knowledge on the shape of lagged coefficients and explicitly model the magnitude and length of carryover effect. We give real data examples to illustrate our method and simulation study was conducted to compare the performance of our proposed BDLAR model with other methods.

email: ziwei.liao.fdu@gmail.com

An Estimating Equation Approach for Recurrent Event Models with Non-Parametric Frailties

Lili Wang*, University of Michigan
Douglas E. Schaubel, University of Pennslyvania

Recurrent event data are frequently modeled using shared frailties to account for the association within clusters, and using correlated multivariate frailties to estimate different event types jointly. For computational and interpretational convenience, the frailties are commonly assumed to follow some well-established distributions like gamma, log-normal, and positive-stable, etc. This distributional assumption is in general arbitrary, and can hardly be verified statistically. We propose a class of semiparametric frailties model which does not limit its frailties to any known distribution. The estimating procedure is based on the estimating equations derived from its first and second-moment conditions. Extensive simulations have shown that the proposed approach can accurately estimate the regression parameters, baseline rates, and variance components, with relatively fast computation time. Moreover, our method can accommodate both shared and correlated frailty models. We illustrate the methods through an analysis of hospitalizations among end-stage renal disease patients.

email: lilywang@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Shape-Preserving Prediction for Stationary Functional Time Series

Shuhao Jiao*, King Abdullah University of Science and Technology
Hernando Ombao, King Abdullah University of Science and Technology

This work presents a novel method for prediction of stationary functional time series, for trajectories sharing a similar pattern with phase variability. Existing prediction methodologies for functional time series only consider amplitude variability. To overcome this limitation, we develop a prediction method that incorporates phase variability. One major advantage of our proposed method is the ability to preserve pattern by treating functional trajectories as shape objects defined in a quotient space with respect to time warping and jointly modeling and estimating amplitude and phase variability. Moreover, the method does not involve unnatural transformations and can be easily implemented using existing software. The asymptotic properties of the least-squares estimator are studied. The effectiveness of the proposed method is illustrated in simulation study and real data analysis on annual sea surface temperatures. It is shown that prediction by the proposed SP (shape-preserving) method captures the common pattern better than the existing prediction method, while providing competitive prediction accuracy.

email: shjiaoqd@163.com

A Class of Dynamic Additive-Multiplicative Models for Recurrent Event Data

Russell S. Stocker*, Indiana University of Pennsylvania

We propose a class of dynamic additive-multiplicative models for recurrent event data that use an effective age process to account for the impact of interventions applied to units after an event occurrence. Estimators are derived and their asymptotic properties are established using results from empirical process theory. Finite sample properties are investigated via a computer simulation study. A real data set is analyzed to illustrate the class of models.

email: rstocker@iup.edu

Causal Dependence between Multivariate Time Series

Yuan Wang*, Washington State University
Louis Scharf, Colorado State University

In this work, we are interested in assessing causal relation among the multivariate time series. Granger causality has been widely used for understanding the inter-dependence between time series or stochastic processes. Broadly speaking, a time series y is causally dependent on another time series x if the past of x contains unique information about y. The uniqueness is with respect to the whole universe and hard to be assessed. In practice, causality can be evaluated with respect to the available information. This causality is often assessed using the multivariate regression model where the predictors are available observables at the current time point. In this work, we will design define a third time series of prima facie evidence in such a way that the question of causality can be resolved from an analysis of partial coherence. Comparing to the conventional methods of assessing causality by comparing two regression models, we will show that partial correlation provides a natural integrated measure of causality.

email: yuan.wang.stat@gmail.com

---

54. MASSIVE DATA: A GIANT PROBLEM?

Irreproducibility in Large-Scale Drug Sensitivity Data

Zoe L. Rehnberg*, University of Michigan
Johann A. Gagnon-Bartsch, University of Michigan

Following the release of several large-scale pharmacogenomic studies, the consistency of high-throughput drug sensitivity data across experiments has been widely discussed. While gene expression data are well replicated, only varying levels of moderate to poor concordance has been found for drug sensitivity measures (half-maximal inhibitory concentration [IC50] and area under the dose-response curve [AUC]) in multiple large databases. In this work, we take advantage of detailed raw data to identify factors ranging from data collection to data analysis contributing to the lack of reproducibility in drug sensitivity studies. We find that many different forms of measurement error and the diversity of biological relationships between cell lines and compounds cause difficulties in reliably summarizing drug efficacy. Additionally, we develop a new method of normalizing raw drug response data that accounts for the presence of measurement error and improves agreement between replicates.

email: zrehnber@umich.edu

A New Integrated Marked Point Process Approach to Analyze Highly Multiplexed Cellular Imaging Data

Coleman R. Harris*, Vanderbilt University Medical Center
Qi Liu, Vanderbilt University Medical Center
Eliot McKinley, Vanderbilt University Medical Center
Joseph Roland, Vanderbilt University Medical Center
Ken Lau, Vanderbilt University Medical Center
Robert Coffey, Vanderbilt University Medical Center
Simon Vandekar, Vanderbilt University Medical Center

There is increasing interest in using high-dimensional multiplexed imaging methods to quantify the heterogeneity of cell populations in healthy and tumor tissue to understand tumor progression and develop improved treatment strategies. After preprocessing steps, these imaging data yield detailed spatial information about tissues at the cellular level. However, current single cell analysis methods rely on clustering and dimension reduction techniques that do not take advantage of the rich spatial information in multiplexed data. Here, we use a two-stage analysis approach that leverages marked point process theory and distance matrix-based methods to study spatial relationships between cell types and marker intensity. We demonstrate the methods by analyzing cells segmented in multiplexed immunofluorescence (MxIF) images of mouse intestine to study how cell architecture changes over tumor development. The resulting inference includes information about the spatial relationships between cell types and immune marker intensity, and may offer improved understanding of the tissue architecture related to tumors.

email: coleman.r.harris@vanderbilt.edu

ABSTRACTS & POSTER PRESENTATIONS

Comparison of Methods to Analyze Clustered Time-to-Event Data with Competing Risks

Yuxuan Wang*, Yale Center for Analytical Sciences
Guanqun Meng, Yale Center for Analytical Sciences
Wenhan Lu, Yale Center for Analytical Sciences
Zehua Pan, Yale Center for Analytical Sciences
Can Meng, Yale Center for Analytical Sciences
Erich Greene, Yale Center for Analytical Sciences
Peter Peduzzi, Yale Center for Analytical Sciences
Denise Esserman, Yale Center for Analytical Sciences

With growing use of pragmatic clinical trials to increase generalizability of results, reduce costs, and emulate real world scenarios, the complexity of their design and analysis has increased. We explored the impact of these design complexities on model convergence, bias and coverage for a cluster randomized clinical trial with a time-to-event outcome and a competing risk of death. We conducted simulations using two methods to generate the clustered data. We varied the event rate, competing risk rate, censoring rate, and amount of clustering. We investigated multiple models including: generalized linear model with a Poisson link function and its extension to allow for clustering; Cox proportional hazards model and its extension to allow for clustering via a sandwich variance estimator; Fine and Grey competing risk model and its extension to allow for clustering (developed by Zhou et al); and a multistate model and its extension to allow for robust variance estimation. We found that only under extreme scenarios did we encounter model convergence issues and the models that ignored clustering had worse bias and coverage compared to those that ignored the competing risk.

email: yuxuan.wang@yale.edu

False Discovery Rates for Second-Generation p-Values in Large-Scale Inference

Valerie Welty*, Vanderbilt University
Jeffrey Blume, Vanderbilt University

The second-generation p-value (SGPV) is a novel alternative to the p-value that accounts for scientific relevance by using a composite null hypothesis to represent null and scientifically trivial effects. A SGPV indicates when the results of a study are compatible with alternative hypotheses (SGPV = 0), null hypotheses (SPGV = 1), or are inconclusive (0 < SPGV < 1). It addresses many of the traditional p-value's undesirable properties, and it is generally easier to interpret as a simple summary indicator. False discovery rates (FDRs) for traditional p-values are well established, and they provide an important assessment of how likely it is that the observed results are mistaken. In this talk, we derive FDRs for SGPVs focusing on the positive false discovery rate (pFDR). Ranking findings on the combination of SGPVs and their estimated pFDR emphasizes precisely estimated clinically meaningful effects, and eschews clinically uninteresting effects that are often captured as statistically significant by traditional methods. We will illustrate our methods with a large-scale example and demonstrate an R package for computing these quantities.

email: valerie.welty@vanderbilt.edu

Drives of Inpatient Readmissions: Insights from Analysis of National Inpatient Database

Haileab Hilafu*, University of Tennessee
Bogdan Bichescu, University of Tennessee

Recent Medicare initiatives linking reimbursements to readmission rates have placed renewed pressure on hospitals to reduce patient readmissions. Using the national readmissions database (NRD), we study the drives of inpatient readmissions. We also develop predicitive models to estimate readmission risk using features for the patient, doctor and hospital, among others. Our study focuses on patients whose primary diagnoses on their index visit to the hospital is wither Heart Failure (HR), Acute Myocardial Infarction (AMI), or Pneumonia (PN).

email: hhilafu@utk.edu

Large Scale Hypothesis Testing with Reduced Variance of the False Discovery Proportion

Olivier Thas*, I-BioStat, Data Science Institute, Hasselt University, Belgium, Ghent University, Belgium and University of Wollongong, Australia
Stijn Hawinkel, Ghent University, Belgium
Luc Bijnens, Janssen Pharmaceuticals

Large scale hypothesis testing aim to control the false discovery rate (FDR), which is the average of the false discovery proportion (FDP) over repeated experiments. Although many methods succeed in controlling the FDR, the FDP variance is often neglected. From a practical perspective it is desirable to have most of the FDP close to the FDR. When the test statistics show strong correlations, it is known that the variance of the FDP is large. We have developed a generic method for controlling the FDR, while reducing the FDP variance. Our method relies on empirical Bayes and on (re)sampling methods. We first properly define the collapsed distribution, which is often referred to as the marginal distribution of the test statistics, and we propose an improved estimator of this distribution. In a simulation study we demonstrate the superiority of our method. Our motivation came from testing for differential abundance in microbiome studies for which the absolute abundance is estimated via flow cytometry. This is a setting which, by design, has strong correlations between test statistics and which therefore can benefit from our procedure.

email: olivier.thas@ugent.be

ABSTRACTS & POSTER PRESENTATIONS

55. HUMAN MICROBIOME STUDIES: NOVEL METHODS AND NEW STUDIES

A Novel Method for Compositional Analysis of the Microbiome Data

Yijuan Hu*, Emory University

Microbiome data are compositional in that the total number of sequencing reads is arbitrarily imposed by the instrument and only the relative abundance of OTUs can be measured. Acknowledging the compositional nature of the microbiome data is important, especially for detecting OTUs that responded to the condition change, not OTUs whose relative abundance changed merely due to the compositional constraint. A common practice of existing compositional analyses is to take a log-ratio transformation of the raw read count data. Because a zero count cannot be log-transformed, people typically add a pseudo count such as 1 or 0.5 to zero counts. However, the zero counts are very prevalent (more than 50%) in the microbiome data and any choice of a pseudo count can have a great impact on the results. We propose a new method that is based on generalized estimating equations and taking the log-ratio transformation on the true, unknown relative abundances and thus avoid the decision of a pseudo count. We show that we can detect differentially abundant OTUs while controlling for false discovery rate.

email: yijuan.hu@emory.edu

Estimating the Overall Contribution of Human Oral Microbiome to the Risk of Developing Cancers Based on Prospective Studies

Jianxin Shi*, National Cancer Institute, National Institutes of Health

The human microbiome is the collection of microbes inhabiting the human body. Advances in high-throughput sequencing allow characterization of the compositions of human microbial communities and make it possible to perform large-scale epidemiological studies. We have recently performed a large- scale case-cohort study with the goal to identify oral microbiome risk factors for a variety of cancers and to improve cancer risk prediction. One key step is to estimate the overall contribution of oral microbiome to the risk of developing cancers. It is also interesting to estimate the effect size distribution of microbiome features for association analysis that would help design and predict the yield of future studies. We develop two statistical methods for these purposes, one based on a linear mixed model assuming log(time-to-event) following a normal distribution and the other based on a high-dimensional Cox proportional hazard model. Preliminary results and implications analyzing the case-cohort oral microbiomestudy will be discussed.

email: jianxin.shi@nih.gov

Multi-Group Analysis of Compositions of Microbiomes with Bias Correction (MANCOM-BC)

Shyamal D. Peddada*, University of Pittsburgh
Huang Lin, University of Pittsburgh

In many applications, researchers are interested in comparing the microbial compositions of more than two study groups or ecosystems. For example, a researcher may be interested in determining the mean temporal changes in the abundance of individual taxa in infant gut. Researchers are also interested in clustering taxa with similar dynamic patterns of abundance. In this talk we generalize ANCOM-BC (Lin and Peddada, 2019 manuscript) methodology to answer such questions while attempting to control the false discovery rates. In some instances, researchers are interested in performing multiple pairwise directional tests to determine if the abundance of a taxon decreased or increased between two ecosystems. We shall extend ANCOM-BC to perform such multiple pairwise directional comparisons for multiple taxa (i.e. multiple comparisons with multiple testing). The resulting methodology is aimed to control mixed directional FDR (mdFDR). We illustrate our methodology using the well-known global gut data of Yatsunenko et al. (2013).

email: sdp47@pitt.edu

A Powerful Microbial Group Association Test Based on the Higher Criticism Analysis for Sparse Microbial Association Signals

Ni Zhao*, Johns Hopkins University
Hyunwook Koh, Johns Hopkins University

A variety of microbial association tests have been proposed in the last few years to evaluate the impact a microbiome group (e.g., community or clade) on a phenotype. The existing microbial group association tests generally fall in the class of aggregation tests which amplify the overall group association by aggregating underlying association signals; as such, they are powerful when many of the nested microbial biomarkers (e.g., OTUs) are associated (i.e., low sparsity). However, in practice, the association signals can be highly sparse, for which the existing approaches lose power substantially. In this paper, we introduce a powerful and data-driven optimal association test for sparse microbial association signals based on the higher criticism analysis framework: microbiome higher criticism analysis (MiHC). Our extensive simulations and real data analysis illustrate that MiHC robustly maintains a high statistical power for different phylogenetic relevance and sparsity levels with correct type I error controls. The high adaptivity of MiHC is attractive especially because the extent of phylogenetic relevance and sparsity is usually known in practice.

email: nzhao10@jhu.edu

ABSTRACTS & POSTER PRESENTATIONS

56. BAYESIAN APPROACHES FOR COMPLEX INNOVATIVE CLINICAL TRIAL DESIGN

Bayesian Clinical Trial Design using Historical Data that Inform the Treatment Effect

Joseph G. Ibrahim*, University of North Carolina, Chapel Hill
Matthew A. Psioda, University of North Carolina, Chapel Hill

We consider the problem of Bayesian sample size determination for a clinical trial in the presence of historical data that inform the treatment effect. Our broadly applicable, simulation-based methodology provides a framework for calibrating the informativeness of a prior while simultaneously identifying the minimum sample size required for a new trial such that the overall design has appropriate power to detect a non-null treatment effect and reasonable type I error control. We develop a comprehensive strategy for eliciting null and alternative sampling prior distributions which are used to define Bayesian generalizations of the traditional notions of type I error control and power. We develop a procedure for generating an appropriately sized Bayesian hypothesis test using a simple partial-borrowing power prior which summarizes the fraction of information borrowed from the historical trial. We present results from simulation studies that demonstrate that a hypothesis test procedure based on this simple power prior is as efficient as those based on more complicated meta-analytic priors. A real dataset in melanoma also examined.

email: ibrahim@bios.unc.edu

Advanced Hierarchical Modeling in Clinical Trials

Kert Viele*, Berry Consultants

Hierarchical Models are one of the workhorses of Bayesian statistics. Within clinical trials, hierarchical models feature in the analysis of basket trials, historical borrowing of information, and smoothing methods such as a normal dynamic linear modeling. When using a hierarchical model, care must be taken in the selection of the hyper prior, allowing for a sufficient amount of borrowing that the methodology gains efficiency, but not so much that the prior acts as a strong hidden assumption (for example implicitly assuming historical clinical trial data has virtually the same parameter as current data). We will discuss recent clinical trial case studies where the calibration of the hierarchical model prior distributions were a fundamental part of the modeling and regulatory review process, as well as discuss recent advances in clustering based methods such as EXNEX, Multi source exchangeability models, and full mixtures of Dirichlet Processes.

email: kert@berryconsultants.net

Bayesian Sequential Monitoring of Clinical Trials

Matthew Austin Psioda*, University of North Carolina, Chapel Hill
Evan Kwiatkowski, University of North Carolina, Chapel Hill
Mat Soukup, U.S. Food and Drug Administration
Eugenio Andraca-Carrera, U.S. Food and Drug Administration

We provide an overview Bayesian sequential monitoring of clinical trials. In such trials, patients are continually enrolled and their data are analyzed as often as is desired/feasible until a hypothesis has been proven or disproven, or until allocated resources for the trial have been exhausted. Such trials does not require a pre-specified sample size or number of analyses. For proving efficacy, the Bayesian collects data until evidence in favor of the investigational treatment is substantial from the perspective of an a priori skeptical judge who doubts treatment efficacy. We also discuss approaches for futility monitoring. This includes stopping enrollment when the posterior probability of treatment efficacy is sufficiently low from the perspective of an a priori enthusiastic judge or when the trial is unlikely to be conclusive based on the predictive distribution of future data given data that has been observed thus far. We propose a framework for elicitation of skeptical and enthusiastic priors that can be applied consistently across applications areas, an appealing property for regulatory bodies who must be concerned with fair and consistent decision making practices.

email: matt_psioda@unc.edu

Bayesian Clinical Trial Designs using SAS

Fang Chen*, SAS Institute Inc.
Guanghan Frank Liu, Merck & Co. Inc.

Although the Bayesian paradigm offers great potential to clinical trial designs as incorporating prior information can lead to lower sample sizes, higher efficiency, and more reliable decisions, the availability of software programs is one of the obstacles that provide easy access to run the design, simulations, and analysis. This talk provides a high-level overview of Bayesian capabilities in SAS/STAT software - the general modeling MCMC procedure and the specialized BGLIMM procedure for Bayesian generalized mixed models - and presents design examples using the software. Application examples include designs using historical data (with Bayesian versions of type I error rate control), dose-finding designs (using pharmacokinetics), and adaptive basket trial designs. We demonstrate that, at a practical level, SAS software provides comprehensive coverage that meets the demand for complex Bayesian designs in real world.

email: FangK.Chen@sas.com

ABSTRACTS & POSTER PRESENTATIONS

57. ACHIEVING REAL-WORLD EVIDENCE FROM REAL-WORLD DATA: RECENT DEVELOPMENTS AND CHALLENGES

Real-World Data and Analytics for Regulatory Decision-Making: FDA/CDRH Experience

Lilly Yue*, U.S. Food and Drug Administration

In medical product development, there has been an increased interest in utilizing real-world data which have become abundant owing to advances in biomedical science, information technology and engineering. High-quality real-world data may be utilized to generate scientific evidence for regulatory or healthcare decision-making using proven analytical methods and techniques, such as propensity score-based methodology. This presentation will focus on practice and perspectives on transforming RWD to scientific evidence in regulatory decision-making for medical devices, illustrated with examples from our pre-market reviews.

email: lilly.yue@fda.hhs.gov

RWD, EHRs, PROs; Using Data to Inform the Patient Trajectory and Experience

Warren A. Kibbe*, Duke University

Now that electronic health records are in use across organizations providing healthcare, how do we more explicitly model patients' individual trajectories and use them to both inform the patient, alert the care team when to intervene, and use these data to understand how changes in care impact patient outcomes? I will present a few examples of these activities and how they are beginning to impact patient care at an institutional level, and one possible path forward for scaling across organizations.

email: warren.kibbe@duke.edu

Addressing Confounding in Real-World Evidence using Propensity Scores

John D. Seeger*, Optum

As real-world evidence acquires greater prominence as a basis for inference about therapeutic interventions, a reminder of threats to the validity of these inferences is needed. In routine care as reflected by real-world data, treatments are not allocated to patients at random, but rather selected for reasons that may be observed or unobserved. Medical practitioners seek to enhance the ratio of benefits to risks when selecting a treatment for a particular patient, and this leads to prescribing medication to patients who are both likely to receive benefit and unlikely to experience harm. While this selective prescribing is good for patients, it leads to confounding when comparing one treatment to another. Propensity score methods can address such confounding, and they are particularly well-suited to situations where the selection (and confounding) results from several variables simultaneously. However, there remains a requirement for appropriate use of propensity scores and attention to their underlying assumptions. This presentation will explore suitable application of propensity score methods with an aim to identify and promote good practices.

email: john.seeger@optum.com

58. NOVEL SPATIAL MODELING APPROACHES FOR AIR POLLUTION EXPOSURE ASSESSMENT

Spatiotemporal Data Fusion Model for Air Pollutants in the Near-Road Environment using Mobile Measurements and Dispersion Model Output

Owais Gilani*, Bucknell University
Veronica J. Berrocal, University of California, Irvine
Stuart A. Batterman, University of Michigan

Concentrations of near-road air pollutants (NRAPs) have increased to very high levels in many urban centers. Adverse health effects of exposure to NRAPs are greater when the exposure occurs in the near-road environment (NRE). Therefore, there is increasing interest in monitoring pollutant concentrations in the NRE. However, due to practical limitations, monitoring pollutant concentrations near roadways is difficult and expensive. Alternatively, deterministic models that provide predictions of pollutant concentrations in the NRE, such as the Research Line-source dispersion model (RLINE), have been developed. A common feature of these models is that their outputs typically display systematic biases and need to be calibrated in space and time using observed pollutant data. We present a non-stationary Bayesian data fusion model that uses novel data on monitored pollutant concentrations (nitrogen oxides and fine particulate matter) in the NRE and, combining it with the RLINE model output, provides predictions at unsampled locations. The model can also be used to evaluate whether including the RLINE model output leads to improved pollutant concentration predictions.

email: owais.gilani@bucknell.edu

ABSTRACTS & POSTER PRESENTATIONS

Multi-Resolution Data Fusion of Air Quality Model Outputs for Improved Air Pollution Exposure Assessment: An Application to PM2.5

Veronica J. Berrocal*, University of California, Irvine

Air quality models are deterministic, numerical models that can generate estimates of air pollutants concentration at predetermined spatial and temporal resolution. Besides their use for regulatory purposes, air quality models have been shown to provide useful information on air pollutants' concentrations that can be leveraged to yield improved estimates of ambient air pollution exposure. However, due to calibration issues and to the spatial mismatch in spatial resolution from grids to point level, air quality model outputs are typically statistically postprocessed. In this paper we propose a Bayesian hierarchical spatio-temporal modeling framework that aims to yield more reliable estimate of PM2.5 ambient exposure by simultaneously integrating multiple outputs of an air quality model run at nested resolutions. The model combines the multiple model outputs by regressing the observed PM2.5 concentration on a new regressor defined through a weighted average of the multiple air quality model outputs with weights depending on the resolution of the output, and defined to be spatially and temporally varying.

email: berrocal@umich.edu

Multivariate Spectral Downscaling for PM2.5 Species

Yawen Guan*, University of Nebraska, Lincoln
Brian Reich, North Carolina State University
James Mulholland, Georgia Institute of Technology
Howard Chang, Emory University

Fine particulate matter (PM2.5) is a mixture of air pollutants that has adverse effects on human health. Understanding the health effects of PM2.5 mixture and its individual species has been a research priority over the past two decades. However, the limited availability of speciated PM2.5 measurements continues to be a major challenge in exposure assessment for conducting large-scale population-based epidemiology studies. The PM2.5 species have complex spatial-temporal and cross dependence structures that should be accounted for in estimating the spatiotemporal distribution of each component. Two major sources of air quality data are commonly used for deriving exposure estimates: point-level monitoring data and gridded numerical computer model simulation, such as the Community Multiscale Air Quality (CMAQ) model. We propose a statistical method to combine these two data sources for estimating speciated PM2.5 concentration. Our method models the complex relationships between monitoring measurements and the numerical model output at different spatial resolutions, and we model the spatial dependence and cross dependence among PM2.5 species.

email: yguan12@unl.edu

Functional Regression for Predicting Air Pollution Concentrations from Spatially Misaligned Data

Meredith Franklin*, University of Southern California
Khang Chau, University of Southern California

Satellite observations have become an instrumental component of air pollution exposure estimation, providing critical spatial and temporal information beyond what can be provided from ground based monitoring networks. Relating satellite observations to ground-level concentrations requires dealing with missing observations, and predictions are generally improved by incorporating external data such as meteorology and land use. However, these multi-source data are all referenced on different spatial scales and supports (areal, point). A functional distance-based weighing approach provides the flexibility to combine spatially misaligned data into a single prediction framework without having to artificially align data to a common reference frame. We demonstrate its applicability in estimating PM2.5 from 1 km satellite observations of aerosol optical depth (AOD), supplemented with meteorology and geographic features.

email: meredith.franklin@usc.edu

---

59. INNOVATIONS IN TWO PHASE SAMPLING DESIGNS WITH APPLICATIONS TO EHR DATA

Optimal and Nearly-Optimal Designs for Studies with Measurement Errors

Gustavo G. C. Amorim*, Vanderbilt University Medical Center
Bryan E. Shepherd, Vanderbilt University Medical Center
Ran Tao, Vanderbilt University Medical Center
Sarah C. Lotspeich, Vanderbilt University Medical Center
Pamela A. Shaw, University of Pennsylvania
Thomas Lumley, University of Auckland

Measurement errors are present in most, if not all, data collection procedures. Severe or even mild errors can harm analyses by biasing estimates of interest. For valid conclusions, we often rely on sampling a small proportion of the data for validation and perform analyses on this validated sample only or using a 2-phase procedure that also includes the error-prone variables. Most studies in the measurement error literature focus on estimation and ways to select the validation sample have usually been ignored. We propose optimal and nearly-optimal designs for selecting this validation sample that lead to more efficient estimates of the parameter of interest when compared to a simple random sampling strategy, which is often applied in practice. We focus on continuous outcomes and allow both predictors and outcomes to be error-prone, with correlated error mechanisms. We discuss sampling designs that target maximum likelihood as well as design based estimators, and show via simulations how they compare to each other. Our results suggest that optimal sampling schemes depend on the analysis method, either likelihood or design-based, and can differ substantially.

email: gustavo.g.amorim@vumc.org

ABSTRACTS & POSTER PRESENTATIONS

The Mean Score and Efficient Two-Phase Sampling for Discrete-Time Survival Models with Error Prone Exposures

Kyunghee Han*, University of Pennsylvania
Thomas Lumley, University of Auckland
Bryan E. Shepherd, Vanderbilt University Medical Center
Pamela A. Shaw, University of Pennsylvania

Increasingly medical research is dependent on data collected for non-research purposes, such as electronic health records data (EHR). EHR data and other large database settings can be prone to measurement error in key exposures. Validating a subset of records is a cost-effective way of gaining information on the error structure, which in turn can be used to adjust analyses for this error and improve inference. We extend the mean score method for the two-stage analysis of discrete-time survival models, which uses the unvalidated covariates as auxiliary variables that can act as surrogates for the unobserved true exposure. This method allows for a two-phase sampling analysis approach that preserves the consistency of the regression model estimates in the validated subset, with increased precision leveraged from the auxiliary data. We evaluate efficiency gains of the mean score estimator from the allocated validation design compared to random and balanced sampling with simulation studies. We also apply the proposed method to a real data setting.

email: kyunghee.stat@gmail.com

Two-Phase Designs Involving Incomplete Life History Processes

Richard J. Cook*, University of Waterloo

We consider a two-phase study design in which interest lies in assessing the effect of an expensive exposure variable on a time to event response. The phase I sample is comprised of individuals with data on a response and inexpensive covariates but here it is constructed by pooling two or more phase I sub-samples, each employing different truncation schemes. Bio-specimens are available for all individuals in the phase I sample and these are to be selected for assay in order to estimate the association between a biomarker of interest and the failure time response. We investigate the efficiency of various selection models for the phase II sub-sample which address the different truncation schemes used in the component phase I sub-samples. A motivating application involves data from a psoriasis registry and a registry of individuals with psoriatic arthritis where the goal is to study genetic markers for the development of psoriatic arthritis among individuals with psoriasis.

email: rjcook@uwaterloo.ca

60. RECENT APPROACHES TO MULTIVARIATE DATA ANALYSIS IN THE HEALTH SCIENCES

A Multivariate Discrete Failure Time Model for the Analysis of Infant Motor Development

Brian Neelon*, Medical University of South Carolina

We develop a multivariate discrete failure time model for the analysis of infant motor development. We use the model to jointly evaluate the time (in months) to achievement of three well-established motor milestones: sitting up, crawling, and walking. The model includes a subject-specific latent factor that reflects underlying heterogeneity in the population and accounts for within-subject dependence across the milestones. The factor loadings and covariate effects are allowed to vary flexibly across milestones, and the milestones are permitted to have unique at-risk intervals corresponding to different developmental windows. We adopt a Bayesian inferential approach and develop a convenient data-augmented Gibbs sampler for posterior computation. We conduct simulation studies to illustrate key features of the model and use the model to analyze data fromthe Nurture study, a birth cohort examining infant health and development during the first year of life.

email: neelon@musc.edu

Incorporating a Bivariate Neighborhood Effect of a Single Neighborhood Identifier in a Hierarchical Model

James O'Malley*, Dartmouth College
Peter James, Harvard T.H. Chan School of Public Health
Todd A. MacKenzie, Dartmouth College
Jinyoung Byun, Dartmouth College
SV Subramanian, Harvard T.H. Chan School of Public Health
Jason B. Block, Harvard Pilgrim Health Care

I describe a situation in which the latent or random effect of a neighborhood should have a bivariate impact on an individuals' outcome. Standard statistical software for hierarchical models does not easily allow for correlation between the components of distinct random effects. To overcome this deficiency, I develop a Bayesian model and accompanying estimation procedure that facilitates correlated bivariate neighborhood effects. I apply the model to the motivating Framingham Heart Study (FHS) linked food establishment data to examine whether proximity to fast-food establishments is associated with Body Mass Index (BMI). In this setting, individuals may reside or work in multiple neighborhoods, individuals may have cross-sectional and longitudinal heterogeneous effects, and there may be serial correlation between repeated observations over time. Simulation studies that vary key model parameters evaluate how well each aspect of the model is identified by the data. Comparisons of the full model to models with restricted versions of the covariance structure illustrate the impact of including each feature of the covariance structure.

email: James.OMalley@Dartmouth.edu

ABSTRACTS & POSTER PRESENTATIONS

A Statistical Framework for the Compositional Data Analysis to Investigate Molecular Mechanisms Associated with Cancer Immunotherapy

Dongjun Chung*, Medical University of South Carolina
Brian Neelon, Medical University of South Carolina

During the last decade, there have been tremendous achievements in cancer immunotherapy. Among those, immune checkpoint blockades, such as Anti-PD1, have completely changed the therapeutic approaches for many types of cancer. However, a significant heterogeneity in the efficacy of these immune checkpoint blockades has been reported and the molecular basis related to such differences still remains to be investigated. Compositional data analysis has been less studied in multivariate analysis literature but recently received significant attention with the emergence of compositional big data such as cancer immune cell composition and microbiome data. In this presentation, I will discuss our recent work on a Bayesian regression framework for the compositional data analysis. This approach allows us to consider correlation among compositional outcomes, identify key covariates associated with the compositional outcomes, and utilize prior biological knowledge. I will illustrate the proposed statistical model with simulation studies and its application to the immune-genomic data of the Immune Landscape of Cancer.

email: chungd@musc.edu

On Nonparametric Estimation of Causal Networks with Additive Faithfulness

Kuang-Yao Lee*, Temple University
Tianqi Liu, Google
Bing Li, The Pennsylvania State University
Hongyu Zhao, Yale University

In this work we propose an additively faithful directed acyclic graph (AFDAG) for causal learning from observational data. Our approach is based on additive conditional independence (ACI), a newly proposed three-way statistical relation that shares many similarities with conditional independence but without resorting to multi-dimensional kernels. This distinct feature strikes a balance between a parametric model and a fully nonparametric model, which makes the proposed model attractive for handling large networks. We develop an estimator for AFDAG based on a linear operator that characterizes ACI, and establish the consistency and convergence rates of this estimator, as well as the uniform consistency of the estimated DAG. Moreover, we introduce a modified PC-algorithm to implement the estimating procedure efficiently, so that its complexity is determined by the level of sparseness rather than the dimension of the network. Finally the usefulness of AFDAG formulation is demonstrated through an application to recovering a MAPK pathway.

email: kuangyao.l@gmail.com

61. SPEED POSTERS: IMAGING DATA/SURVIVAL ANALYSIS/SPATIO-TEMPORAL

61a. A Geometric Approach Towards Evaluating fMRI Preprocessing Pipelines

Martin Lindquist*, Johns Hopkins Bloomberg School of Public Health

The preprocessing pipelines typically used in resting-state fMRI (rs-fMRI) analysis are modular in nature, as they are composed of a number of separately developed components performed in a flexible order. We illustrate the shortcomings of this approach, as we introduce a geometrical framework to illustrate how later preprocessing steps can reintroduce artifacts that had previously been removed from the data in a prior step of the pipeline. These issues can arise in practice when any combination of common preprocessing steps such as nuisance regression, scrubbing, CompCor, and temporal filtering are performed in a modular fashion. We illustrate the problem using a few concrete examples and conclude with a general discussion of how different preprocessing steps interact with one another. These results highlight the fact that special care needs to be taken when performing preprocessing on rs-fMRI data, and the need to critically revisit previous work on rs-fMRI data that may not have adequately controlled for these types of effects.

email: mlindqui@jhsph.edu

61b. Non-Parametric Estimation of Spearman's Rank Correlation with Bivariate Survival Data

Svetlana K. Eden*, Vanderbilt University
Chun Li, Case Western Reserve University
Bryan Shepherd, Vanderbilt University

We study rank-based approaches to estimate the correlation between two right-censored variables. With end-of-study censoring, it is often impossible to non-parametrically identify the complete bivariate survival distribution, and therefore it is also impossible to non-parametrically compute Spearman's rank correlation. As a solution, we propose two measures that can be non-parametrically identified and estimate Spearman-like quantities. The first can be thought of as computing Spearman's correlation after assigning the highest rank value to observations censored at the maximum follow-up times. The second is Spearman's correlation in a restricted region where the conditional bivariate distribution can be identified. We describe population parameters for these measures and illustrate how they are similar to and different from Spearman's correlation. We propose consistent estimators of these measures and study their use through simulations. We illustrate our methods with a study assessing the correlation between the time to viral failure and the time to regimen change among persons living with HIV in Latin America who start antiviral therapy.

WITHDRAWN

email: svetlana.eden@vanderbilt.edu

ABSTRACTS & POSTER PRESENTATIONS

61c. Nonparametric Tests for Semi-Competing Risks Data under Markov Illness-Death Model

Jing Li*, Indiana University
Giorgos Bakoyannis, Indiana University
Ying Zhang, University of Nebraska Medical Center
Sujuan Gao, Indiana University

It is generally of interest to explore if the risk of death would be modified by medical conditions (e.g., illness) that occurred prior. This situation is known as semi-competing risks data, a mixture of competing risks and progressive state data. It occurs when a non-terminal event can be censored by a terminal event, but not vice versa. In this work, we adopt a Markov illness-death model under multi-state modeling framework. Numerous research has been done on nonparametric estimation for multi-state models however not on nonparametric testing for transition intensities. Hence, we propose three nonparametric tests including a linear test, a Kolmogorov-Smirnov-type test, and a L2-distance-type test to directly assess whether the status of non-terminal event alters the risk of terminal event. The asymptotic distributions of the proposed test statistics under the null hypothesis are established using empirical process theory. The performance of these tests in finite samples is numerically evaluated under various scenarios through simulation. This research is applied to the Indianapolis-Ibadan Dementia Project to explore whether dementia changes mortality risk.

email: JL204@iu.edu

61d. Parsimonious Covariate Selection for Interval Censored Data

Yi Cui*, State University of New York at Albany
Xiaoxue Gu, North Dakota State University
Bo Ye, State University of New York at Albany

Interval censored outcomes widely arise in many clinical trials and observational studies. In many cases, subjects are only followed-up periodically. As a result, the event of interest is known only to occur within a certain interval. We provided a method to select the parsimonious set of covariates associated with the interval censored outcome. First, the iterative sure independence screening (ISIS) method was applied to all interval censored time points across subjects to simultaneously select a set of potential important covariates; then multiple testing approaches were used to improve the selection accuracy through refining the selection criteria, i.e. determining a refined common cutoff value. We compared the improvement of selection accuracy by using both familywise error rate (FWER) and generalized FWER (gFWER) methods. Our method shows good performance in simultaneously in selecting non-zero effects and deselecting zero-effects, respectively.

email: ycui3@albany.edu

61e. Identifying Amenity Typologies in the Built Environment: A Bayesian Non-Parametric Approach

Adam T. Peterson*, University of Michigan
Veronica Berrocal, University of California, Irvine
Brisa Sánchez, Drexel University

The presence of specific built environment resources such as food vendors or recreation centers may impact diet and physical activity, thereby affecting the subsequent risk of important health conditions like cardiovascular disease. Of particular interest to investigators are types of spatial patterns that describe the availability of a particular resource near a location where subjects frequently spend their time, residential neighborhoods or schools, for example. To describe these types of spatial patterns, we model the spatial distribution of fast food restaurants around each school as a one-dimensional Inhomogeneous Poisson Process (IPP). We cluster the spatially varying intensities of the IPPs using the Nested Dirichlet Process which allows us to group schools into environment types. Our modeling approach is non-parametric in that the number of clusters is not chosen a priori, more clusters can be added in accordance with the best marginal density fit and the shape of the intensity function across distance from the school is data-driven.

email: atpvyc@umich.edu

61f. Estimation of a Buffering Window in Functional Linear Cox Regression Models for Spatially-Defined Environmental Exposure

Jooyoung Lee*, Harvard T.H. Chan School of Public Health
Donna Spiegelman, Yale School of Public Health
Molin Wang, Harvard T.H. Chan School of Public Health

In environmental health research, it is of interest in understanding the effect of the neighborhood environment on health. Typically, neighborhood environmental exposures are measured within radial buffer zones from a residential address and identification of a buffer window is of importance, which is so called “uncertain geographic context” problem. We propose to address geographic uncertainty through developing methods for estimating the buffering window in a functional linear Cox proportional hazard model. The theoretical properties of our proposed method are studied and simulation studies are conducted. The method is illustrated in a study of the effect of walkability on cardiovascular disease in the Nurses’ Health Study.

email: jooyounglee@hsph.harvard.edu

ABSTRACTS & POSTER PRESENTATIONS

61g. An Alternative Sensitivity Analysis for Informative Censoring

Patrick O'Connor*, University of Arizona
Chiu-Hsieh Hsu, University of Arizona
Denise Roe, University of Arizona
Chengcheng Hu, University of Arizona
Jeremy M.G. Taylor, University of Michigan

Most existing survival analysis methods work under the independent censoring assumption. When censoring is informative of event times, those methods will produce biased survival estimates. We propose a sensitivity analysis approach via nonparametric multiple imputation to impute event times for censored observations, in which the sensitivity analysis parameter is based on the magnitude of informative censoring (i.e. the correlation between event and censoring times). Specifically, Kendall's tau measures the correlation between event and censoring times and is used as the sensitivity analysis parameter due to its standardized range and interpretability. For each censored observation, tau will define the size and direction of the imputing risk set based on the at-risk subjects. As tau gets more extreme, the imputing risk set gets smaller. The proposed approach is illustrated using data from an HIV-prevention trial. Simulation show the nonparametric imputation method produces survival estimates similar to the true values with coverage rates similar to the nominal 95% level. These findings hold even for high censoring rates or high levels of informative censoring.

email: patrickaoconnor@email.arizona.edu

61h. Displaying Survival of Patient Groups Defined by Covariate Paths: Extensions of the Kaplan-Meier Estimator

Melissa Jay*, University of Iowa
Rebecca Betensky, New York University

Extensions of the Kaplan-Meier estimator have been developed to illustrate the relationship between a time-varying covariate of interest and survival. In particular, Snapinn et al. and Xu et al. developed estimators to display survival of a hypothetical group of patients who always have a certain time-varying covariate value. These estimators properly handle time-varying covariates, but their interpretation is not always clinically meaningful. It might be of greater clinical interest to compare survival for patients who lie on a covariate path, to identify potential milestones on a patient's course of treatment. We present extensions of Snapinn et al. and Xu et al.'s estimators, providing a crude and covariate-adjusted approximation of the survival function for patient groups defined by covariate paths. We demonstrate the utility of these estimators with a medical example and simulation results.

email: melissa-jay@uiowa.edu

61i. Semiparametric Transformation Model for Clustered Competing Risks Data

Yizeng He*, Medical College of Wisconsin
Soyoung Kim, Medical College of Wisconsin
Lu Mao, University of Wisconsin, Madison
Kwang Woo Ahn, Medical College of Wisconsin

Competing risks outcomes are common in cancer research. When competing risks are present, an event from a competing risk precludes an event of the primary interest from occurring. Such competing risks data often suffer from cluster effects from matched pair design, study center effect, among others. To directly evaluate covariate effects on the cumulative incidence function for clustered competing risks data, the marginal proportional subdistribution hazards model and frailty models are commonly used in practice. However, most work in the current literature is based on inverse probability censoring weighting which requires estimating the censoring distribution. Thus, these methods may yield invalid statistical inference if the censoring distribution is not correctly modeled. To address this limitation, we propose a marginal semiparametric transformation model based on Mao and Lin (2017). The proposed method does not rely on models for the censoring distribution, accommodates non-proportional subdistribution hazards structure, and provides a platform for joint inference of all causes.

email: yizhe@mcw.edu

61j. Partial Linear Single Index Mean Residual Life Models

Peng Jin*, New York University School of Medicine
Mengling Liu, New York University School of Medicine

Mean residual life (MRL) function is an alternative to the hazard function for characterizing the time-to-event data. The MRL can be interpreted as the remaining life expectancy of a subject who has survived to a time point. In biomedical research, the proportional and the additive MRL models have been primarily focused to study the association between risk factors and disease in multiplicative and additive scale. When the risk factors have complex correlation structures or are high-dimensional, simple linear relationship between the risk factors and MRL functions may not be sufficient. The single index model offers flexibility capturing nonlinear covariate effects and provides a solution to reduce dimensionality. In this paper, we propose a partially linear single index MRL model, which contains a nonparametric single index component and a linear component. The polynomial spline technique is implemented to model the nonparametric single index function and we estimate the parameters using an iterative algorithm based on the estimating equations. The finite sample performance is evaluated through numerical simulations. A real data application is presented for illustration.

email: pj691@nyu.edu

ABSTRACTS & POSTER PRESENTATIONS

61k. Evaluating the Diagnostic Accuracy of a New Biomarker for Prostate Cancer: Challenges in Small Samples

Joshua I. Banks*, Thomas Jefferson University
Jungreem Woo, Thomas Jefferson University
Sandra Santasusagna, Thomas Jefferson University
Benjamin Leiby, Thomas Jefferson University
Josep Domingo-Domenech, Thomas Jefferson University

Prostate Cancer (PCa) is a very common cancer in men in the US. The established method of detecting prostate cancer is by the use of the prostate specific antigen (PSA). However, a limitation of PSA is its low specificity. Prior studies have shown that the molecule GATA2 is expressed in PCa, however its complementary use as a biomarker has not been studied. We investigated how GATA2 measured in urine extracellular vesicles may improve the accuracy of PCa diagnosis. We applied standard logistic regression methods to determine how well GATA2 performed as a marker of a positive biopsy as a single marker and in combination with PCA3 and TMPRRSSR2-ERG which are non-clinically established biomarkers. We generated ROC curves as well sensitivity, specificity, negative predictive values, and positive predictive values to compare the predictive ability of various combinations of GATA2 with other biomarkers in this study. We discuss the challenges of evaluating the utility of a biomarker in relatively small training and validation cohorts and potential approaches for addressing them.

email: joshua.banks@jefferson.edu

61l. Identifying Spatio-Temporal Variation in Breast Cancer Incidence Among Different Age Cohorts Using Bayesian Hierarchical Modeling

Amy E. Hahn*, University of Iowa
Jacob Oleson, University of Iowa
Paul Romitti, University of Iowa

We aim to learn more about breast cancer incidence in Iowa by examining annual cohorts based on locations at birth and diagnosis, year of diagnosis, and age at diagnosis. Women ages 20-39 years diagnosed from 1992-2016 were identified from the Iowa Surveillance, Epidemiology, and End Results (SEER) Cancer Registry. Locations were generated using a comprehensive geographic information systems database and geocoded into latitude and longitude coordinates, as opposed to aggregated counts by area. With this level of granularity, we adopt a Bayesian Poisson point process model that maintains the continuous spatial nature of the data for more specific examination of spatial variation in risk. Data on dates of birth and diagnosis allow us to model spatial variation over time. Applying space-time interaction, we can further identify spatio-temporal variation based on these cohorts. Integrated nested Laplace approximation (INLA) is implemented due to the size of the dataset and the number of parameters introduced. By applying a Bayesian hierarchical model, we aim to identify areas of increased risk and changes in risk over space and time.

email: amy-hahn@uiowa.edu

61m. One-to-One Feature Matching with Application to Multi-Level Modeling

David Degras*, University of Massachusetts, Boston

When a study involves multiple nested levels of analysis there is not always an obvious correspondence between the features obtained for each unit at a given level. In functional MRI research for example, brain activity can be described by functional connectivity (FC) graphs that correspond to cognitive or behavioral processes. FC graphs, however, are not always easily comparable across subjects in group-level analyses. It is therefore essential to accurately match features across units of one level (e.g. subjects) before proceeding to the next level of analysis (e.g. groups). To match features across units in a one-to-one fashion, we seek permutations of feature labels that minimize empirical discrepancies between units. We propose an algorithm that combines k-means clustering algorithm with the Hungarian algorithm for bipartite graphs. Theoretical insights on the algorithm's convergence properties are provided. We assess the algorithm in numerical experiments and present an application to the ABIDE database, a large collection of resting-state fMRI datasets from individuals with autism spectrum disorders as well as controls.

email: ddegrasv@gmail.com

---

62. IMAGING AND STREAMING DATA ANALYSIS

Generalizable Two-Stage PCA for Confounding Adjustment

Sarah M. Weinstein*, University of Pennsylvania
Kristin A. Linn*, University of Pennsylvania
Russell T. Shinohara*, University of Pennsylvania

Two major challenges in biomedical research are dimension reduction and confounding adjustment. Recent methods have sought to simultaneously handle dimension reduction and confounding adjustment through simple modifications to principal component analysis (PCA). Although such methods perform well in-sample, their generalizability is limited when applying the rotations obtained from PCA in one dataset to a new dataset where the distribution of the confounder may have changed. We propose a generalizable two-stage approach to confounding adjustment and dimension reduction. Using our method on simulated image data, we find that variation explained by the confounder can be reduced or even eliminated, depending on the size of the confounding effect.

email: sarah.weinstein@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Permutation-Based Inference for Spatially Localized Signals in Longitudinal MRI Data

Jun Young Park*, University of Minnesota
Mark Fiecas, University of Minnesota

Alzheimer's disease is a neurodegenerative disease in which the degree of cortical atrophy in the brain serves as a useful imaging biomarker. A massive-univariate analysis, a simplified approach that fits a univariate model for every vertex along the cortex, is insufficient to model cortical atrophy because it does not account for spatial relatedness of cortical thickness from magnetic resonance imaging (MRI), and it can suffer from Type I error rate control. Using the longitudinal structural MRI from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we develop a permutation-based inference procedure to detect spatial clusters of vertices showing statistically significant differences in the rates of cortical atrophy. The proposed method uses spatial information to combine signals adaptively across nearby vertices, yielding high statistical power while maintaining an accurate family-wise error rate. When the global null hypothesis is rejected, we use a region selection algorithm to search for clusters of signals. We validate our method using simulation studies and apply it to the real data to show its superior performance over existing methods.

email: park1131@umn.edu

Geostatistical Modeling of Positive Definite Matrices: An Application to Diffusion Tensor Imaging

Zhou Lan*, The Pennsylvania State University
Brian Reich, North Carolina State University
Joseph Guinness, Cornell University
Dipankar Bandyopadhyay, Virginia Commonwealth University

Diffusion tensor imaging (DTI), a neuroimaging characterizing the brain's anatomical structure, produces a positive definite matrix for each voxel. Currently, only a few geostatistical models for positive definite matrices have been proposed. In this paper, we use the spatial Wishart process, a spatial random field where each positive definite matrix-variate marginally follows a Wishart distribution, and spatial dependence between random matrices is induced by latent Gaussian processes. This process is valid on an uncountable collection of spatial locations and is almost-surely continuous, leading to a reasonable means of modeling spatial dependence. We propose a spatial matrix-variate regression model based on the spatial Wishart process. We propose an approximation method to obtain a feasible Cholesky decomposition model and show that the Cholesky decomposition model is asymptotically equivalent to the spatial Wishart process model. The simulation studies and real data analysis demonstrate that the Cholesky decomposition process model produces reliable inference and improved performance compared to other methods.

email: zlan@psu.edu

Length Penalized Probabilistic Principal Curve with Application to Pharmacologic Colon Imaging Study

Huan Chen*, Johns Hopkins Bloomberg School of Public Health

The classical Principal Curve algorithm was developed as a nonlinear version of principal component analysis to model curves. However, existing principal curve algorithms with classical penalties, such as smoothness or ridge penalties, lack the ability to deal with complex curve shapes. In this manuscript, we introduce a robust and stable length penalty which solves issues of unnecessary curve complexity, such as the self-looping, that arise widely in principal curve algorithms. A novel probabilistic mixture regression model is formulated. A modified penalized EM Algorithm was applied to the model to obtain the penalized MLE. Two applications of the algorithm were performed. In the first, the algorithm was applied to the MNIST dataset of handwritten digits to find the centerline, not unlike defining a TrueType font. We demonstrate that the centerline can be recovered with this algorithm. In the second application, the algorithm was applied to construct a three dimensional centerline through single photon emission computed tomography images of the colon arising from the study of pre-exposure prophylaxis for HIV. The centerline in this application is crucial for understanding the distribution of the antiviral agents in the colon for HIV prevention. The new algorithms improves on previous applications of principal curves to this data.

email: hchen130@jhmi.edu

Image-on-Scalar Regression Via Interpretable Regularized Reduced Rank Regression

Tianyu Ding*, University of Pittsburgh
Dana Tudorascu, University of Pittsburgh
Annie Cohen, University of Pittsburgh
Robert Krafty, University of Pittsburgh

We propose an interpretable penalized multivariate high-dimensional method for image-on-scalar regression. The method is implemented within a reduced-rank regression framework. An encoder decoder based fused group lasso penalty is formulated that regularizes spatial smoothness, sparsity and group structure on the reduced dimension while maintaining interpretation on the image space. An algorithm is developed for model fitting that uses alternating direction method of multipliers (ADMM) to perform optimization within subspaces after dimension reduction. The estimators possess favorable statistical properties, even when the number of image response variables is much larger than the sample size. Empirical properties of the proposed approach are examined and compared to existing methods in simulation studies and in the analysis of PET images from subjects in a study of Alzheimer's disease.

email: tid16@pitt.edu

ABSTRACTS & POSTER PRESENTATIONS

Automatic Transformation and Integration to Improve Visualization and Discovery of Latent Effects in Imaging Data

Johann A. Gagnon-Bartsch*, University of Michigan
Gregory J. Hunt, William & Mary

Data transformation is an essential part of analysis, and choosing appropriate transformations for variables can enhance visualization, improve efficacy of analytical methods, and increase data interpretability. However determining appropriate transformations of variables from highly-multiplexed imaging data poses new challenges. Imaging data produces hundreds of covariates, each of which may have a different distribution and need a potentially different transformation. Determining an appropriate transformation for each of them is infeasible by hand. We explore simple, robust, and automatic transformations of highly-multiplexed image data. A central application of is to microenvironment microarray bioimaging data from the NIH LINCS program. Our robust transformations enhance visualization and improve the discovery of substantively relevant latent effects. The transformations enhance analysis of image features individually and also improve data integration approaches when combining together multiple features. Similar advantages may also be realized in the analysis of data from other highly-multiplexed technologies such as Cell Painting or Cyclic Immunofluorescence.

email: johanngb@umich.edu

63. CAUSAL INFERENCE AND PROPENSITY SCORE METHODS

Generalizing Randomized Trial Findings to a Target Population using Complex Survey Population Data

Benjamin Ackerman*, Johns Hopkins Bloomberg School of Public Health
Catherine R. Lesko, Johns Hopkins Bloomberg School of Public Health
Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Randomized trials are considered the gold standard for estimating causal effects. Trial findings are often used to inform health policy, yet their results may not generalize well to the target population of interest due to potential differences in effect moderators between the trial and population. Statistical methods have been developed to improve generalizability by combining trial and population data and weighting the trial to resemble the population on baseline covariates. Large health surveys with complex survey designs are a common source for population data; however, there is currently no best practice for incorporating survey weights when generalizing trial findings to a complex survey. We propose a two-stage weighting approach to properly incorporate the survey weights in this context, and examine the performance of this method through simulation. We then apply the methods to generalize findings from PREMIER, a lifestyle intervention trial, to a target population using an NHANES sample. The work highlights the importance in properly accounting for survey sampling design when generalizing trial findings to a population using data from a complex survey sample.

email: backer10@jhu.edu

Weak-Instrument Robust Tests in Two-Sample Summary-Data Mendelian Randomization

Sheng Wang*, University of Wisconsin, Madison
Hyunseung Kang, University of Wisconsin, Madison

Mendelian randomization (MR) is a popular method in genetic epidemiology to estimate causal effects of an exposure on an outcome of interest using genetic variants as instrumental variables (IV). Recently, the two-sample summary data setting is increasingly common in MR. Unfortunately, many existing methods in MR, while tailored for this data setting, cannot robustly handle weak instruments, where the correlation between the IVs and exposure is small. In this work, we leverage recent works in econometrics and propose a set point estimators and test statistics that (i) are robust to weak instruments and (ii) work with two-sample summary-level data. For point estimation, we extend a popular method in econometrics called limited information maximum likelihood (LIML) and an unbiased estimator under known signs. For tests for the exposure effect, we extend the Anderson-Rubin, Kleibergen, and conditional likelihood ratio tests and derive their asymptotic properties when the instruments are arbitrary weak. We conclude with a numerical study where we show that the proposed tests control size and have better power than current methods when instruments are weak.

email: swang676@wisc.edu

Propensity Score Matching Methods in Unbalanced Studies with Optimal Caliper Choice

Ziliang Zhu*, University of North Carolina, Chapel Hill
Toshio Kimura, Regeneron Pharmaceuticals, Inc.
Xinyi He, Agios Pharmaceuticals, Inc.
Zhen Chen, Regeneron Pharmaceuticals, Inc.

Propensity score (PS) matching is commonly used for confounding effect adjustment in causal inference. Within PS matching, we explored alternatives to traditional matching approaches which are often 1) one-time match using fixed k-to-1 match, and 2) pre-defined selection of caliper width with limited justification for this width selection. We proposed other PS matching methods for unbalanced studies, where the treatment arm and placebo arm have extremely disproportional amount of data. We studied and compared performance based on the criteria of least squared loss. Instead of one-time match, we explored multiple sampling with uniform distribution and dynamic sample size k-to-1 matching. We concluded that a Dynamic Sample Size approach, which includes all subjects falling in a given caliper, as the optimal method for causal effect estimation. Additionally, we evaluated methods and determined the optimal caliper width. We showed that Dynamic Sample Size approach with optimal caliper will achieve near-optimal rate under mild conditions in an unbalanced study situation. These findings were confirmed by simulation studies under various scenarios.

email: ziliang@live.unc.edu

ABSTRACTS & POSTER PRESENTATIONS

Mendelian Randomization with Statistical Warranty of All Core Assumptions

Zhiguang Huo*, University of Florida

Observational studies investigate the association between exposures and outcome variables, which is usually distorted by unknown confounders and reverse causations. Mendelian randomization (MR) is an effective method to estimate the causal relationship from exposures to outcome variables, by utilizing genetic variants as instrumental variables. It is becoming increasingly popular in large epidemiology studies, where genetic data are routinely collected. However, some core assumptions underlying MR are not testable, and are often assumed to be true with reasonable biological justifications. Leveraging the rich omics data in epidemiology studies, our goal is to establish a MR procedure, in which all underlying assumptions can be warranted via statistical tests. The proposed method is shown to have superior performance comparing to vanilla versions of MR in simulation. The algorithm is also evaluated in a large-scale multi-omics data.

email: zhuo@ufl.edu

Improved Propensity Score for Matching

Ernesto Ulloa*, University of Washington
Marco Carone, University of Washington
Alex Luedtke, University of Washington

When estimating causal effects of a binary exposure using observational data one needs to balance the distribution of confounders across treatment arms. Nearest neighbor matching with the propensity score (PS) effectively removes the effect of confounding allowing estimation of causal effects (Rosenbaum and Rubin). The asymptotic properties and distribution of the PS matching estimator have recently been derived (Abadie and Imbens). In particular, it has been shown that matching on the estimated propensity score is asymptotically more efficient compared to matching on the known PS. Nevertheless, the efficiency gain is not sufficient for the PS matching estimator to attain the asymptotic efficiency bound. In this work, we propose adding an additional covariate to the propensity score model that allows the estimator to maximize its efficiency gain. Further, we show that by adding this covariate the proposed PS matching estimator attains the efficiency gain asymptotically as one increases the number of matches. Finally, we evaluate the proposed matching estimator in several simulation settings and compare it to other similar estimators.

email: ulloae@uw.edu

A Likelihood Ratio Test for Multi-Dimensional Mediation Effects

Wei Hao*, University of Michigan
Peter X.K. Song, University of Michigan

Mediation analysis has become a widely used tool in studying whether the effect of an exposure on an outcome is mediated through some intermediate variables. When multiple mediators are of interest, the statistical inference on the joint mediation effect is challenging due to the composite null hypothesis involving many cases. We propose a likelihood ratio test and derive the asymptotic null distribution. We examine the performance of our method via extensive simulations, and compare it with two recent product significance test approaches. The simulation results show that the proposed method controls type one error rates and has better or similar power rates when comparing with the existing methods. We applied our method to a dataset from “Early Life Exposures in Mexico to ENvironmental Toxicants” study, to examine whether any of the 34 clusters of metabolites with number of mediators ranging from 5 to 32, mediate the effect of the fat intake on the HOMA-CP (homeostatic model assessment of insulin resistance using C-peptide) scores for the 203 study participants.

email: weihao@umich.edu

64. LONGITUDINAL DATA AND JOINT MODELS OF LONGITUDINAL AND SURVIVAL DATA

Estimation of the Joint Distribution of Survival Time and Mark Variable in the Presence of Dependent Censoring

Busola O. Sanusi*, University of North Carolina, Chapel Hill
Michael G. Hudgens, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill

In biomedical studies, there is often an interest to examine the relationship between survival time and mark variable, which is only observable at failure times. The mark and failure times could be correlated. In this framework of joint modeling, previous literature assumed that censoring process is independent of the failure process and as a result, the estimated joint distribution function may yield biased results when there is dependent censoring. We consider non-parametrically estimating the joint distribution of a survival time and mark variable while allowing for dependent censoring, by implementing the inverse probability of censoring weighting technique. We show that the proposed estimator is consistent and asymptotically normal. We examine the performance of the proposed estimator in finite samples via simulations and illustrate it in real data example.

email: sanus1bo@email.unc.edu

ABSTRACTS & POSTER PRESENTATIONS

A Multilevel Mixed Effects Varying Coefficient Model with Multilevel Predictors and Random Effects for Modeling Hospitalization Risk in Patients on Dialysis

Yihao Li*, University of California, Los Angeles
Danh V. Nguyen, University of California, Irvine
Esra Kurum, University of California, Riverside
Connie M. Rhee, University of California, Irvine
Yanjun Chen, University of California, Irvine Institute of Clinical and Translational Science
Kamyar Kalantar-Zadeh, University of California, Irvine
Damla Senturk, University of California, Los Angeles

For patients on dialysis, hospitalizations remain a major risk factor for mortality and morbidity. We use data from United States Renal Data System to model time-varying effects of hospitalization risk factors as functions of time since initiation of dialysis. To account for the three-level hierarchical structure in the data where hospitalizations are nested in patients and patients are nested in dialysis facilities, we propose a multilevel mixed effects varying coefficient model where multilevel (patient- and facility-level) random effects are used to model the dependence structure. The proposed model also includes multilevel covariates, where baseline demographics and comorbidities are among the patient-level factors, and staffing composition and facility size are among the facility-level risk factors. To address the challenge of high-dimensional integrals due to the hierarchical structure of the random effects, we propose a novel two-step approximate EM algorithm based on the fully exponential Laplace approximation. Inference for the varying coefficient functions and variance components is achieved via derivation of the standard errors using score contributions.

email: liyihao@ucla.edu

Structural Joint Modeling of Longitudinal and Survival Data

Bryan Blette*, University of North Carolina, Chapel Hill
Peter Gilbert, Fred Hutchinson Cancer Research Center
Michael Hudgens, University of North Carolina, Chapel Hill

Joint modeling of longitudinal and time-to-event data is an increasingly popular method for analyzing randomized trials to assess associations of longitudinal variables with failure events. Recently, it has also been used for observational data, including cohort studies and electronic health records data. Despite its documented utility, an analogous causal modeling framework has yet to be described. We introduce the structural joint model, a causal joint modeling approach which adjusts for potential confounding. We provide estimators for the model, explore their large sample properties, show that they are generalizations of those of a conditional-score joint model and a marginal structural Cox model, and compare these three models in a simulation study. In finite samples, only the structural joint model estimator was approximately unbiased with confidence intervals having nominal coverage in the presence of both confounding and longitudinal exposure measurement error. In contrast to a joint model, the structural joint model can estimate causal treatment effect in counterfactual scenarios defined by longitudinal exposure values which differ from those in the observed data.

email: blette@live.unc.edu

Bayesian Models for Joint Longitudinal and Competing Risks Data

Allison KC Furgal*, University of Michigan
Ananda Sen, University of Michigan
Jeremy M.G. Taylor, University of Michigan

Joint models are useful in analyzing data with a survival time and an associated longitudinal marker. When the survival outcome can have multiple causes, competing risks techniques must be incorporated in the joint model. Research in joint modeling of longitudinal and competing risks data has almost exclusively used cause-specific hazard functions. Such modeling is unable to capture the explicit effect of the association between the risk components and the longitudinal marker. We explore Bayesian joint model within a latent failure time framework using parametric models based on a multivariate Weibull and log-Normal distributions. Flexibility is added through nonparametric Dirichlet process priors. We evaluate our model via simulations. We illustrate the approach with an application to data from adrenocortical carcinoma patients at the University of Michigan measuring morphomic markers of body composition over time as well as time to cancer progression or death.

email: acullen@umich.edu

Joint Model for Survival and Multivariate Sparse Functional Data with Application to a Study of Alzheimer's Disease

Cai Li*, Yale University
Luo Xiao, North Carolina State University

Studies of Alzheimer's disease (AD) often collect multiple longitudinal clinical outcomes, which are correlated and can be predictive of AD progression. It is of great interest to investigate the association between the outcomes and time to AD onset. While joint modeling has received much attention in recent years, most works either assume parametric frameworks or focus on only a single longitudinal outcome. We model the multiple longitudinal outcomes as multivariate sparse functional data and propose a novel functional joint model. In particular, we propose a multivariate functional mixed model (MFMM) to identify the shared progression pattern and outcome-specific progression patterns of the outcomes, which enables more interpretable modeling of the association between outcomes and AD onset. The proposed method is applied to the Alzheimer's Disease Neuroimaging Initiative study (ADNI) and the functional joint model sheds new lights on inference of five longitudinal outcomes and their association with AD onset.

email: cli9@ncsu.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Semiparametric Joint Models to Study Growth and Islet Autoimmunity in Subjects at High Risk for Type 1 Diabetes

Xiang Liu*, University of South Florida
Roy Tamura, University of South Florida
Kendra Vehik, University of South Florida
Jeffrey Krischer, University of South Florida

In the pathogenesis of Type 1 diabetes (T1D), the accelerator and overload hypotheses postulate that rapid growth and high weight speed up both beta cell insufficiency and insulin resistance. Islet autoimmunity (IA) precedes clinical T1D. In order to study the effect of weight on the risk of IA, joint modeling of longitudinal and time-to-event data is necessary to consider the growth process and the development of IA, while accounting for their interrelationship. A child's weight trajectory in early life is nonlinear. In addition, longitudinal data collected in large observational studies introduces burdens in computation. In order to address the nonlinearity and also the computational challenge, we introduce a Bayesian semiparametric joint model, in which a partial linear mixed sub-model was used to model the weight trajectories. Splines were used to approximate the nonlinear weight trajectories and the joint model was estimated within a Bayesian framework. We used data from the Environmental Determinants of Diabetes in the Young (TEDDY) study to illustrate its use and showed that weight is associated with the risk of IA.

email: xiang.liu@epi.usf.edu

Marginal Inference in Transition Models with Generalized Estimating Equations: What is Being Estimated?

Danping Liu*, National Cancer Institute, National Institutes of Health
Joe Bible, Clemson University
Paul S. Albert, National Cancer Institute, National Institutes of Health
Bruce G. Simons-Morton, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Generalized estimating equations (GEEs) are commonly used to estimate transition models. When the Markov assumption does not hold but first-order transition probabilities are still of interest, the transition inference is sensitive to the choice of working correlation. With a random process transition model as the true data generating mechanism, two types of transition probabilities at the population level are defined: naive transition probabilities that average across all the transitions and population-average transition probabilities that average the subject-specific transition probabilities. Through asymptotic bias calculations and simulations, we demonstrate that the unstructured working correlation estimates the population-average transition probabilities while the independence working correlation estimates the naive transition probabilities. We further demonstrate that the sandwich estimator fails for unstructured GEE and recommend either jackknife or bootstrap variance estimates. The proposed method is motivated by and applied to the NEXT Generation Health Study, to estimate the population-average transition probabilities of alcohol use in adolescents.

email: danping.liu@nih.gov

65. PERSONALIZED MEDICINE AND BIOMARKERS

Synergistic Self-Learning of Individualized Dietary Supplement Rules from Multiple Health Benefit Outcomes

Yiwang Zhou*, University of Michigan
Peter X.K. Song, University of Michigan

Deriving individualized dietary supplement rules (IDSRs) based on the effects of nutrients on health is a hard problem in precision nutrition. Challenges arise in estimating IDSRs with multiple variables pertinent to health benefit: (i) outcomes are of different clinical relevance to the underlying health benefit; (ii) outcomes are measured with different sample sizes. This paper is motivated by a clinical trial aiming to assess the effect of calcium supplement for pregnant women in reducing fetal lead exposure. We propose to integrate different types of blood lead concentrations of varying clinical relevance and sample sizes in training IDSRs. We develop a new support vector machine (SVM) that allows synergizing heterogeneous data sources in a weighted self-learning paradigm. As an extension to outcome weighted learning (OWL), our proposed synergistic self-learning (SS-learning) incorporates multiple outcomes in the optimization, resulting in an optimal solution of individual rules for dietary supplements. We establish the algorithmic convergence of SS-learning and illustrate its performance through both simulation studies and real data analysis of the motivating trial.

email: yiwangz@umich.edu

Integrative Network Learning for Multi-Modality Biomarker Data

Shanghong Xie*, Columbia University
Donglin Zeng, University of North Carolina, Chapel Hill
Yuanjia Wang, Columbia University

The biomarker networks measured by different modalities of data (e.g., structural magnetic resonance imaging (sMRI), diffusion tensor imaging (DTI)) may share the same underlying biological model. In this work, we propose a node-wise biomarker graphical model to leverage the shared mechanism between multi-modality data to provide a more reliable estimation of target modality network and account for the heterogeneity in networks due to differences between subjects and networks of external modality. Latent variables are introduced to represent the shared unobserved biological network and the information from the external modality is incorporated to model the distribution of the underlying biological network. An approximation approach is used to calculate the posterior expectations of latent variables to reduce time. The performance of the proposed method is demonstrated by extensive simulation studies and an application to construct gray matter brain atrophy network of Huntington's disease by using sMRI data and DTI data. The estimated network measures are shown to be meaningful for predicting follow-up clinical outcomes in terms of patient stratification and prediction.

email: sx2168@cumc.columbia.edu

ABSTRACTS & POSTER PRESENTATIONS

An Optimal Design of Experiments Approach to Closed-Loop Target-Controlled Induction of Anesthesia for Robustness to Interpatient PK/PD Variability: A Simulation Study

Ryan T. Jarrett*, Vanderbilt University
Matthew S. Shotwell, Vanderbilt University

Closed-loop control systems in health care adjust inputs on the basis of feedback received from the patient, typically to maintain a set point specified for the system. We propose a framework for closed-loop control over the induction of anesthesia based on replacing the fixed set point with a target function that can be optimized for clinical goals. Further, we employ principles of optimal experimental design to increase robustness to unobserved variability in patient pharmacokinetics and pharmacodynamics (PK-PD). Closed-loop control is established through Bayesian updates of the prior distribution to continually refine a patient-specific PK-PD model. We demonstrate the utility of or method through simulations and show that it results in reduced target overshoot and a shorter time until patients stably enter the target control region. Our framework allows a clinician to prioritize rapid patient sedation relative to minimizing overshoot, as may be required by the clinical context. Our results suggest that the proposed method provides a viable approach to closed-loop control of induction and is robust to unobserved inter-patient PK-PD variability.

email: ryan.t.jarrett@vanderbilt.edu

Utilization of Residual Lifetime Quantiles to Optimize Personalized Biomarker Screening Intervals

Fang-Shu Ou*, Mayo Clinic
Phillip J. Schulte, Mayo Clinic
Martin Heller, Private Practitioner

In oncology surveillance, a biomarker may be measured routinely to monitor disease progression. This type of surveillance was scheduled in pre-specified intervals following treatment guidelines regardless of the actual value of the biomarker. Taking measurements too often does not lead to much information gain yet increases medical costs and patient discomfort. Whereas, recording insufficient measurements may result in a delay in detection. Personalizing the screening interval based on previously observed marker values would optimize the screening interval and strike a balance between information gained and measurement costs. We frame this as a quantile regression problem conditional on current survival time and observed biomarker values. The screening time for the next biomarker assessment is subject to a preselected risk threshold using the quantile of conditional survival for the progression event and observed biomarker values. The proposed methods are applied to a study of patients following radical prostatectomy with screening of prostate specific antigen in the presence of survival from prostate cancer recurrence.

email: ou.fang-shu@mayo.edu

Precision Medicine Using MixedBART for Repeated Measures

Charles K. Spanbauer*, Medical College of Wisconsin
Rodney Sparapani, Medical College of Wisconsin

Bayesian Additive Regression Trees (BART) is a Bayesian non-parametric machine learning ensemble model with excellent out of sample predictive capabilities; yet, its use in biostatistical analyses and biomedical research is in its infancy. One barrier is the implicit assumption of independence between observations that is inherent in the original BART model. In contrast, repeated measures data is common in biomedical research and so using BART may not be justified in such a scenario. For instance, BART can be applied to clinical trial data with the goal of finding treatment response heterogeneity, or individualized treatment rules, based on patient characteristics. Therefore, we are proposing mixedBART which utilizes parametric random effect terms to relax the independence assumption. A general random effect design matrix is allowed allowing more complicated models beyond the random intercept model such as growth curve models. In addition, time to event outcomes are considered using the accelerated failure time model. Simulation studies and an application to clinical research and precision medicine are presented.

email: cspanbauer@mcw.edu

A Statistical Method to Estimate Sleep Duration from Actigraphy Data

Jonggyu Baek*, University of Massachusetts Medical School
Margaret Banker, University of Michigan
Erica C. Jansen, University of Michigan
Karen E. Peterson, University of Michigan
E. Andrew Pitchford, Iowa State University
Peter X. K. Song, University of Michigan

Sleep duration is a recognized determinant of mental health, obesity and CVD, cognition and memory across the lifespan. Due to convenience and cost, sleep duration is often measured with self-report; yet, tends to be biased. Recently, actigraphs have been recommended as an objective measure of sleep duration. Various sleep evaluation methods have been developed upon regression methods with coefficients constructed on minute-by-minute data measured at a specific placement. Because activity counts per minute may be affected by various factors (e.g., physical characteristics), regression-based methods within specific populations may not be appropriate for wider use. To fill these gaps, we propose a statistical method to obtain robust and consistent sleep duration estimates. Our proposed method is built upon fast changepoint detection method: pruned dynamic programming (PDP). First, we identify temporal segments identified by PDP; then develop a calling algorithm to capture sleep periods. Our proposed method is applied in the Multi-Ethnic Study of Atherosclerosis (MESA) Sleep and the Early Life Exposure in Mexico to ENvironmental Toxicants (ELEMENT) studies.

email: jonggyu.baek@umassmed.edu

ABSTRACTS & POSTER PRESENTATIONS

66. STATISTICAL GENETICS: SINGLE-CELL SEQUENCING DATA

SMNN: Batch Effect Correction for Single-Cell RNA-seq Data via Supervised Mutual Nearest Neighbor Detection

Gang Li*, University of North Carolina, Chapel Hill
Yuchen Yang, University of North Carolina, Chapel Hill
Huijun Qian, University of North Carolina, Chapel Hill
Kirk C. Wilhelmsen, University of North Carolina, Chapel Hill
Yin Shen, University of California, San Francisco
Yun Li, University of North Carolina, Chapel Hill

Batch effect correction has been recognized to be indispensable when integrating single-cell RNA-sequencing (scRNA-seq) data. A recent study proposed an effective correction using mutual nearest neighbors (MNN). However, the proposed MNN method is unsupervised in that it ignores cluster label information of single cells. Such information can further improve effectiveness of batch effect correction, particularly when true biological differences are not orthogonal to batch effect. Under this motivation, we propose SMNN which performs supervised mutual nearest neighbor detection for batch effect correction. SMNN either takes cluster/cell-type information as input, or, in the absence of such information, infers cell types by performing clustering. It then detects mutual nearest neighbors within matched cell types and corrects batch effect accordingly. Our extensive evaluations show that SMNN provides improved merging, leading to reduced differentiation across batches over MNN. Furthermore, SMNN retains more cell type-specific features after correction. SMNN is implemented in R, and available at https://yunliweb.its.unc.edu/SMNN/.

email: frankleegangli@gmail.com

Multiple Phenotype-Multiple Genotype Testing with Principal Components

Andy Shi*, Harvard University
Ryan Sun, University of Texas MD Anderson Cancer Center
Xihong Lin, Harvard University

The increasing popularity of large-scale genetic compendiums has driven interest in (1) testing sets of genotypes against a single phenotype and (2) testing sets of phenotypes against a single genotype. Using correlated sets of variants and outcomes can boost power to detect novel associations, reduce the multiple testing burden, and produce more interpretable conclusions about the genetic etiology of complex diseases by incorporating prior biological knowledge into set definitions. However, less work has focused on the testing problem when sets are formed for both genotypes and phenotypes. In this paper we propose and study the performance of principal components-based methods that can jointly test for the association between multiple phenotypes and multiple genotypes. We demonstrate how the properties of each test are determined by the correlation structures of the genotypes and phenotypes and the direction of effects linking the two sets. We also propose an omnibus test that is robust to the direction of the effects. Simulations demonstrate that our method controls Type-I error. We apply our method to analyze correlated biomarkers in a study of cardiovascular outcomes.

email: andyshi@g.harvard.edu

Single-Cell ATAC-seq Signal Extraction and Enhancement with SCATE

Zhicheng Ji*, Johns Hopkins University
Weiqiang Zhou, Johns Hopkins University
Hongkai Ji, Johns Hopkins University

Single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) is the state-of-the-art technology for analyzing genome-wide regulatory landscape in single cells. Single-cell ATAC-seq data are sparse and noisy. Analyzing such data is challenging. Existing computational methods cannot accurately reconstruct activities of individual cis-regulatory elements (CREs) in individual cells or rare cell subpopulations. We present a new statistical framework, SCATE, that adaptively integrates information from co-activated CREs, similar cells, and publicly available regulome data to substantially increase the accuracy for estimating activities of individual CREs. We show that using SCATE, one can better reconstruct the regulatory landscape of a heterogeneous sample.

email: zhichengji@gmail.com

A Neural Network Based Dropout Correction for Single-Cell RNA-Seq Data with High Sparsity

Lingling An*, University of Arizona
Xiang Zhang, University of Arizona
Siyang Cao, University of Arizona

High dimensional single-cell RNA sequencing (scRNA-seq) data become available recently with high sparsity. The high sparsity of the scRNA-seq data is an obstacle for the downstream analysis, such as clustering and classification analyses. Most existing imputation methods cannot handle the data with extremely high sparsity very well. In this paper, we propose a neural network-based imputation for data with high sparsity, where autoencoder is used with a novel loss function to recognize and impute the dropouts in scRNA-seq data. The simulations and experiments show that the new method outperforms the existing state-of-the-art methods. It can process a dataset of millions of entries on GPUs or CPUs.

e-mail: anling@email.arizona.edu

ABSTRACTS & POSTER PRESENTATIONS

A Novel Surrogate Variable Analysis Framework in Large-Scale Single-Cell RNA-seq Data Integration

Chao Huang*, Florida State University
Yue Julia Wang, Florida State University
Madison Layfield, Florida State University

More and more large-scale single-cell RNA-seq datasets have been seen in recent years. These datasets are produced in different laboratories and at different times, introducing batch effects that may compromise the integration and interpretation of the data. Existing single-cell RNA-seq analysis methods incorrectly assume that the composition of cell populations is either known or identical across batches. In this talk, we present a novel surrogate variable analysis strategy for detecting the significant covariates of interest while adjusting potential confounders. In addition, a clustering analysis is incorporated which provides a powerful tool for biological subpopulation identification. The performance of our proposed method is assessed via simulation studies and one real data analysis on the single-cell RNA-seq datasets collected from College of Medicine at Florida State University.

e-mail: chuang7@fsu.edu

Robust Normalization of Single-Cell RNA-seq Data using Local Smoothing and Median Ratio

Chih-Yuan Hsu*, Vanderbilt University Medical Center
Qi Liu, Vanderbilt University Medical Center
Yu Shyr, Vanderbilt University Medical Center

Single-cell RNA-seq (scRNAseq) normalization is an essential step to correct unwanted biases caused by sequencing depth, capture efficiency, dropout and other technical factors. Most scRNAseq normalization methods mainly focus on adjusting the effect resulted from sequencing depth by modeling count-depth relationship and/or assuming a specific distribution for read counts. However, the existence of cell-specific technical biases other than sequencing depth and the unfaithful model assumption will lead to over or under-correction. We present a new normalization method for correcting any known or hidden technical cofounders without any model assumption. The method obtains the size factors by using local smoothing and median-ratio normalization. Evaluation with simulated and real scRNAseq data validated that our method is more robust and powerful than existing methods.

email: chih-yuan.hsu@vumc.org

Subpopulation Identification for Single-Cell RNA-Sequencing Data Using Functional Data Analysis

Kyungmin Ahn*, RIKEN Center for Biosystems Dynamics Research, Japan
Hironobu Fujiwara, RIKEN Center for Biosystems Dynamics Research, Japan

In single-cell RNA-sequencing (scRNA-seq) data analysis, a number of statistical tools in multivariate data analysis (MDA) have been developed to help analyze the gene expression data. In this paper, we propose a functional data analysis (FDA) approach on scRNA-seq data whereby we consider each cell as a single function. To avoid a large number of dropouts and reduce the high dimensionality of the data, we first perform a principal component analysis (PCA) and assign PCs to be the amplitude of the function. For the phase components, we propose two criteria: we use the PCs directly from PCA, and we sort the PCs by the genetic spatial information. For the latter, we embed the spatial information of genes by aligning the genomic gene locations to be the phase of the function. To demonstrate the robustness of our method, we apply several existing FDA clustering algorithms to the gene expression data to improve the accuracy of the classification of the cell types against the conventional clustering methods in MDA. As a result, the FDA clustering algorithms achieve superior accuracy on simulated data as well as real data.

email: kyungmin.ahn@riken.jp

67. SEMIPARAMETRIC AND NONPARAMETRIC METHODS AND APPLICATIONS

A Semiparametric Alternative Method to Conditional Logistic Regression for Combining Biomarkers under Matched Case-Control Studies

Wen Li*, University of Texas Health Science Center at Houston
Ruosha Li, University of Texas Health Science Center at Houston
Ziding Feng, Fred Hutchinson Cancer Research Center
Jing Ning, University of Texas MD Anderson Cancer Center

Incorporating promising biomarker into the current cancer screening practice for early-detection is increasingly appealing since the performance of current screenings is still far from satisfactory. The matched case-control studies are commonly implemented to evaluate the discriminatory power of biomarker candidates, with an intention to eliminate some confounding effects. Methods for analyzing data from matched case-control studies have been focused on the conditional logistic regression, which assumes a logit link function. We propose a distribution-free method for identifying an optimal combination including biomarker information to discriminate cases and controls. We are particularly interested in combinations with clinically and practically acceptable specificities to avoid a large number of subjects undergoing unnecessary and possibly intrusive diagnosis in general population screening. We establish desirable properties for the derived combination and confirm its improved finite sample performance in simulations. We illustrate the proposed method on a real data set.

email: liwenmoi@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Exponential and Super-Exponential Convergence of Misclassification Probabilities in Nonparametric Modeling

Richard Charnigo*, University of Kentucky
Cidambi Srinivasan, University of Kentucky

We establish several results for limits of scaled log misclassification probabilities, when data arise from one of finitely many candidate nonparametric regression models. With normality and identifiability assumptions, misclassification probabilities converge to zero exponentially. If a certain symmetry prevails, then we also have the specific result that $\lim_{n \to \infty} n^{-1} \log {\mathbb P}(\mbox{ misclassification }) = -\int_0^1 |\eta_1(w)|^2 dw / (8 \sigma^2)$, where σ^2 is the noise variance and η_1 is the difference between true and spurious mean response functions. Even in the absence of normality, but with some other conditions, it is possible to obtain an exact expression for the aforementioned limit. Surprisingly, there are some noise distributions with which the misclassification probabilities decay super-exponentially. Our results are potentially relevant to scientific applications at the interface of classification and nonparametric regression, including Raman spectroscopy and nanoparticle characterization.

email: richard.charnigo@uky.edu

Zero-Inflated Quantile Rank-Score Based Test (ZIQRank) with Application to scRNA-seq Differential Gene Expression Analysis

Wodan Ling*, Fred Hutchinson Cancer Research Center
Ying Wei, Columbia University
Wenfei Zhang, Sanofi

Differential gene expression analysis with scRNA-seq data is challenging. First, multimodality of gene expression and heterogeneity among different cell conditions lead to crossings of the gene distributions. Thus, existing mean-based parametric approaches are limited, while the quantile regression that examines various locations in the distribution gives a higher power. However, scRNA-seq data is zero-inflated, then we cannot directly apply the quantile regression method since its basic assumptions are violated. We propose a quantile rank-score based test for differential distribution detection of zero-inflated outcomes to handle both multimodality and zero-inflation. It comprises a test in logistic model for the zero-inflation, and rank-score based tests on multiple quantiles of the positive part adjusting for zero-inflation. All p-values are combined by MinP or Cauchy procedures. The proposed tests are evaluated with simulation studies and have a higher power in detecting true differentially expressed genes (DEGs) than existing methods. We also apply it to a real scRNA-seq data, and successfully detect a group of crucial genes associated with human glioma.

email: wl2459@columbia.edu

A Nonparametric MC-SIMEX Method

Lili Yu*, Georgia Southern University
Congjian Liu, Georgia Southern University
Jingjing Yin, Georgia Southern University
Jun Liu, Georgia Southern University

MC-SIMEX proposed by Kuchenho et al. is a general method to handle the categorical data with measurement error. It consists of two steps, simulation step and extrapolation step. In the simulation step, it simulates observations with varying degrees of measurement error. Then parameter estimators for varying degrees of measurement error are obtained based on these observations. In the extrapolation step, it uses parametric extrapolation function to obtain the parameter estimators for data with no measurement error. However, as shown in many studies, the parameter estimators are still biased due to the parametric extrapolation function used in the MC-SIMEX method. Therefore, we proposed a nonparametric MC-SIMEX method in which we use a nonparametric extrapolation function. It uses fractional polynomial method with cross validation to choose the appropriate fractional polynomial terms. Simulations show it results in unbiased parameter estimators.

email: lyu@georgiasouthern.edu

Nonparametric Regression for Error-Prone Homogeneous Pooled Data

Dewei Wang*, University of South Carolina

This presentation introduces new nonparametric regression methods for homogeneous pooling in the presence of measurement error. Pooling is a cost-effective method for detecting rare infectious diseases and measuring biomarker concentration levels from irreplaceable specimens. Homogeneous pooling refers to the case where specimens with similar covariates are pooled together. This type of pooling helps nonparametric regression achieve the same convergence rate as in the case of no pooling despite the cost savings. However, when the observed covariate contains measurement error, current methods fail to deliver a consistent estimator. The approach I shall introduce is the only nonparametric method available.

email: deweiwang@stat.sc.edu

ABSTRACTS & POSTER PRESENTATIONS

Testing for Uniform Stochastic Ordering among k Populations

Chuan-Fa Tang*, University of Texas, Dallas
Dewei Wang, University of South Carolina

We develop nonparametric goodness-of-fit (GOF) tests for k uniform stochastic ordered distributions. The new tests extend the GOF testing procedures proposed by Tang et al. (2017) with data-driven critical values and achieve the correct sizes asymptotically without determining the least favorable configuration. We provide the finite sample performances and apply the tests to a data set involving the biomarker microfibrillar-associated protein 4 (MFAP4).

email: chuan-fa.tang@utdallas.edu

k-Tuple Partially Rank-Ordered Set Sampling

Kaushik Ghosh*, University of Nevada, Las Vegas
Marvin C. Javier, University of Nevada, Las Vegas

Ranked Set Sampling (RSS), introduced by McIntyre (1952), and other related methods, such as Partially Rank-Ordered Set Sampling (PROSS) developed by Ozturk (2011) have shown that inclusion of a ranking mechanism produces estimators with lower variance than their simple random sample (SRS)-based counterparts. Like RSS, PROSS takes only one measurement from each partially ranked-ordered set. This article considers selecting multiple observations from each partially rank-ordered set. Such sampling may become necessary due to cost constraints and our inability to rank units perfectly. Estimation of the mean, variance and distribution function as well as properties of the resulting estimators are investigated. The proposed method is illustrated using an application to forestry data.

email: kaushik.ghosh@unlv.edu

68. CHALLENGES AND OPPORTUNITIES IN METHODS FOR PRECISION medicine

Subgroup-Effects Models (SGEM) for Analysis of Personal Treatment Effects

Peter X.K. Song*, University of Michigan
Ling Zhou, University of Michigan
Shiquan Sun, University of Michigan
Haoda Fu, Eli Lilly and Company

The emerging field of precision medicine is transforming statistical analysis from the classical paradigm of population-average treatment effects into that of personal treatment effects. This new scientific mission has called for adequate statistical methods to assess heterogeneous covariate effects in regression analysis settings. We focus on a subgroup analysis that consists of two primary analytic tasks: identification of treatment effect subgroups and individual group memberships, and statistical inference on treatment effects by subgroup. We propose an approach to synergizing supervised clustering analysis via ADMM algorithm and statistical inference on subgroup effects via EM algorithm. Our proposed procedure, termed as Hybrid Operation for Subgroup Analysis (HOSA), enjoys computational speed and numerical stability with interpretability and reproducibility. We establish key theoretical properties for both proposed clustering and inference procedures. Numerical illustration includes extensive simulation studies and an analysis of motivating data from a randomized clinical trial to compare two drugs treating patients with type II diabetes.

email: pxsong@umich.edu

Kernel Optimal Orthogonality Weighting for Estimating Effects of Continuous Treatments

Michele Santacatterina*, Cornell University

Many scientific questions require estimating the effects of continuous treatments, which relationships with an outcome are usually described by dose-response curves. Outcome modeling and methods based on the generalized propensity score are the most commonly used methods to evaluate continuous effects. However, these methods may be sensitive to model misspecification. In this paper, we propose Kernel Optimal Orthogonality Weighting (KOOW), a convex optimization-based method, for estimating the effects of continuous treatments. KOOW finds weights that minimize the penalized weighted functional covariance between the continuous treatment and the confounders. By minimizing this quantity while simultaneously penalizing the weights, KOOW successfully provides weights that optimally orthogonalize confounders and the continuous treatment. We describe its properties and valuate its comparative performance in a simulation study. Using data from the Women's Health Initiative observational study, we apply KOOW to evaluate the effect of red meat consumption on blood pressure.

email: santacatterina@cornell.edu

Inference on Individualized Treatment Rules from Observational Studies with High-Dimensional Covariates

Yingqi Zhao*, Fred Hutchinson Cancer Research Center
Muxuan Liang, Fred Hutchinson Cancer Research Center
Young-Geun Choi, Sookmyung University
Yang Ning, Cornell University
Maureen Smith, University of Wisconsin, Madison

Individualized treatment rules (ITR) assign treatments according to different patient's characteristics. As recent clinical datasets consist of a large number of pretreatment covariates, numerous variable selection methods have been proposed to identify prescriptive variables (variables contributing to the true optimal treatment rules). However, much less attention has been given to statistical inference for those prescriptive variables. Furthermore, when the dataset is observational, it is particularly challenging to infer the estimated rules. We propose a hypothesis testing procedure for the high-dimensional ITRs. Specifically, from a general framework that directly optimizes overall treatment bene t, we construct a local test for testing low dimensional components of the ITRs. The procedure can apply to observational studies by taking into account the additional variability from the estimation of propensity score. The proposed methodology is illustrated with numerical studies and a real data example on electronic health records of patients with Type-II Diabetes.

email: yqzhao@fredhutch.org

ABSTRACTS & POSTER PRESENTATIONS

Integrative Analysis of Electronic Health Records for Precision Medicine

Yuanjia Wang*, Columbia University
Jitong Luo, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill

Electronic health records (EHRs) automatically capture patients' health informa- tion through normal medical practice and has increasingly become an important source of patient data for personalized medicine. In EHRs, disease biomarkers from the same patient are recorded longitudinally at clinical encounters and these correlated biomark- ers can be either continuous, binary or counts. In order to comprehensively assess patient's disease comorbidity and susceptibility, it is necessary to characterize these biomarkers over time in an integrative way, while at the same time, to tackle the challenges in EHRs such as the biomarkers are sparsely measured at irregular and in- formative clinical encounters. In this talk, we propose integrative analysis for mixed types/modes of biomarkers through latent multivariate Gaussian temporal processes. These processes are used to capture between-patient and between-marker heterogeneity and optimize individualized treatment strategies for precision medicine.

email: yw2016@cumc.columbia.edu

69. RECENT DEVELOPMENTs IN RISK ESTIMATION AND BIOMARKER MODELING WITH A FOCUS IN ALZHEIMER'S DISEASE

Analyzing Semi-Competing Risks Data as a Longitudinal Bivariate Process

Sebastien Haneuse*, Harvard T.H. Chan School of Public Health
Daniel Nevo, University of Tel Aviv

Semi-competing risks refers to the setting where interest lies in some non-terminal time-to-event outcome, the occurrence of which is subject to a terminal event (usually death). Key to semi-competing risks data is that they provide an opportunity to learn about whether and how the two events co-vary. Existing analysis approaches, however, fail to take advantage of this. We propose a novel framework for the analysis of semi-competing risks data that views the two outcomes through the lens of a longitudinal bivariate process on a partitioning of the time scale. At the core of the framework are three time interval-specific regression models, each specified in a manner analogous to a generalized linear model, with time-varying components represented via B-splines. Key to the framework is that it captures two distinct forms of dependence, "local" and "global" dependence, both of which have intuitive clinical interpretations. Estimation and inference is performed via penalized maximum likelihood, and can accommodate both right censoring and left truncation (as needed). The methods are motivated by and illustrated with data from the Adult Changes in Thought study.

email: shaneuse@hsph.harvard.edu

Biomarker Models for Early Alzheimer's Disease Risk Prediction Before Symptoms Appear

Zheyu Wang*, Johns Hopkins University

The recognition of the decade-long asymptomatic stage of AD has greatly impact the research and therapeutic development to focus on the preclinical stage of AD pathogenic process, at which time disease-modifying therapy is more likely to be effective. However, the preclinical stage imposes a major challenge in investigating biomarkers for early AD detection, because 1) using the clinical diagnosis as the reference point can be in error, especially in the early course of the disease; and 2) most AD studies do not have autopsy data to confirm diagnoses. Until technology advance allows for brain examination with "autopsy level" clarity, an appropriate statistical method that directly address the unobservable nature of preclinical AD progression is necessary for any rigorous AD biomarker evaluation and for efficient analyzing AD study data where only clinical data are available and neuropathology data are not yet available. We propose a latent variable model to study the underlying AD pathophysiology process revealed by multidimensional markers and apply the model to two different AD data sets.

email: wangzy@jhu.edu

A Statistical Test on the Ordering of Changes in Biomarkers for Preclinical Alzheimer's Disease

Chengjie Xiong*, Washington University in St. Louis

Multivariate cross sectional and longitudinal data arise in studies of many chronic diseases such as Alzheimer's Disease (AD). An important scientific question in AD research is to locate the biomarkers that may exert the earliest changes in the disease process, likely many years prior to the onset of clinical symptoms. The answer to this question is crucial as it may help identify the target for earliest possible prevention or therapeutic intervention. We develop a statistical methodology to address these very questions by jointly modeling multiple biomarkers, and define their orderings of changes by using the estimated rates of changes. Specifically, we choose a multivariate random intercept and random slope model, and propose a confidence interval estimate and a statistical test for the ordering of changes across multiple biomarkers whose sampling distribution depends on the modified Bessel functions of the second type. Finally, we demonstrate the proposed methodology by applying it to a biomarker study of AD including cerebrospinal fluid (CSF) and neuroimaging biomarkers, and infer on the likely orderings of the biomarker changes during the preclinical stage of AD.

email: chengjie@wustl.edu

ABSTRACTS & POSTER PRESENTATIONS

Changepoint Estimation for Biomarkers of Alzheimer's Disease

Laurent Younes*, Johns Hopkins University

Based on a two-phase regression model, we identify changepoints in several biomarkers prior to the clinical onset of Alzheimer's disease (AD). After describing the model and sketching the estimation algorithm, we will present results examining nine measures based on cerebrospinal fluid, magnetic resonance imaging and cognitive testing, obtained from 306 cognitively normal individuals, a subset of whom subsequently progressed to the symptomatic phase of AD. We determined significant changepoints for all nine measures, most of them estimated 10–15 years prior to symptom onset. These findings highlight the long period of time prior to symptom onset during which AD pathology is accumulating in the brain.

email: laurent.younes@jhu.edu

70. CLINICAL TRIAL DESIGNS IN A NEW ERA OF IMMUNOTHERAPY: CHALLENGES AND OPPORTUNITIES

Immune-Oncology Agents: Endpoints and Designs

Hao Wang*, Johns Hopkins University School of Medicine
Gary Rosner, Johns Hopkins University School of Medicine

As our understanding of the immune system and how it and cancers interact have led to huge advances in oncology. The discovery and development of immune checkpoint inhibition and drugs that target the tumor's effect on checkpoints have revolutionized the treatment of multiple cancers, particularly melanoma and non-small cell lung cancer. At the same time, researchers are developing vaccines to treat cancers. Observations of difference in behaviors of tumors after patients receive these agents have led researchers to reconsider traditional study endpoints and develop new ones. These innovations have also led to new designs. In this talk, we discuss several of the endpoints for evaluating effect and effectiveness of immunotherapies in oncology. We look at endpoints that seek to elucidate immune responses in patients, as well as clinical endpoints. We point out several concerns related to the endpoints and consider some ways one might address these concerns.

email: hwang76@jhmi.edu

Adaptive Dose Finding Based on Safety and Feasibility in Early-Phase Clinical Trials of Adoptive Cell Immunotherapy

Nolan A. Wages*, University of Virginia
Camilo E. Fadul, University of Virginia

Dose feasibility is a challenge that may arise in the development of adoptive T cell therapies for cancer. In early-phase clinical trials, dose is quantified either by a fixed or per unit body weight number of cells infused. It may not be feasible, however, to administer a patient's assigned dose due to an insufficient number of cells harvested or functional heterogeneity of the product. The study objective becomes to identify the maximum tolerated dose with high feasibility of being administered. This talk describes a new dose-finding method that adaptively accounts for safety and feasibility endpoints in guiding dose allocation. We apply the proposed methodology in a single simulated trial and evaluate its operating characteristics through extensive simulation studies. A design that incorporates feasibility, as a function of the quantity and quality of the product manufactured, in addition to safety will have an impact on recommended phase II doses in studies that evaluate patient outcomes.

email: nwages@virginia.edu

Novel Bayesian Phase I/II Designs for Identifying Safe and Efficacious Treatments for Immunotherapy

J. Jack Lee*, University of Texas MD Anderson Cancer Center

The primary goal in drug development is to find safe and efficacious treatments. Several adaptive Phase I/II designs have been proposed to identify the optimal biological dose. Various dose-finding methods such as the family of Bayesian optimal interval designs (BOIN) are used to monitor the toxicity outcomes. The joint distribution of the toxicity and efficacy outcome can be captured by inducing the association between the two or by applying the multinomial-Dirichlet model. Bayesian optimal phase 2 design (BOP2) can be used for monitoring and providing early stopping for phase 2 studies with multiple toxicity and/or efficacy endpoints. In addition, the best treatment can be selected by the utility-based approach to enable the efficacy and toxicity tradeoff. Furthermore, a short-term endpoint can be used to predict the delayed efficacy outcome often observed in immunotherapy to facilitate the timely decision making. Simulation studies will be reported to compare the performance of various methods. User-friendly Shiny applications are developed to facilitate the design and implementation of such studies.

email: jjlee@mdanderson.org

Impact of Design Misspecification in Immuno-Oncology Trials

Jennifer Le-Rademacher*, Mayo Clinic
Quyen Duong, Mayo Clinic
Tyler Zemla, Mayo Clinic
Sumithra J. Mandrekar, Mayo Clinic

Immuno-Oncology (IO) therapy has recently been approved for various cancers. IO is effective in a small subset of patients in whom the response can be durable. The efficacy of IO is often delayed compared to other types of cancer treatment. Traditional trial designs using log-rank test may no longer be appropriate. The difference in restricted mean survival time (RMST) and the cure rate model have been proposed as alternative endpoint for IO trials. As trials are designed based on a set of assumptions about the survival patterns, when these assumptions deviate from the true pattern, the trial conclusions may be erroneous. The objective of this work is to evaluate how misspecified design assumptions affect trial conclusions. Specifically, we will provide a summary of the survival patterns reported in IO trials and how they differ from the assumptions used for design. We will compare the impact of assumption misspecification on the statistical power among the log-rank test, the RMST test, and cure rate models. Finally, we will provide recommendations for endpoint and model considerations under various scenarios.

email: Le-Rademacher.Jennifer@mayo.edu

ABSTRACTS & POSTER PRESENTATIONS

71. THE THREE M'S: MEETINGS, MEMBERSHIPS, AND MONEY!

Panel Discussion:

Jeff Goldsmith, Columbia University
Donna LaLonde, American Statistical Association
Nandita Mitra, University of Pennsylvania Perelman School of Medicine
Sarah Ratcliffe, University of Virginia

A panel discussion to educate emerging and new statisticians on how best to navigate conferences and take advantage of opportunities for networking, funding, and more through professional associations for successful careers in statistics and biostatistics.

72. RECENT ADVANCES IN JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA

Assessing Importance of Biomarkers: A Bayesian Joint Modeling Approach of Longitudinal and Survival Data with Semicompeting Risks

Ming-Hui Chen*, University of Connecticut
Fan Zhang, University of Connecticut
Xiuyu Julie Cong, Boehringer Ingelheim (China) Investment Co., Ltd.
Qingxia Chen, Vanderbilt University

Motivated from a head and neck cancer clinical trial, we develop a class of trajectory-based models for longitudinal and survival data with disease progression. Specifically, we propose a class of mixed effects regression models for longitudinal measures, a cure rate model for the disease progression (TP) time, and a time-varying covariates model for the overall survival (OS) time to account for TP and treatment switching. The properties of the proposed models are examined in details. In addition, we derive the decompositions of the deviance information criterion (DIC) and the logarithm of the pseudo marginal likelihood (LPML) to assess the fit of the longitudinal component of the model and the fit of each survival component, separately. We further develop ΔDIC and ΔLPML to determine the importance and contribution of the longitudinal data to the model fit of the TP and OS data. Moreover, efficient Markov chain Monte Carlo sampling algorithms are developed to carry out posterior computation. We apply the proposed methodology to analyze the real data from a head and neck cancer clinical trial.

email: ming-hui.chen@uconn.edu

Inference with Joint Models Under Misspecified Random Effects Distributions

Sanjoy Sinha*, Carleton University
Abdus Sattar, Case Western Reserve University

Joint models are commonly used in clinical studies for analyzing survival data with time-dependent covariates or biomarkers. It is often assumed that the latent processes that are used to describe the association between longitudinal and survival outcomes follow a multivariate normal distribution. While the joint likelihood gives valid inferences under correctly specified latent processes or random effects distributions, but the likelihood can give biased estimators and hence invalid inferences under misspecified random effects distributions. We propose a robust method to address uncertainties in random effects distributions. An empirical study demonstrates that our proposed method is able to provide consistent and efficient estimators of the model parameters under various types of misspecified random effects distributions. An application is provided using clinical data obtained from a group of sepsis patients.

email: sinha@math.carleton.ca

Personalized Decision Making for Biopsies in Prostate Cancer Active Surveillance Programs

Dimitris Rizopoulos*, Erasmus University Medical Center

Low-risk prostate cancer patients enrolled in active surveillance programs commonly undergo biopsies for examination of cancer progression. Biopsies are conducted as per a fixed and frequent schedule. Since biopsies are burdensome, patients do not always comply with the schedule, which increases the risk of delayed detection of cancer progression. Our aim is to better balance the number of biopsies (burden) and the delay in detection of cancer progression, by personalizing the decision of conducting biopsies. We use patient data from with 5270 patients, 866 cancer progressions, and an average of nine prostate-specific antigen (PSA) and five digital rectal examinations (DRE) per patient. Using joint models for time-to-event and longitudinal data, we model the historical DRE and PSA measurements, and biopsy results of a patient at each follow-up visit. This results in a visit and patient-specific cumulative risk of cancer progression. If this risk is above a certain threshold, we schedule a biopsy. We compare this personalized approach with the currently practiced biopsy schedules via an extensive and realistic simulation study.

email: d.rizopoulos@erasmusmc.nl

ABSTRACTS & POSTER PRESENTATIONS

Quantifying Direct and Indirect Effect for Longitudinal Mediator and Survival Outcome Using Joint Modeling Approach

Cheng Zheng*, University of Wisconsin, Milwaukee
Lei Liu, Washington University in St. Louis

It is usual practice to record biomarkers over time to monitor disease progression in biomedical studies. Previous work has shown that by comparing different joint models of such longitudinal biomarkers and survival outcomes, one can determine the potential causal mechanisms between the biomarkers and the survival outcomes. In this work, we further quantify the strength of such causal effect by estimating the direct and indirect effects for the treatment when the mechanism suggests that the biomarker is a mediator. The shared effects from the joint modeling framework allow us to control for the potential individual level time-independent unmeasured confounding between the biomarker process and survival time. We derived the formula and algorithm for calculating the direct and indirect effects after a joint model fitting with controlling for such unmeasured confounding. We illustrate our methods by analyzing data from two clinical trials: an AIDS study and a liver cirrhosis study.

email: zhengc@uwm.edu

73. RECENT ADVANCES IN NETWORK META-ANALYSIS WITH FLEXIBLE BAYESIAN APPROACHES

Data-Adaptive Synthesis of Historical Information through Network-Meta-Analytic-Predictive Priors

Jing Zhang*, University of Maryland
Hwanhee Hong, Duke University School of Medicine
Yong Chen, University of Pennsylvania
Cher Dallal, University of Maryland

Data-adaptive borrowing of historical information according to the consistency between the historical information and the new experimental data is gaining popularity in Bayesian clinical trial designs. It resolves the problems of reckless borrowing such as larger biases, higher type I error, and a lengthier and costlier trial, especially when prior-data conflict appears. In this article, we proposed a novel network-meta-analytic-predictive prior (NMAPP) method by incorporating a network meta-analysis element in the synthesis of historical information. Advantages of the proposed NMAPP method include that it (1) facilitates the design of multiple-arm trials; (2) avoids extracting single-arm information from randomized controlled trials; and (3) gains statistical efficiency thus further reduces sample size, cost, time and ethical hazard. Multi-component mixtures of conjugate priors are used as approximations to cope with analytic unavailability. This mixture gains robustness and offers data-adaptive borrowing. The conjugacy eases the calculation of posteriors. We illustrated the proposed methodology with case studies and simulation studies.

email: jzhang86@umd.edu

Bayesian Flexible Hierarchical Skew Heavy-Tailed Multivariate Meta Regression Models for Individual Patient Data with Applications

Sung Duk Kim*, National Cancer Institute, National Institutes of Health
Ming-Hui Chen, University of Connecticut
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Arvind K. Shah, Merck Research Laboratories
Jianxin Lin, Merck Research Laboratories

A flexible class of multivariate meta-regression models are proposed for Individual Patient Data (IPD). The methodology is well motivated from 26 pivotal Merck clinical trials that compare statins in combination with ezetimibe and statins alone on treatment-naive patients and those continuing on statins at baseline. The research goal is to jointly analyze the multivariate outcomes, LDL-C, HDL-C, and TG. These outcome measures are correlated and shed much light on a subject's lipid status. The proposed multivariate meta-regression models allow for different skewness parameters and different degrees of freedom for the multivariate outcomes from different trials under the general class of skewed t-distributions. The theoretical properties of the proposed models are examined and an efficient MCMC algorithm is developed for sampling from the posterior distribution under the proposed multivariate meta-regression model. A detailed analysis of the IPD meta data from the 26 Merck clinical trials is carried out to demonstrate the usefulness of the proposed methodology.

email: kims2@mail.nih.gov

Bayesian Network Meta-Analysis for Estimating Population Treatment Effects

Hwanhee Hong*, Duke University School of Medicine

Comparative effectiveness research relies heavily on the results of network meta-analysis (NMA) of randomized controlled trials (RCTs) to evaluate the efficacy and safety of multiple interventions. However, the NMA results may not generalize to all people in a target population of interest in which we want to make decisions regarding treatment implementation, because individual RCTs may not be representative of the target population. In this talk, we introduce NMA using Bayesian composite likelihood methods to estimate population treatment effects. First, to make RCT samples look like the population, we estimate the probability of being in RCTs given baseline characteristics and then calculate weights for all RCT participants. Second, these weights are integrated in a Bayesian network meta-analysis model with a composite likelihood function. These two steps can be conducted independently (two-step approach) or simultaneously (one-step approach), where the latter fully incorporates the uncertainty of weights. We apply these methods to generalize NMA results comparing antipsychotics treatments on schizophrenia to the US population of adults with schizophrenia.

email: hwanhee.hong@duke.edu

ABSTRACTS & POSTER PRESENTATIONS

74. ELECTRONIC HEALTH RECORDS DATA ANALYSIS

Estimating Individualized Treatment Rules for Multicategory Type 2 Diabetes Treatments Using Electronic Health Records

Jitong Lou*, University of North Carolina, Chapel Hill
Yuanjia Wang, Columbia University
Lang Li, The Ohio State University
Donglin Zeng, University of North Carolina, Chapel Hill

Type 2 diabetes (T2D) has caused severe health problems to millions of patients. In recent years, electronic health records (EHRs) and precision medicine have played important roles in recommendations of T2D treatments. However, current methods have limitations in jointly analyzing various types of patient-specific characteristics in EHRs and estimating optimal individualized treatment rules (ITRs) for multicategory treatments. In this article, we propose a latent process model to deal with data challenges in EHRs and analyze mixed-type along with correlated biomarkers in an integrative way. Furthermore, we cluster patients based on their health status, which can be captured by the latent variables. Within each patient group, we estimate optimal ITRs by extending a matched learning model (Wu et al. 2019) for comparing binary treatments to handle multicategory treatments using a one-versus-one approach. We apply our method to estimate ITRs for T2D patients in a large sample of EHRs from the Ohio State University Wexner Medical Center and show its utility to select the optimal treatments from four classes of drugs and achieve better T2D controls than any universal rules.

email: jitong@live.unc.edu

Modeling Heterogeneity and Missing Data in Electronic Health Records

Rebecca Anthopolos*, New York University
Qixuan Chen, Columbia University Mailman School of Public Health
Ying Wei, Columbia University Mailman School of Public Health

In electronic health records (EHRs), unobservable groups of patients may exhibit distinctive patterning in longitudinal health trajectories. For such data, growth mixture models (GMMs) enable classifying patients into different latent classes based on individual longitudinal trajectories and risk factors associated with class membership. However, the application of GMMs in EHRs is hindered by two patient-led missing data processes: the visit process and the response process for each EHR variable given a patient visits the clinic. If either process is associated with the underlying process for the longitudinal health outcomes, then valid inferences must account for a missing not at random mechanism. We propose a Bayesian shared parameter model that links GMMs of the longitudinal outcomes, visit process, and response process given a clinic visit using a discrete latent class variable. Our focus is on longitudinal health outcomes with a clinically prescribed visit schedule. We apply our model in EHR weight and height measurements. We show that failure to account for nonignorable visit and response processes may result in biased group-specific or population-averaged inferences.

email: Rebecca.Anthopolos@nyulangone.org

Modeling Valid Drug Dosage in the Presence of Conflicting Information Extracted from Electronic Health Records

Michael L. Williams*, Vanderbilt University Medical Center
Hannah L. Weeks, Vanderbilt University Medical Center
Cole Beck, Vanderbilt University Medical Center
Elizabeth McNeer, Vanderbilt University Medical Center
Leena Choi, Vanderbilt University Medical Center

Diverse medication-based studies including population pharmacokinetic and pharmacodynamic studies require longitudinal drug dose information. Electronic health records (EHRs) have great potential to provide such information as detailed medication dose information can be extracted from clinical notes using natural language processing (NLP) systems. However, multiple mentions of a drug in the same clinical note can yield conflicting dosages. To address this challenge, we extracted dose information for two test drugs, tacrolimus and lamotrigine, from Vanderbilt EHRs using our own NLP system, medExtractR. A random forest classifier was used to estimate the probability of correctness for each extracted dose on the basis of subject longitudinal dosing patterns and extracted EHR note context. Using this feasibility measure and other features such as subject dosing pattern, we developed statistical models to predict the true dose on the basis of the extracted doses: a separate random forest regression, a transition model, a boosting model, and a Bayesian hierarchical model. We compared model-predicted doses to physician-validated doses to evaluate model performance.

email: michael.l.williams@vanderbilt.edu

ABSTRACTS & POSTER PRESENTATIONS

Case Contamination in Electronic Health Records-Based Case-Control Studies

Jill Schnall*, University of Pennsylvania
Lu Wang, University of Pennsylvania
Scott Damrauer, University of Pennsylvania
Michael Levin, University of Pennsylvania
Jinbo Chen, University of Pennsylvania

One challenge in using electronic health records (EHRs) for research is that the true phenotype status of an individual must be derived using information present in the EHR. Phenotyping algorithms are often used to define cases and controls, but it is difficult to balance the accuracy of phenotype classification with sample size. We explore the use of an estimating equation (EE) approach that allows for more relaxed phenotype definitions and corrects the bias introduced by case contamination. The EE approach relies on drawing a validation subset from a contaminated case pool and training a phenotyping model to distinguish cases from non-cases. Through simulation studies, we assess the performance of the EE method for bias correction, evaluate the robustness of the method to specification of the phenotyping model, and evaluate the performance of the method when the phenotyping model is fit using high-dimensional data methods. Finally, we apply the method to an EHR-based study of dilated cardiomyopathy. We find that our method outperforms other methods used for bias correction and can also perform well when high-dimensional data methods are needed to fit the phenotyping model.

email: jschnall@pennmedicine.upenn.edu

Quantile Rank Test for Dynamic Heterogeneous Genetic Effect in Longitudinal Electronic Health Record Analysis

Tianying Wang*, Columbia University
Ying Wei, Columbia University
Iuliana Ionita-Laza, Columbia University
Zixu Wang, Columbia University
Chunhua Weng, Columbia University

Over the past few years, an increasing number of sequence-based association studies evaluated the group-wise effects of rare and common genetic variants and identified significant associations between a gene and a phenotype of interest. Utilizing the longitudinal trajectory of outcomes can help us explore the dynamic genetic effect and improve the test power. However, limited researches have been done in such area especially for electronic health records data. In this paper, we propose a generalized integrated rank score test based on quantile regression, which considering the quantile effect of the entire sample. A perturbation method is used to overcome the deflated/inflated type I error issue for longitudinal studies. Using simulations studies and the Electronic Medical Records and Genomics (eMERGE) Network data, we show that the proposed test complements the mean-based analysis and improves efficiency and robustness, not only to the heterogeneous associations among the population but also to the misspecification of within-subject correlation.

email: tw2696@cumc.columbia.edu

Leveraging Electronic Health Data for Embedded Pragmatic Clinical Trials within Health Care Systems: Lessons Learned from the NIH Collaboratory

Andrea J. Cook*, Kaiser Permanente Washington Health Research Institute

Pragmatic clinical trials embedded within health care systems provide an important opportunity to evaluate new interventions and treatments. Networks have recently been developed to support practical and efficient studies. Pragmatic trials will lead to improvements in how we deliver health care and promise to more rapidly translate research findings into practice. The National Institutes of Health (NIH) Health Care Systems Collaboratory was formed to conduct pragmatic clinical trials and to cultivate collaboration across research areas and disciplines to develop best practices for future studies. We will outline general themes and challenges with proposed solutions when conducting embedded pragmatic clinical trials that utilize electronic health records in the study design, unit of randomization, sample size and statistical analysis. Our findings are applicable to other pragmatic clinical trials conducted within health care systems.

email: Andrea.J.Cook@kp.org

75. REBEL WITHOUT A CAUSE: SESSIONS ON CAUSAL INFERENCE

A New Method for Estimating a Principal Stratum Causal Effect Conditioning on a Post-Treatment Intermediate Response

Xiaoqing Tan*, University of Pittsburgh
Judah Abberbock, GlaxoSmithKline
Priya Rastogi, University of Pittsburgh
Gong Tang, University of Pittsburgh

In neoadjuvant trials on early stage breast cancer, patients are usually randomized into a control group and a treatment group with an additional target therapy. Early efficacy of the new regimen is assessed via the binary pathological complete response (pCR) and the eventual efficacy is assessed via a long-term clinical outcome such as survival. Although pCR is strongly associated with survival, it has never been confirmed as a surrogate endpoint. To fully understand its clinical implication, it is important to establish causal estimands such as the causal effect in survival for patients who would achieve pCR under the new regimen. Under some mild conditions, current works under the principal stratification framework specify a model for a counterfactual response given observed data and draw conclusions on the causal estimand via sensitivity analyses by varying values of the model parameters (Shepherd et al., 2006). Here we propose to estimate those model parameters and subsequently the causal estimand using the empirical data under the same assumptions. The proposed method is applied to a recent clinical trial and its performance is evaluated via simulation studies.

email: xiaoqingtan@pitt.edu

ABSTRACTS & POSTER PRESENTATIONS

Detecting Heterogeneous Treatment Effect with Instrumental Variables

Michael W. Johnson*, University of Wisconsin, Madison
Jiongyi Cao, The University of Chicago
Hyunseung Kang, University of Wisconsin, Madison

There is an increasing interest in estimating heterogeneity in causal effects in randomized and observational studies. However, little research has been conducted to understand effect heterogeneity in an instrumental variables study. In this work, we present a method to estimate heterogeneous causal effects using an instrumental variable. The method has two parts. The first part uses subject-matter knowledge and interpretable machine learning techniques, such as classification and regression trees, to discover potential effect modifiers. The second part uses closed testing to test for statistical significance of each effect modifier while strongly controlling the the familywise error rate. We apply this method on the Oregon Health Insurance Experiment, estimating the effect of Medicaid on the number of days an individual's health does not impede their usual activities using a randomized lottery as an instrument. Our method revealed Medicaid's effect was most impactful among older, English-speaking, non-Asian males and younger, English-speaking individuals with at most high school diplomas or GEDs.

email: mwjohnson8@wisc.edu

A Groupwise Approach for Inferring Heterogeneous Treatment Effects in Causal Inference

Chan Park*, University of Wisconsin, Madison
Hyunseung Kang, University of Wisconsin, Madison

There is a growing literature on nonparametric estimation of the conditional average treatment effect given a specific value of covariates. However, this estimate is often difficult to interpret if covariates are high dimensional and in practice, effect heterogeneity is discussed in terms of subgroups of individuals with similar attributes. The paper propose to study treatment heterogeneity under a groupwise framework where effects are divided into meaningful subgroups. Our method is simple, only based on linear regression and sample splitting, and is semiparametrically efficient under assumptions. We also discuss ways to conduct multiple testing. We conclude by reanalyzing a get-out-the-vote experiment during the 2014 U.S. midterm elections.

email: cpark229@wisc.edu

Estimating Complier Quantile Causal Treatment Effects with Randomly Censored Data and A Binary Instrumental Variable

Bo Wei*, Emory University
Limin Peng, Emory University
Mei-jie Zhang, Medical College of Wisconsin
Jason Fine, University of North Carolina, Chapel Hill

The causal effect of a treatment is of fundamental interest in many biomedical studies. Instrumental variable (IV) methods are commonly used to estimate causal treatment effects in the presence of unmeasured confounding. In this work, we study a new IV framework with randomly censored outcomes to quantify complier quantile causal effect (CQCE). Compared to complier average causal effect, CQCE has better identifiability in censored outcomes, and provides more dynamic insight about the potential outcome difference under different treatments. Employing the special feature of IV and the principle of conditional score, we uncover a simple weighting scheme that can be incorporated into the standard censored quantile regression procedure to estimate CQCE. We develop robust nonparametric estimation of the derived weights in the first stage, which permits to implement the second stage estimation based on existing software. We establish rigorous asymptotic properties for the estimator. Extensive simulation studies and an application to a dataset from the Center for International Blood and Marrow Transplant Research illustrate the validity and practical utility of the proposed method.

email: bwei8@emory.edu

Causal Effects in Twin Studies: The Role of Interference

Bonnie Smith*, Johns Hopkins Bloomberg School of Public Health
Elizabeth Ogburn, Johns Hopkins Bloomberg School of Public Health
Saonli Basu, University of Minnesota
Matthew McGue, University of Minnesota
Daniel Scharfstein, Johns Hopkins Bloomberg School of Public Health

The use of twin designs to address causal questions is becoming increasingly popular. A standard assumption is that there is no interference between twins--that is, that no twin's exposure has a causal impact on their co-twin's outcome. However, there may be settings in which this assumption would not hold, and this would (1) impact the causal interpretation of parameters obtained by commonly used existing methods; (2) change which effects are of greatest interest; and (3) impact the conditions under which we may estimate these effects. We explore these issues, and we derive semi-parametric efficient estimators for causal effects in the presence of interference between twins. Using data from the Minnesota Twin Family Study, we apply our estimators to assess whether twins' consumption of alcohol in early adolescence may have a causal impact on their co-twins' substance use later in life.

email: bsmit179@jhmi.edu

ABSTRACTS & POSTER PRESENTATIONS

Causal Inference from Self-Controlled Case Series Studies Using Targeted Maximum Likelihood Estimation

Yaru Shi*, Merck & Co., Inc.
Fang Liu, Merck & Co., Inc.
Jie Chen, Merck & Co., Inc.

There has been an increasing interest in causal inference using real-world data (RWD) for regulatory and healthcare decision-making. One common objective of such practice is to provide additional insights for approved drugs in reporting post-marketing safety. This paper extends the causal inference approach for case-control studies by Rose and van der Laan (2009) to self-controlled case series (SCCS) studies. First introduced in 1995, the SCCS method uses cases as their own controls in which all time-invariant confounders are automatically controlled, rendering the possibility of the causality assessment for time-varying effects that can be efficiently carried out by using targeted maximum likelihood estimation (TMLE). The proposed approach is applied to a real-world dataset to investigate the causal relationships between an immunotherapy and a rare AE observed post therapy.

email: yaru.shi@merck.com

Caution Against Examining the Role of Reverse Causality in Mendelian Randomization

Sharon M. Lutz*, Harvard Medical School and Harvard Pilgrim Health Care Institute
Ann C. Wu, Harvard Medical School and Harvard Pilgrim Health Care Institute
Christoph Lange, Harvard T.H. Chan School of Public Health

Recently, Mendelian Randomization (MR) has gained in popularity as a concept to assess the causal relationship between phenotypes in genetic association studies. The MR Steiger approach has been proposed as a tool that claims to be able infer the causal direction between two phenotypes. Through simulation studies, we examine the ability of the MR Steiger approach to correctly determine the mediator and outcome, i.e. effect direction. Our results show that the Steiger approach generally is not able to infer causality based on study/observational data. We examine the scenarios in detail for which the MR Steiger approach is unable to correctly determine causality. We also applied several popular MR approaches and the MR Steiger method to the COPDGene study, a case-control study of chronic obstructive pulmonary disease (COPD) in current and former smokers, to examine the role of smoking on lung function. We have created an R package on Github called reverseDirection which runs simulations for user-specified scenarios in order to examine when the MR Steiger approach can correctly determine the direction of the arrow between the mediator and outcome.

email: smlutz@hsph.harvard.edu

76. HYPOTHESIS TESTING: KNOWLEDGE IS POWER

A Score Based Test for Functional Linear Concurrent Regression

Rahul Ghosal*, North Carolina State University
Arnab Maity, North Carolina State University

We propose a score based test for testing the null hypothesis of no effect of a covariate on the response in the context of functional linear concurrent regression. We establish an equivalent random effects formulation of our functional regression model under which our testing problem reduces to testing for zero variance component for random effects. For this purpose, we use a one-sided score test approach, which is an extension of the classical score test. We provide theoretical justification as to why our testing procedure has the right levels (asymptotically) under null using standard assumptions. Using numerical simulations, we illustrate that our testing method has the desired type I error rate and gives higher power compared to a bootstrapped F test currently existing in the literature. Our model and testing procedure are shown to give good performances even when the data is sparsely observed, and the covariate is contaminated with noise. Applications of the proposed testing method are demonstrated on gait study and a dietary calcium absorption data.

email: rghosal@ncsu.edu

Differential Expression Analysis in Single-Cell RNA Sequencing with G-modeling-based Two-Sample Test

Jingyi Zhai*, University of Michigan
Hui Jiang, University of Michigan

Many real data analyses involve two-sample comparisons in locations or in distributions. Most existing methods focus on problems where observations are independently identically distributed in each group. However, in some applications, such as the differential expression analysis on single-cell RNA sequencing (scRNA-seq) data, the observations may not be identically distributed. To address this challenge, we propose a new test as a combination of the g-modeling density estimation and two-sample Kolmogorov-Smirnov test, and we estimate the statistical significance using bootstrap. We model the scRNA-seq data using zero-inflated Poisson distributions. Simulations show that our proposed new test has high accuracy and statistical power. We also apply the proposed method to a real scRNA-seq dataset and compare with five other methods for differential expression analysis.

email: jyzhai@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Detect with BERET

Duyeol Lee*, University of North Carolina, Chapel Hill
Kai Zhang, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

Recently, the binary expansion testing framework was introduced to test the independence of two continuous random variables by utilizing symmetry statistics which are complete sufficient statistics for dependence. We develop a new test through an ensemble method utilizing both the sum of squared symmetry statistics and distance correlation. Simulation studies suggest that the proposed method improves the power while preserving the clear interpretation of the binary expansion testing. We further extend this method to tests of independence of random vectors in arbitrary dimension. The proposed binary expansion randomized ensemble test (BERET) transforms the multivariate independence testing problem into a univariate one through random projections. The power of the proposed method is illustrated with many simulated and real data examples.

email: duyeol@live.unc.edu

Resampling-Based Stepwise Multiple Testing Procedures with Applications to Clinical Trial Data

Jiwei He*, U.S. Food and Drug Administration
Feng Li, U.S. Food and Drug Administration
Yan Gao, The University of Illinois at Chicago
Mark Rothmann, U.S. Food and Drug Administration

The commonly used multiple testing procedures (MTPs) in clinical trials do not take into consideration the correlation between test statistics. Romano and Wolf (2005) have constructed a resampling-based stepdown method that incorporates the correlation structure of test statistics and is shown to be more powerful than the usual stepdown Holm's method. However, there is at present little experience with applications of such methods in analyzing clinical trial data. We have extended Romano and Wolf (2005)'s approach to a step-up procedure and examined the performance of both stepdown and stepup methods under a variety of correlation structures and distribution types. Results from our simulation studies support the use of the resampling-based methods under various scenarios, including binary data and small samples, with strong control of FWER. Under positive dependence and for binary data even under independence, the resampling-based methods are more powerful than the Holm's and Hochberg methods. Lastly, we illustrate the advantage of such methods with two clinical trial data examples: a cardiovascular outcome trial and an oncology trial respectively.

email: jiwei.he@fda.hhs.gov

Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models

Rong Ma*, University of Pennsylvania
T. Tony Cai, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

High-dimensional logistic regression is widely used in analyzing data with binary outcomes. In this paper, global testing and large-scale multiple testing for the regression coefficients are considered in both single- and two-regression settings. A test statistic for testing the global null hypothesis is constructed using a generalized low-dimensional projection for bias correction and its asymptotic null distribution is derived. A lower bound for the global testing is established, which shows that the proposed test is asymptotically minimax optimal. For testing the individual coefficients simultaneously, multiple testing procedures are proposed and shown to control the false discovery rate (FDR) and falsely discovered variables (FDV) asymptotically. Simulation studies are carried out to examine the numerical performance of the proposed tests and their superiority over existing methods. The testing procedures are also illustrated by analyzing a data set of a metabolomics study that investigates the association between fecal metabolites and pediatric Crohn's disease and the effects of treatment on such associations.

email: rongm@upenn.edu

Hypothesis Testing to Determine if Two Penalties Are Better Than One: Should Second Order Terms have the Same Penalty as Main Effects?

Todd A. MacKenzie*, Dartmouth College
Iben Ricket, Dartmouth College
Jiang Gui, Dartmouth College
Kimon Bekelis, Dartmouth College

Machine learning users criticize linear models because they assume linearity. This stems from the misconception that linear models cannot fit non-linear associations. At the least, the addition of second order terms (quadratic terms and interactions) should be considered when prediction is done using generalized linear models. The addition of higher order terms can lead to a feature space whose dimension rivals the sample size which statisticians often address using penalized optimands such as in LASSO or Ridge regression or Elastic Net. Typically, a single penalty parameter is employed in LASSO or Ridge regression but the use of separate penalties for different ordered terms should be explored. In this study we employ LASSO allowing for distinct penalties for first and second order terms and propose a test statistic for evaluating if two penalty parameters are superior to one. The latter is accomplished by recognizing that penalty tuning using K-fold cross-validation sum of squares (or likelihood) is similar to non-linear least squares (or maximum likelihood estimation). We report the distribution of the test statistic using theory and simulations.

email: todd.a.mackenzie@dartmouth.edu

ABSTRACTS & POSTER PRESENTATIONS

77. MISSING (DATA) IN ACTION

Identifying Treatment Effects using Trimmed Means when Data are Missing Not at Random

Alex J. Ocampo*, Harvard University

Patients often discontinue treatment in a clinical trial because their health condition is not improving. If we only analyze the patients who complete the trial, then this biases the estimator of a medication's efficacy because study outcomes are missing not at random (MNAR). One way to overcome this problem - the trimmed means approach for missing data - sets missing values as slightly worse than the worst observed outcome and then trims away a fraction of the distribution from each treatment arm before calculating differences in treatment efficacy. In this paper we derive sufficient and necessary conditions for when this approach can identify the average population treatment effect. Numerical studies show the trimmed means approach's ability to effectively estimate treatment efficacy when data are MNAR and missingness is strongly associated with an unfavorable outcome. Lastly, If the reasons for discontinuation in a clinical trial are known analysts can improve estimates with a combination of multiple imputation (MI) and the trimmed means approach when the assumptions of each missing data mechanism hold.

email: ocampo@g.harvard.edu

A Bayesian Multivariate Skew-Normal Mixture Model for Longitudinal Data with Intermittent Missing Observations: An Application to Infant Motor Development

Carter Allen*, Medical University of South Carolina
Brian Neelon, Medical University of South Carolina
Sara E. Benjamin-Neelon, Johns Hopkins Bloomberg School of Public Health

In studies of infant growth, a crucial research goal is to identify latent clusters of infants with delayed motor development — a risk factor for adverse outcomes. However, there are many statistical challenges in modeling such outcomes: the data are typically skewed, exhibit intermittent missingness, and are correlated across repeated measurements. Using data from the Nurture study, we develop a Bayesian mixture model for analysis of infant motor development data. First, we model developmental trajectories using matrix skew normal distributions with cluster-specific parameters to accommodate dependence and skewness. Second, we model the cluster membership probabilities using Pólya-Gamma data-augmentation, which improves cluster allocation. Lastly, we impute missing responses from conditional multivariate skew normal distributions. Through simulation studies, we show that the proposed model yields improved inferences over models that ignore skewness or adopt conventional imputation methods. In the Nurture data, we discovered two developmental clusters, as well as detrimental effects of food insecurity on development.

email: allecart@musc.edu

Estimation, Variable Selection and Statistical Inference in a Linear Regression Model under an Arbitrary Missingness Mechanism

Chi Chen*, State University of New York at Buffalo
Jiwei Zhao, State University of New York at Buffalo

Nonignorable missing data is a common issue in clinical studies and usually the missingness mechanism is unknown to investigators. Traditional analysis with nonignorable missing data tend to involve bias in conclusions. We propose an unconventional likelihood method for parameter estimation in linear regression analysis. Meanwhile, we consider adaptive LASSO for purpose of variable selection, with BIC to select tuning parameter. The objective function is reformed to boost computation efficiency. The method is effective under arbitrary missingness mechanism where most parameters of clinical interest are estimable. The large sample properties of estimates and variable selection are discussed. We validate our method in simulation studies by comparing results to that using fully observed data only. We apply our method to a chondral lesions and meniscus procedures (ChAMP) study.

email: chenchi0526@outlook.com

Influence Function Based Inference in Randomized Trials with Non Monotone Missing Binary Outcomes

Lamar Hunt*, Johns Hopkins Bloomberg School of Public Health
Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health

In randomized trials in which patients are to be assessed at regular intervals, patients may miss assessments in an irregular fashion, yielding non-monotone missing data patterns. While missing at random (MAR) has been considered a reasonable benchmark assumption for studies with monotone missing data patterns, Robins (1997) and Little and Rubin (2014) have argued that MAR is implausible for studies with non-monotone patterns. Towards this end, we investigate inference about the treatment-specific mean outcome at each assessment under a missing not random (MNAR) benchmark assumption. This "block-conditional" assumption, first introduced by Zhou, Little and Kalbfleisch (2010), posits that the missingness of the outcome at each assessment time depends on the past history of outcomes (observed or not). We derive the class of all influence functions for the parameter of interest. We discuss how to utilize these influence functions for inference in settings where the outcomes are binary. We also propose a global sensitivity analysis to evaluate the robustness of inference to deviations from the benchmark assumption.

email: lhunt13@jhmi.edu

ABSTRACTS & POSTER PRESENTATIONS

Multiple Imputation Variance Estimation in Studies with Missing or Misclassified Inclusion Criteria

Mark J. Giganti*, Center for Biostatistics in AIDS Research
Bryan E. Shepherd, Vanderbilt University

In observational studies using routinely collected data, a variable with a high level of missingness or misclassification may determine whether an observation is included in the analysis. In settings where inclusion criteria are assessed after imputation, the popular multiple imputation variance estimator proposed by Rubin (Rubin's rules [RR]) is biased. While alternative approaches exist, most analysts are not familiar with them. Using partially validated data from an HIV cohort, we illustrate the calculation of an imputation variance estimator proposed by Robins and Wang (RW) in a scenario where the study exclusion criteria are based on a variable that must be imputed. The corresponding imputation variance estimate for the log odds was 29% smaller using RW (0.046) relative to RR (0.065). A simulation study showed the coverage probabilities of 95% confidence intervals based on RR were too high and became worse as more observations were imputed and the number of subjects excluded from the analysis increased. The RW imputation variance estimator performed much better and should be employed when there is incompatibility between imputation and analysis models.

email: mgiganti@sdac.harvard.edu

Missing Data in Deep Learning

David K. Lim*, University of North Carolina, Chapel Hill
Naim U. Rashid, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

The field of deep learning has boomed in popularity in recent years, fueled initially by its performance in the classification and manipulation of image data, and, more recently, in areas of public health, medicine, and biology. However, the presence of missing data in these latter areas is much more common, and involves more complicated mechanisms of missingness. While a rich statistical literature exists regarding the characterization and treatment of missing data in traditional statistical models, it is unclear how such methods may extend to deep learning. In this study, we provide a formal treatment of missing data in the context of Variational Autoencoders (VAEs), and show through simulations and real data analyses that the extension of missing data is critical to ensure the performance of common tasks, such as data reconstruction and latent structure discovery in populations. We utilize several simulation examples as well as datasets from EHR and single cell data to illustrate the impact of missingness on such tasks, and compare the performance of several proposed methods in handling missing data.

email: deelim@live.unc.edu

An Approximated Expectation-Maximization Algorithm for Analysis of Data with Missing Values

Gong Tang*, University of Pittsburgh

We consider regression analysis of data with nonresponse. Standard statistical methods, including likelihood-based methods and weighted estimating equations, require correct specification of a model for the missing-data mechanism. Misspecification of the missing-data model often causes biased estimates and wrongful conclusions. The expectation-maximization (EM) algorithm is an iterative algorithm that is often used to find the maximum likelihood estimate for the likelihood-based methods. In the E-steps, given a current estimate and a model for the missing-data mechanism, the conditional expectations of the sufficient statistics are calculated. Under the premise that the current estimate is consistent, we approximate those conditional expectations from the empirical data without modeling the missing-data mechanism. The resulted algorithm can be applied to analysis of data with nonresponse regardless of the potential missing-data mechanism. The consistent initial value can be either obtained from an external complete dataset, complete recall on a subset or a robust method requiring less assumption on the missing-data model. Various versions of this algorithm will be discussed.

email: got1@pitt.edu

78. BACK TO THE FUTURE: PREDICTION AND PROGNOSTIC MODELING

High Dimensional Classified Mixed Model Prediction

Mengying Li*, University of Miami
J. Sunil Rao, University of Miami

Today, more and more practical problems involving high dimensional data are moving beyond the variable selection framework to focus on prediction at the subject or (small) sub-population level. Significant increases in prediction accuracy can be achieved by identifying a group that a new subject belongs to as was done in the Classified Mixed Model Prediction work of Jiang et al. (2018). Here, we propose a new method, called High Dimensional Classified Mixed Model Prediction, which allows classified mixed predictions for high dimensional predictors. A four-step algorithm is used which includes a mixed model variable screening step, followed by penalized estimation of both fixed and random effects, followed by re-estimation for the restricted model, followed by CMMP prediction. Importantly, this work also extends the methodology to allow for the challenging case of unknown grouping of observations. Asymptotic and empirical studies are carried out which demonstrate favorable properties of HDCMMP. Finally, an analysis of breast cancer genomic data from The Cancer Genome Atlas (TCGA) repository clearly demonstrates the utility of the new methodology in practice.

email: mengyingli80@yahoo.com

ABSTRACTS & POSTER PRESENTATIONS

Connecting Population-Level AUC and Latent Scale-Invariant R-square via Semiparametric Gaussian Copula and Rank Correlations

Debangan Dey*, Johns Hopkins Bloomberg School of Public Health
Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health

We employ Semiparametric Gaussian Copula (SGC) to model joint dependence between observed binary outcome and observed continuous predictor via correlation of latent standard normal random variables. Under SGC, we show how, both population-level AUC and latent scale-invariant R-square, defined as a squared latent correlation, can be estimated using any of the four rank statistics calculated on binary-continuous pairs: Wilcoxon rank-sum, Kendall's Tau, Spearman and Quadrant rank correlations. We then focus on three implications and applications: i) we show that under SGC, the population-level AUC and the population-level latent R-square are related via a monotone function that depends on the population-level prevalence rate, ii) we propose Quadrant rank correlation as a robust semiparametric version of AUC; iii) we demonstrate how, under complex-survey designs, Wilcoxon rank-sum statistics, Spearman and Quadrant rank correlations provide estimators of population-level AUC using only single-participant survey weights. We illustrate these applications using five-year mortality and continuous predictors from 2003-2006 National Health and Nutrition Examination Survey.

email: ddey1@jhu.edu

Artificial Intelligence and Agent-Based Modeling - Prediction and Simulation Issue

Nicolas J. Savy*, Toulouse Institute of Mathematics
Philippe Saint-Pierre, Toulouse Institute of Mathematics

Artificial Intelligence has shown wonderful power for prediction. A lot of strategies are available to predict an outcome from data. Various tools have been built to automatically perform those models (superlearner or caret R-packages). Agent-based modeling consists in a set of models whose aim to mimic the behavior of individuals in a random environment. In health context, agent-based modeling may be used to simulate the effect of a treatment on patients. To do so, virtual patients are randomly generated and models are used to predict their medical outcomes under different scenarios of treatment effects. By comparing these scenarios it is possible to derive an estimation of the effect of treatment on the medical outcome. Strategy usually called "In Silico Clinical Trial" (ISCT). In this talk we will browse the milestones of this strategy focusing our attention on the main methodological pitfall: to set up an agent-based modeling, a predictive model is not enough a sharp modeling of the error of prediction is necessary. By a series of simulation studies, we will discuss the role of the error prediction modeling on the results of the ISCT and we will quantify it.

email: Nicolas.Savy@math.univ-toulouse.fr

Improving Survival Prediction Using a Novel Feature Selection and Feature Reduction Framework Based on the Integration of Clinical and Molecular Data

Lisa Neums*, University of Kansas Medical Center and University of Kansas Cancer Center
Richard Meier, University of Kansas Medical Center and University of Kansas Cancer Center
Devin C. Koestler, University of Kansas Medical Center and University of Kansas Cancer Center
Jeffrey A. Thompson, University of Kansas Medical Center and University of Kansas Cancer Center

The accurate prediction of a cancer patient's risk of progression or death can guide clinicians in the selection of treatment. Ideally, predictive models will use multiple sources of data (e.g., clinical, molecular, etc.). However, there are many challenges associated with data integration, such as over fitting and redundant features. Here, we developed a novel feature selection and feature reduction framework that can handle correlated data. Selected genes, in combination with clinical data, are used to build a predictive model for survival. We tested our framework using kidney cancer, lung cancer and bladder cancer. Across all data sets, our approach outperformed the clinical data alone in terms of predictive power. Further, we were able to show increased predictive performance of our method compared to lasso-penalized cox proportional hazards models fit to both gene expression and clinical data, as well as increased or comparable predictive power compared to ridge regression models. Therefore, our score for clinical independence improves prognostic performance as compared to modeling approaches that do not consider combining non-redundant data.

email: lneums@kumc.edu

Quantile Regression for Prediction of High-Cost Patients

Scott S. Coggeshall*, VA Puget Sound

Accurate predictions of which patients will incur high healthcare costs are important for healthcare systems. Understanding which patients are likely to incur high costs allows for better planning and potential intervention to reduce costs. Standard classification methods require dichotomizing the measure of health care cost prior to modeling. Quantile regression provides an alternative method for obtaining binary predictions that simultaneously allows modeling the measure of cost on its original, non-binary scale. In this paper, we discuss implementing this quantile regression methodology for prediction at the scale of modern EHR systems. We show how a penalized quantile regression model with both parametric and non-parametric components can be fit using the Alternating Direction Method of Multipliers (ADMM) algorithm. The ADMM algorithm allows for straightforward distributed fitting across multiple computing resources, which is particulary important when fitting these models to EHR-scale data.

email: sscogges@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Joint Prediction of Variable Importance Rank from Binary and Survival Data via Adaptively Weighted Random Forest

Jihwan Oh*, Merck & Co., Inc.
John Kang, Merck & Co., Inc.

Discovering important genes plays a key-role in genome-based biomarker development for the treatment of cancer patients. The random forest is one of a most popular and powerful technique in predictive analyses and can provide the importance ranks of genes among many of them. In this study, we propose an extension to the random forest framework that allows each tree to be constructed over multiple heterogeneous outcomes including binary, continuous and/or survival responses. Our method adaptively finds the weight to combine different types of responses which leads to the better variable selection than single-response random forests and other multi-response random forests.

email: jihwan05@gmail.com

External Validation Study of SMART Vascular Event Prediction Model Using UK Primary Care Data Between 2000-2017

Laura H. Gunn*, University of North Carolina, Charlotte & Imperial College London
Ailsa McKay, Imperial College London
Azeem Majeed, Imperial College London
Kosh Ray, Imperial College London

Among those with atherosclerotic cardiovascular disease (ASCVD), future vascular event risk varies widely. Ability to predict individual risks enables more refined approaches to risk management. This validation study assesses performance of the SMART prediction model in predicting 10-year vascular event risks. Data from the Clinical Practice Research Datalink consists of adults registered with UK National Health Service primary care providers diagnosed with coronary, cerebrovascular, peripheral, and/or aortic ASCVD. Exposure variables include demographics, medical history, and clinical measurements, while the outcome is first post cohort-entry occurrence of myocardial infarction, stroke, or cardiovascular death. The calibration and discrimination achieved by the SMART model was not dissimilar to performance at internal validation. It slightly under-predicted risk among lower risk groups, but clinical utility was apparent across potential treatment thresholds. Results remained consistent in sensitivity analyses. The SMART model has utility in the context of routine UK primary care-based secondary prevention of cardiovascular disease.

email: laura.gunn@uncc.edu

79. M&M: MEASUREMENT ERROR AND MODELING

Statistical Analysis of Data Reproducibility Measures

Zeyi Wang*, Johns Hopkins Bloomberg School of Public Health
Eric Bridgeford, Johns Hopkins Bloomberg School of Public Health
Joshua T. Vogelstein, Johns Hopkins University
Brian Caffo, Johns Hopkins Bloomberg School of Public Health

In the field of functional magnetic resonance imaging, it is crucial, yet challenging, to quantify the reproducibility of the generated data because of its high dimensional nature and the unusually complex measuring and preprocessing procedures. Novel data reproducibility measures have been brought up in the context where a set of subjects are measured twice or more, including fingerprinting, rank sums, discriminability, and generalizations of the intraclass correlation. However, the relationships between and the best practices among these measures remains largely unknown. In this manuscript, we systematically analyze the most natural reproducibility statistics associated with different statistical models. We show that the rank sum statistic is deterministically linked to an estimator of discriminability. We theoretically prove the relation between discriminability and the intraclass correlation under random effect models, with univariate or multivariate measurements. The power of permutation tests derived from these measures are compared numerically under Gaussian and non-Gaussian settings, and with batch effects. Recommendations are given for each setting accordingly.

email: zwang107@gmail.com

An Approximate Quasi-Likelihood Approach to Analyzing Error-Prone Failure Time Outcomes and Exposures

Lillian A. Boe*, University of Pennsylvania
Pamela A. Shaw, University of Pennsylvania

Measurement error arises commonly in clinical research settings that rely on data from electronic health records or large observational cohorts. In particular, self-reported outcomes are typical in cohort studies for chronic diseases such as diabetes in order to avoid the burden of expensive diagnostic tests. Dietary intake, which is also commonly collected by self-report and prone to error, is a major factor linked to diabetes and a number of other chronic diseases. These errors can bias exposure-disease associations that ultimately can mislead clinical decision-making. We have extended an existing semiparametric likelihood-based method for handling error-prone, discrete failure time outcomes to also address covariate measurement error. We conduct a numerical study to evaluate the proposed method in terms of bias and efficiency in the estimation of the regression parameter of interest. This method is applied to data from the Women's Health Initiative in order to assess the association between energy and protein intake and the risk of incident diabetes, correcting for the errors in both the self-reported outcome and dietary exposures.

email: boel@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Improving the Efficiency of Generalized Raking Estimators to Address Correlated Covariate and Failure-Time Outcome Error

Eric J. Oh*, University of Pennsylvania
Thomas Lumley, University of Auckland
Bryan E. Shepherd, Vanderbilt University
Pamela A. Shaw, University of Pennsylvania

Large clinical studies, particularly those utilizing electronic health records, can be subject to correlated measurement error in covariates and a failure time outcome. Instead of reviewing all medical records to correct errors, we can utilize a two-phase design and perform data validation on a subset to inform statistical methods that adjust estimates for the error structure. For such two-phase designs, generalized raking has been shown to be a robust method that yields consistent and asymptotically normal estimates; however, there are aspects of the raking estimators that can be improved to yield more efficient estimates. Namely, the assumption of a linear relationship between the target of inference and auxiliary variables and utilization of efficient sampling designs to select the phase two subjects are two areas we explore. We present numerical studies of the relative efficiency of various choices of the auxiliary variable and phase two sampling strategy for the generalized raking estimator under varying level of censoring, error structure, validation subset size, and strength of association. We further examine our proposed approaches with an application to real data.

email: ericoh@pennmedicine.upenn.edu

Impact of Design Considerations in Sensitivity to Time Recording Errors in Pharmacokinetic Modeling

Hannah L. Weeks*, Vanderbilt University
Matthew S. Shotwell, Vanderbilt University

Pharmacokinetic models are important clinical tools used to monitor drug response and modify treatment in the presence of individual heterogeneity. Such models rely on accurately recorded time data, in particular when blood draws are taken or medication is administered. Using an approximate Bayesian two compartment model for an intravenously administered antibiotic, we estimate patient-specific summaries of target attainment. We perform a simulation study to compute estimates in the presence of blood draw or infusion time errors. In particular, we focus on design considerations such as length of infusion (30 minutes or 4 hours), number of blood draws, and relative time at which blood draws are taken. We then compute bias between estimates with and without timing errors. Preliminary results show that for a single blood draw, longer infusions are less susceptible to bias with time recording errors. Additionally, blood draws taken during an infusion lead to more bias in estimates than those taken not during an infusion. By understanding conditions which lead to more bias in the presence of time recording errors, researchers can develop more robust pharmacokinetic studies.

email: hannah.l.weeks@vanderbilt.edu

Surrogate-Assisted Subsampling in Logistic Regression with Outcome Misclassification

Chongliang Luo*, University of Pennsylvania
Arielle Marks-Anglin, University of Pennsylvania
Yong Chen, University of Pennsylvania

Association analysis using Logistic Regression in EHR data often faces the outcome misclassification problem. The misclassified outcome, namely the surrogate outcome, can lead to biased estimation of the effect size. To quantify the misclassification and improve the estimation, chart review is usually conducted for a small subsample. We develop an approach for chart review Optimal Subsampling Assisted by the Surrogate Outcome (OSASO). The resulting estimator is asymptotically unbiased and have improved efficiency compared to the existing methods. The performance of the proposed approach is demonstrated by simulation and real data examples.

email: luocl3009@gmail.com

---

80. PRESIDENTIAL INVITED ADDRESS

Medical Product, Healthcare Delivery, and Road Safety Policies: Seemingly Unrelated Regulatory Questions

Sharon-Lise Normand, Ph.D., S. James Adelstein Professor of Health Care Policy (Biostatistics), Department of Health Care Policy, Harvard Medical School, Department of Biostatistics, Harvard T.H. Chan School of Public Health

The evaluations of medical product effectiveness and safety, the quality of hospital care, and the safety of U.S. roadways involve the use of large, complex observational data to make policy decisions. Careful design and analysis of such data are critical given the large populations impacted. While increasing access to data of increased size and type permit, in theory, richer evaluations, study design should assume a more prominent role. This talk will describe three different policy problems: the impact of the hospital readmission reduction program, the effectiveness of seemingly similar drug eluting coronary stents, and the safety of U.S. motor carriers. Statistical issues common across these problems, including clustered data, multiple treatments, multiple outcomes, high-dimensional data, and lack of randomization, are highlighted and solutions discussed.

e-mail: sharon@hcp.med.harvard.edu

ABSTRACTS & POSTER PRESENTATIONS

81. STATISTICAL ANALYSIS OF BIOLOGICAL SHAPES

Manifold-Valued Data Analysis of Brain Networks

Ian L. Dryden*, University of Nottingham
Simon P. Preston, University of Nottingham
Katie E. Severn, University of Nottingham

Networks are used in many biomedical applications, for example to indicate associations between brain regions or genes. It is of interest to develop statistical methodology in situations where samples of networks are available. We represent networks as graph Laplacians and compare the use of different metrics to perform statistical analysis. We define embeddings, tangent spaces and a projection from Euclidean space into the space of graph Laplacians. Using the metrics and projection we provide a general framework to perform extrinsic statistical analysis, such as calculating means, performing principal component analysis and regression. We also develop a hypothesis test for the equality of means between two samples of networks, and investigate its power. We apply the methodology to samples of resting-state brain networks.

email: ian.dryden@nottingham.ac.uk

Shape Analysis for Mitochondria Data

Todd Ogden*, Columbia University
Ruiyi Zhang, Florida State University
Martin Picard, Columbia University
Anuj Srivastava, Florida State University

For many biological objects, the shape of the object is related to its functionality. We describe a framework for analysis of two-dimensional shapes based on distances between objects considered to exist in a non-Euclidean shape space. For this there are several metrics available, and based on these metrics, it is possible to study the variability of the distribution of shapes, analogous to standard ANOVA in Euclidean space . We apply this approach to data on shapes of mitochondria taken from skeletal muscles of mice and address the question of which of several factors are related to these shapes.

email: to166@columbia.edu

Geometric Methods for Image-Based Statistical Analysis of Shape and Texture of Glioblastoma Multiforme Tumors

Sebastian Kurtek*, The Ohio State University
Karthik Bharath, University of Nottingham
Veera Baladandayuthapani, University of Michigan
Arvind Rao, University of Michigan

Biomedical studies are a common source of rich and complex imaging data. Statistical analysis of such data requires novel methods due to two main challenges: 1. the functional nature of the data objects under study and 2. the nonlinearity of their representation spaces. We consider the task of quantifying and analyzing two different types of tumor heterogeneity. The first, which is represented by a probability density function (pdf), summarizes the tumor's texture information. We use the nonparametric Fisher-Rao Riemannian framework to define intrinsic statistical methods on the space of pdfs for summarization and inference. The second type, which is represented by a parameterized, planar closed curve, captures the tumor's shape information. A key component of analyzing tumor shapes is a suitable metric that enables efficient comparisons, provides tools for computing descriptive statistics and implementing principal component analysis on the tumor shape space. We demonstrate the utility of our framework on a dataset of Magnetic Resonance Images of patients diagnosed with Glioblastoma Multiforme, a malignant brain tumor with poor prognosis.

email: kurtek.1@stat.osu.edu

Fiber Bundles in Probabilistic Models

Lorin Crawford*, Brown University
Bruce Wang, Princeton University
Timothy Sudijono, Brown University
Henry Kirveslahti, Duke University
Tingran Gao, The University of Chicago
Doug M. Boyer, Duke University
Sayan Mukherjee, Duke University

The recent curation of large-scale databases with 3D surface scans of shapes has motivated the development of tools that better detect global-patterns in morphological variation. Studies which focus on identifying differences between shapes have been limited to simple pairwise comparisons and rely on pre-specified landmarks (that are often known). We present SINATRA: the first probabilistic pipeline for analyzing collections of shapes without requiring any correspondences. A key insight is that, one can use invertible tools from differential topology to transform objects represented as meshes into a collection of vectors (with little to no loss of information about their natural structure). Our novel algorithm takes in two classes of shapes and highlights the physical features that best describe the variation between them. We use a rigorous simulation framework to assess our approach. Lastly, as a case study, we use SINATRA to analyze mandibular molars from four different suborders of primates and demonstrate its ability recover known morphometric variation across phylogenies.
email: lorin_crawford@brown.edu

ABSTRACTS & POSTER PRESENTATIONS

IMPROVING THE DEVELOPMENT AND VALIDATION OF SCREENING TESTS FOR RARE DISEASES

From Prediction to Policy: Risk Stratification to Improve the Efficiency of Early Detection for Cancer

Ruth Etzioni*, Fred Hutchinson Cancer Research Center

In recent years a lot of energy has been devoted to discovering risk strata for cancer incidence and outcome, based on demographics, egermline genomics and even biomarker measurements. The rationale behind these efforts is that understanding which subsets of the population are at higher risk of disease can inform targeted screening policies that focus efforts and resources on the higest-risk strata. But all too often the line from risk stratification to screening policy is not a straight one. First, the objective of risk-based screening policies must be clearly articulated. In practice, ithe preferred policy will depend on the mechanism underlying the observed heterogeneity in the risk of disease diagnosis across population strata. . In this presentation we show how statistical and simulation models of disease natural history can be used to connect the dots from prediction to policy. We consider specifically the case of risk stratification for prostate cancer screening based on information about race, germline BRCA status, and baseline PSA at age 45.

e-mail: retzioni@fredhutch.org

A Simple Framework to Identify Optimal Cost-Effective Risk Thresholds for a Single Screen: Comparison to Decision Curve Analysis

Hormuzd Katki*, National Cancer Institute, National Institutes of Health
Ionut Bebu, The George Washington University

Decision Curve Analysis (DCA) is a popular approach for assessing biomarkers, but does not require costs and thus cannot identify optimal risk thresholds. Full decision analyses can identify optimal thresholds, but typically used methods are complex and often hard to understand. We develop a simple framework to calculate the Incremental Net Benefit for a single-time screen as a function of costs (for tests and treatments) and effectiveness (life-years gained). We provide simple expressions for the optimal cost-effective risk-threshold and, equally importantly, for the monetary value of life-years gained associated with the risk-threshold. We consider the controversy over the risk-threshold to screen women for mutations in BRCA1/2. Importantly, most, and sometimes even all, of the thresholds identified by DCA are infeasible based on their associated dollars per life-year gained. Our simple framework facilitates sensitivity analyses to cost and effectiveness parameters. Our approach estimates optimal risk thresholds in a simple and transparent manner, provides intuition about which quantities are critical, and may serve as a bridge between DCA and a full decision analysis.

email: katkih@mail.nih.gov

Sample Weighted Semiparametric Estimation of Cause-Specific Cumulative Risk and Incidence Using Left or Interval-censored Data from Electronic Health Records

Noorie Hyun*, Medical College of Wisconsin
Hormuzd A. Katki, National Cancer Institute, National Institutes of Health
Barry I. Graubard, National Cancer Institute, National Institutes of Health

Electronic health records (EHRs) can be a cost-effective data source for forming cohorts and developing risk models in the context of disease screening. However, important issues need to be handled: competing outcomes, left-censoring of prevalent disease, interval-censoring of incident disease, and uncertainty of prevalent disease when accurate disease ascertainment is not conducted at baseline. Furthermore, novel tests that are costly and limited in availability can be conducted on stored biospecimens selected as samples from EHRs by using different sampling fractions. We propose sample-weighted semiparametric mixture models for estimating cause-specific risks. And a numerical algorithm for nonparametrically calculating the maximum likelihood estimates for subdistribution hazard functions and regression parameters. We apply our methods to a cohort assembled from EHRs at a health maintenance organization where we estimate cumulative risk of cervical pre-/cancer and incidence of infection-clearance by HPV genotype among human papilloma virus (HPV) positive women.

email: nhyun@mcw.edu

A Statistical Review: Why Average Weighted Accuracy, not Accuracy or AUC?

Qing Pan*, The George Washington University
Yunyun Jiang, The George Washington University
Scott Evans, The George Washington University

Sensitivity and specificity are key aspects in evaluating the performance of diagnostic tests. Accuracy and AUC are commonly used composite measures that incorporate sensitivity and specificity. AWA is motivated by the need for a statistical measure of diagnostic yield that can be used to compare diagnostic tests from the medical costs and clinical impact point of view, while incorporating the relevant prevalence range of the disease as well as the relative importance of false positive versus false negative cases. We derive testing procedures in four different scenarios: (i) one diagnostic test vs. the best random test, (ii) two diagnostic tests from two independent samples, (iii) two diagnostic tests from the same sample, and (iv) more than two diagnostic tests from different or the same samples. The impacts of sample size, prevalence, and relative importance on power and average medical costs/clinical loss are examined through simulation studies. The use of AWA is illustrated on a three-arm clinical trial evaluating three different assays in detecting Neisseria gonorrhoeae (NG) and Chlamydia trachomatis (CT) in the rectum and pharynx.

email: panqing94@yahoo.com

ABSTRACTS & POSTER PRESENTATIONS

83. CAUSAL INFERENCE AND HARMFUL EXPOSURES

Envisioning Hypothetical Interventions on Occupational Exposures to Protect Worker Health: Applications of the Parametric G-formula

Andreas M. Neophytou*, Colorado State University

Assessing potential harmful effects of occupational exposures on health outcomes is often hampered by the healthy worker survivor bias, in the form of time-varying confounding by underlying health (or employment) status which is in turn affected by previous exposure. Whereas traditional regression approaches are ill suited to tackle this issue, g-methods are well equipped to address the issue of time-varying confounding affected by previous exposure within a potential outcomes framework. We demonstrate applications of one such method, the parametric g-formula, in separate occupational epidemiology studies: particulate matter exposures and lung function in the aluminum industry, and occupational crystalline silica exposure and mortality in the diatomaceous earth industry. In both settings we envision hypothetical interventions to lower harmful occupational exposures and assess the reduction in risk in the outcome of interest under these interventions compared to no intervention (what actually happened). Required assumptions for causal inferences are discussed, while approaches to account for censoring and competing events are also presented.

email: andreas.neophytou@colostate.edu

A Causal Inference Framework for Cancer Cluster Investigations Using Publicly Available Data

Rachel C. Nethery*, Harvard T.H. Chan School of Public Health
Yue Yang, Harvard T.H. Chan School of Public Health
Anna J. Brown, The University of Chicago
Francesca Dominici, Harvard T.H. Chan School of Public Health

In response to a notification of high cancer rates in a community, the CDC recommends performing a standardized incidence ratio (SIR) analysis to test whether the observed cancer incidence is higher than expected. We instead propose a causal inference approach to cancer cluster investigations. Assuming that a source of hazard in the community is identified a priori, we introduce a new estimand called the causal SIR (cSIR). The cSIR is a ratio defined as the expected cancer incidence in the exposed population divided by the expected cancer incidence under the counterfactual scenario of no exposure. To estimate the cSIR we overcome two challenges: 1) identify unexposed populations similar to the exposed one to inform estimation under the counterfactual scenario of no exposure, and 2) make inference on cancer incidence in these unexposed populations using publicly available data that are available at a higher level of spatial aggregation than desired. We overcome the first challenge by applying matching and the second by developing a hierarchical model that borrows information from other sources to impute cancer incidence at the desired finer level of spatial aggregation.

email: rnethery@hsph.harvard.edu

Estimating the Effects of Precinct Level Policing Policies Through Causal Inference with Interference

Joseph Antonelli*, University of Florida
Brenden Beck, University of Florida

Due to a number of controversial police tactics, there have been calls for police reform to tackle discriminatory policing. Frequently, policing changes are adopted at the precinct level while precincts self-select to follow the new recommendations or not. There is substantial interest in evaluating the impacts of these policies both in the treated precinct's jurisdiction, but also in neighboring areas. To study these questions, we adopt novel statistical approaches to estimate the effects of policy changes within a precinct, as well as potential and unintended spillover effects into surrounding areas. We apply our approach to a comprehensive data set of police and crime activity for the New York City Area over the years 2000-2015, to evaluate the effectiveness of a number of policy changes that have been adopted over the years.

email: jantonelli@ufl.edu

Exploring Evidence of Residual Confounding in Tropical Cyclone Epidemiology Using a Negative Exposure Control Analysis

Brooke Anderson*, Colorado State University
Meilin Yan, Peking University

Natural disasters can severely impact human health. Several study designs, including time series and case-crossover designs, are commonly used to investigate the health risks associated with ambient environmental exposures that vary from day to day. However, for disaster exposures with severe and lasting health effects, these study designs may result in bias if the period of potential disaster effects is mis-specified in the model, as study days affected by the disaster could be mis-specified as control days. The use of controls from other years can prevent this, but may in turn introduce residual confounding from long-term trends in exposure patterns and the health outcome of interest. To test for such residual confounding, we demonstrate the use of a negative exposure control analysis. This control analysis is applied to a multi-storm analysis of the association between tropical cyclone exposure and Medicare hospitalizations, in which we conduct a negative control analysis based on using the date two weeks before each true storm exposure as the negative controls, finding little evidence of residual confounding by long-term trends in the case study.

email: brooke.anderson@colostate.edu

ABSTRACTS & POSTER PRESENTATIONS

84. STATISTICAL METHODS FOR EMERGING DATA IN ENVIRONMENTAL HEALTH RESEARCH

Bayesian Joint Modeling of Chemical Structure and Dose Response Curves

Kelly R. Moran*, Duke University
David Dunson, Duke University
Amy H. Herring, Duke University

The Toxic Substances Control Act, enacted in 1976, regulates new and existing chemicals in the United States. Today there are approximately 85,000 chemicals on the list, with around 2,000 new chemicals introduced each year. It is impossible to screen all of these chemicals for potential toxic effects. Our goal is to accurately predict chemical toxicity based on chemical structure alone by taking advantage of a large database of chemicals that have already been tested in high-throughput screening programs. Additionally, we aim to learn a distance between chemicals targeted to toxicity. Our Bayesian partially Supervised Sparse and Smooth Factor Analysis (BS3FA) model is able to predict a functional response given some set of descriptors that may be continuous, binary, or count. Simultaneously, a low dimensional representation of the descriptors is learned that, via supervision from the functional response, is decomposed into response-relevant and response-irrelevant components. A distance metric can be generated based on those features important to the functional response. An R package for the method is available online at https://github.com/kelrenmor/bs3fa.

email: krmoran@g.clemson.edu

Source-Specific Exposure Assessment by using Bayesian Spatial Multivariate Receptor Modeling

Eun Sug Park*, Texas A&M Transportation Institute

A major difficulty with assessing source-specific health effects is that source-specific exposures cannot be measured directly; rather, they need to be estimated by a source apportionment method such as Positive Matrix Factorizations (PMF). The uncertainty in estimated source-specific exposures (source contributions) has been largely ignored in previous studies. Also, most previous studies examining health effects of source-specific air pollution have used monitor-specific estimated source contributions as an indicator of individual exposures, which are subject to non-ignorable spatial misalignment error. We present a Bayesian spatial multivariate receptor modeling (BSMRM) approach that incorporates spatial correlations in multisite multipollutant data into the estimation of source composition profiles and contributions. The BSMRM can predict unobserved source-specific exposures at any location and time along with their uncertainty, which can greatly reduce spatial misalignment errors. The proposed method is illustrated with real multipollutant data obtained from multiple monitoring stations. Maps of estimated source-specific exposures are also presented.

email: e-park@tti.tamu.edu

The Impact of Complex Social and Environmental Mixtures on Educational Outcomes in Young Children

Kathy B. Ensor*, Rice University
Mercedes Bravo, Research Triangle Institute and Rice University
Daniel Kowal, Rice University
Henry Leong, Rice University
Marie Lynn Miranda, Rice University

Although it is widely agreed that child health and well-being are determined by multiple forces, surprisingly little is known about the interactions of those forces. Environmental exposures often cumulate in particular geographies, and the nature of the complex mixtures that characterize these exposures remains understudied. In addition, adverse environmental exposures often occur in communities facing multiple social stressors such as deteriorating housing, inadequate access to health care, poor schools, high unemployment, crime, and poverty. We assess the impact of complex mixtures of both social stress and environmental exposures on the educational outcomes of young children evaluated at the population level for the State of North Carolina. Our analytical approach involves models that leverage the unique spatio-temporal dataset that we have constructed, as well as data reduction methods to determine the key drivers of educational outcomes. The analytical approach is designed to identify productive strategies for targeted interventions to achieve improved outcomes for children.

email: ensor@rice.edu

Accounting for Mixtures in Risk Assessment

Chris Gennings*, Icahn School of Medicine at Mount Sinai

Fundamental to regulatory guidelines is to identify chemicals that are implicated with adverse human health effects and inform public health risk assessors about acceptable ranges of such environmental exposures (e.g., consumer products, pesticides). The process is made more difficult when accounting for complex human exposures to multiple environmental chemicals. We will describe a new class of nonlinear statistical models for human data that incorporate and evaluate regulatory guideline values into analyses of health effects of exposure to chemical mixtures. The method will be illustrated using prenatal concentrations of mixtures of 11 chemicals with suspected endocrine disrupting properties and two health effects: birth weight and language delay at 2.5 years. Based on the application of this new method we conclude that the guideline values need to be lower than those for single chemicals when the chemicals are observed in combination to achieve a similar level of protection as was aimed for the individual chemicals. The proposed modeling may thus suggest data-driven uncertainty factors for single chemical risk assessment that takes environmental mixtures into account.

email: chris.gennings@mssm.edu

ABSTRACTS & POSTER PRESENTATIONS

85. BAYESIAN ANALYSIS IN FUNCTIONAL BRAIN IMAGING

Functional Regression Methods for Functional Neuroimaging

Jeffrey Scott Morris*, University of Pennsylvania
Hongxiao Zhu, Virginia Tech University
Michelle Miranda, University of Victoria
Neel Desai, Rice University
Veera Baladandayuthapani, University of Michigan
Philip Rausch, Humboldt University

Much of neuroimaging data can be considered as complex functional data objects characterized by various types of potentially nonstationary spatial and temporal correlation. Many common analytical approaches either do not model the entire data set but only extracted summaries, or model the entire data set but do not flexibly account for its characteristic structure. Either approach has statistical consequences, with missed discoveries resulting from reductionistic results and/or inaccurate uncertainty quantification. In this talk, we discuss the adaptation of functional regression methods designed for structured complex functional data for some common types of analyses for neuroimaging data, including event-related potential data from electroencephalography or magnetoencephalography, simulataneously determining differential activation and functional connectivity in fMRI data, and connectivity regression with fMRI data to assess how functional connectivity across brain regions varies with covariates.

email: Jeffrey.morris@pennmedicine.upenn.edu

A Grouped Beta Process Model for Multivariate Resting-State EEG Microstate Analysis on Twins

Mark Fiecas*, University of Minnesota
Brian Hart, UnitedHealthGroup
Stephen Malone, University of Minnesota

EEG microstate analysis investigates the collection of distinct temporal blocks that characterize the electrical activity of the brain. We propose a Bayesian nonparametric model that estimates the number of microstates and their underlying behavior. We use a Markov switching vector autoregressive (VAR) framework, where a hidden Markov model (HMM) controls the non-random state switching dynamics of the EEG activity and a VAR model defines the behavior of all time points within a given state. We analyze resting state EEG data from twin pairs collected through the Minnesota Twin Family Study. We fit our model at the twin pair level, sharing information within epochs from the same participant and within epochs from the same twin pair. We capture within twin pair similarity by using a Beta process Bernoulli process to consider an infinite library of microstates and allowing each participant to select a finite number of states from this library. The state spaces of highly similar twins may completely overlap while dissimilar twins could select distinct state spaces. In this way, our Bayesian nonparametric model defines a sparse set of states which describe the EEG data.

email: mfiecas@umn.edu

Bayesian Analysis of Multidimensional Functional Data

John Shamshoian*, University of California, Los Angeles
Donatello Telesca, University of California, Los Angeles
Damla Senturk, University of California, Los Angeles

Multi-dimensional functional data arises in numerous modern scientific experimental and observational studies. In this paper we focus on longitudinal functional data, a structured form of multidimensional functional data. Operating within a longitudinal functional framework we aim to capture low dimensional interpretable features. We propose a computationally efficient nonparametric Bayesian method to simultaneously smooth observed data, estimate conditional functional means and functional covariance surfaces. Statistical inference is based on Monte Carlo samples from the posterior measure through adaptive blocked Gibbs sampling. Several operative characteristics associated with the proposed modeling framework are assessed comparatively in a simulated environment. We illustrate the application of our work in two case studies. The first case study involves age-specific fertility collected over time for various countries. The second case study is an implicit learning experiment in children with Autism Spectrum Disorder (ASD).

email: jshamsho@gmail.com

Encompassing Semiparametric Bayesian Inference for Stationary Points in Gaussian Process Regression Models with Applications to Event-Related Potential Analysis

Meng Li*, Rice University
Cheng-Han Yu, Rice University
Marina Vannucci, Rice University

Stationary points embedded in the derivatives are often critical for a model to be interpretable and may be considered as key features of interest in many applications. We propose a semiparametric Bayesian model to efficiently infer the locations of stationary points of a nonparametric function, while treating the function itself as a nuisance parameter. We use Gaussian processes as a flexible prior for the underlying function and impose derivative constraints to control the function's shape via conditioning. We develop an encompassing strategy to bypass the daunting task to specify the number of stationary points. We show a generalized Bernstein–von Mises theorem for the posterior distribution of stationary points, which converges to Gaussian mixtures under the total variation distance, allowing convenient inference in practice. In an application to analyzing event-related potentials (ERP) derived from electroencephalography (EEG) signals, our proposed method automatically identifies characteristic components and their latencies at the individual level, which avoids excessive averaging across subjects that is routinely done in the field to obtain smooth curves.

email: xylimeng@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

86. HUMAN DATA INTERACTION: GAINING AN UNDERSTANDING OF THE DATA SCIENCE PIPELINE

Tools for Analyzing R Code the Tidy Way

Lucy D'Agostino McGowan*, Wake Forest University

With the current emphasis on reproducibility and replicability, there is an increasing need to examine how data analyses are conducted. In order to analyze the between researcher variability in data analysis choices as well as the aspects within the data analysis pipeline that contribute to the variability in results, we have created two R packages: matahari and tidycode. These packages build on methods created for natural language processing; rather than allowing for the processing of natural language, we focus on R code as the substrate of interest. The matahari package facilitates the logging of everything that is typed in the R console or in an R script in a tidy data frame. The tidycode package contains tools to allow for analyzing R calls in a tidy manner. We demonstrate the utility of these packages as well as walk through two examples.

email: lucydagostino@gmail.com

Domain Specific Languages for Data Science

Hadley Wickham*, RStudio

Domain specific languages (DSLs) for data science (e.g. ggplot2, dplyr, rvest) provide flexible environments tailored for specific challenges. I'll discuss why I think DSLs strike the right balance between helping the user find the golden path of success while still giving them the ability to go off-roading when it's really needed.

email: h.wickham@gmail.com

The Challenges of Analytic Workflows: Perspectives from Data Science Educators

Sean Kross*, University of California, San Diego

As modern data science practices spread throughout academic and industrial research institutions, the demand for training in these methods is growing to new heights. At the same time the range of tools and depth of technical expertise required to successfully execute a data analysis appears to be increasingly overwhelming and difficult to navigate. To explore these expanding tensions we reflect on a study of 20 data science educators working in academia and industry. Despite the widely varying settings in which these instructors teach we found that: 1) instructors must scaffold the correct mental models, while also providing supportive programming environments, so students can properly navigate the data science technology stack, 2) instructors highly value teaching authentic workflows which integrate code, data, and communication, 3) significant challenges exist for instructors when helping their students cope with the uncertainty inherent in doing data analytic work. The results from this study can inform the design of future data analysis tools, and they support new paradigms for teaching data science across domains.

e-mail: seankross@ucsd.edu

Building a software package in tandem with machine learning methods research can result in both more rigorous code and more rigorous research

Nick Strayer*, Vanderbilt University

Often a machine learning research project has the following timeline: the researcher begins brainstorming, writes one-off scripts while the idea forms, and - once they are ready to publish - writes a software package to disseminate the product. This talk advocates rethinking this process by spreading software development across the entire research process. When used from the start of a project, proper development processes for building a package can produce not only robust code but robust research. I will demonstrate these principles using a newly developed research product: a native R package written to fit and investigate the results of Stochastic Block Models for uncertainty-aware clustering. By going over the ups and downs of this process, I hope to leave the audience with inspiration for moving the package writing process closer to the start of their projects and melding research and code more closely to improve both.

e-mail: n.strayer@vanderbilt.edu

---

87. SPATIAL AND SPATIAL-TEMPORAL DATA ANALYSIS

Bayesian Spatial-Temporal Accelerated Failure Time Models for Survival Data from Cancer Registries

Ming Wang*, The Pennsylvania State University
Zheng Li, Novartis
Lijun Zhang, The Pennsylvania State University
Yimei Li, University of Pennsylvania
Vern M. Chinchilli, The Pennsylvania State University

Prostate cancer is the most common cancer among U.S. men, and the second leading cause of cancer death in U.S. men. The incidence rate and mortality vary substantially across geographical regions and over time. The widely-used Cox Proportional Hazards model does not apply due to the violation of the proportional hazards assumption. In this work, we propose to fit Bayesian accelerated failure time (AFT) models to analyze prostate cancer survival and take spatial-temporal variation into account by incorporating random effects with multivariate conditional autoregressive priors. The parameter estimation and inference are based on the Monte Carlo Markov Chain technique under the Bayesian framework. Extensive simulations are performed to examine and compare the performances of various Bayesian AFT candidate models with goodness of fit check via the deviance information criterion. Finally, we apply our method into the 2004-2014 Pennsylvania Prostate Cancer Registry data which includes newly diagnosed prostate cancer patients and their vital status.

email: mwang@phs.psu.edu

ABSTRACTS & POSTER PRESENTATIONS

Where did All the Good Fish Go? Spatio- Temporal Modelling of Research Vessel Data with R

Ethan Lawler*, Dalhousie University
Joanna Mills Flemming, Dalhousie University

A major problem with modern fishing is bycatch -- incidental catch of undesired species -- which can have devastating effects on at risk or endangered species. In addition to fishing gear restrictions, one of the main tools used to combat excessive bycatch is a fisheries closure (or a marine protected area). A fisheries closure is a restriction on fishing activity in a spatial area, either fixed or changing through time. Assessing the potential effectiveness of proposed fisheries closures requires a spatio-temporal species distribution map for both target species and bycatch species. We present the staRVe R package developed to quickly fit flexible hierarchical spatio-temporal models to large point-referenced datasets. Covariate effects in the mean and covariance structures, non-Gaussian responses, and forecasting are all supported. The sf and raster R packages are tightly integrated to support existing workflows. We showcase the package with the analysis of spatially-referenced research vessel data on the Scotian Shelf spanning roughly 40 years.

email: lawlerem@dal.ca

Assessing Meteorological Drivers of Air Pollution in the Eastern United States via a Bayesian Quantile Regression Model with Spatially Varying Coefficients

Stella Coker Watson Self*, University of South Carolina
Christopher S. McMahan, Clemson University
Brook Russell, Clemson University
Derek Andrew Brown, Clemson University

Meteorological covariates such as windspeed and air temperature are known to influence spikes in air pollution levels. This work develops a Bayesian quantile regression model to assess the effect of such covariates on spikes in PM2.5, or particulate matter smaller than 2.5 micrometers in diameter. As the relationship between the response and the covariates is known to change with location, spatially varying coefficients are used to generate an effect surface over the entire study area for each covariate. Various spline-based estimators are compared for these effect surfaces. Spatio-temporal random effects are included in the model to account for the spatio-temporal dependence in the data. The effect of imposing different temporal correlation structures on these random effects is explored. Regions of significance are identified for each covariate.

email: stellaw@clemson.edu

Spatio-Temporal Mixed Effects Single Index Models

Hamdy F. F. Mahmoud*, Virginia Tech and Assiut University, Egypt
Inyoung Kim, Virginia Tech

The goal of this paper is to introduce semiparametric single index models for spatially-temporally correlated data. Two models are introduced. In the first model, spatial effects are integrated into the single index function and temporal effects are additive to the single index function, and in the second model, both spatial and temporal effects are additive to the single index function. Two algorithms based on Monte Carlo Expectation Maximization (MCEM) algorithm are introduced to estimate the model's parameters, spatial effects and temporal effects. Several simulation studies are conducted to evaluate the performance of the proposed algorithms. The advantage of our proposed models is demonstrated using of six major cities mortality data in South Korea (Seoul, Busan, Daegu, Incheon, Gwangju, and Daejeon) that cover eight years, from 2000 to 2007. Interesting results are found and the two proposed models are compared to select the appropriate model for this data set. It is found that Busan has the highest spatial mortality and the second model is more appropriate in terms of fitting and prediction.

e-mail: ehamdy@vt.edu

Bayesian Spatial Blind Source Separation via Thresholded Gaussian Processes

Ben Wu*, University of Michigan
Ying Guo, Emory University
Jian Kang, University of Michigan

The goal of blind source separation (BSS) is to separate latent source signals from their mixtures. The typical methods of BSS include principal components analysis (PCA), singular value decomposition (SVD) and independent component analysis (ICA). For spatially dependent signals such as neuroimaging data, most existing BSS methods do not directly account for spatial dependence among signals and do not explicitly model the sparsity of signals. To address the limitations, we propose a Bayesian nonparametric model for BSS of spatial processes. We assume observed data as a linear mixture of multiple sparse and piece-wise smooth latent source processes, for which we construct a new class of prior models by thresholding Gaussian processes. Under regularity conditions, we show the prior enjoys large support; and we establish the consistency of the posterior distribution with a divergent number of voxels in images. We demonstrate that the proposed method outperforms existing BSS methods for brain network separation and brain activation region detection via simulations. We apply the proposed method to rs-fMRI data, showing a promising recovery of latent brain functional networks.

e-mail: bewu@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Incorporating Spatial Structure into Bayesian Spike-and-Slab Lasso GLMs

Justin M. Leach*, University of Alabama at Birmingham
Inmaculada Aban, University of Alabama at Birmingham
Nengjun Yi, University of Alabama at Birmingham

Spike-and-slab priors model predictors as coming from one of two distributions: predictors that should (slab) or should not (spike) be included in the model. The spike-and-slab lasso (SSL) model is a variant of this framework where the prior distributions for each part of the mixture are double exponential, which allows for more flexible penalties to be imposed on parameters, rather than the single penalty of the lasso. This framework has been extended for use in Generalized Linear Models, specifically for application to genetic studies. However, while these extensions can handle large numbers of potentially highly correlated predictors, spatially dependent data may provide more useful information if spatial structure is incorporated into the model. One approach is to incorporate spatial structure into the prior probabilities of inclusion in the model, which affect the shrinkage of parameter estimates; thus, parameters spatially near to each other will be more likely to have similar shrinkage penalties. We develop and test an EM algorithm to fit SSL GLM's for spatially dependent predictors where the spatial structure is modeled with conditional autoregressions.

e-mail: jleach@uab.edu

88. EARLY PHASE CLINICAL TRIALS AND BIOMARKERS

Building an Allostatic Load Scale using Item Response Theory

Shelley H. Liu*, Icahn School of Medicine at Mount Sinai
Kristen Dams-O'Connor, Icahn School of Medicine at Mount Sinai
Julie Spicer, Icahn School of Medicine at Mount Sinai

Allostatic load is a latent construct representing poor health and dysregulation across multiple physiological systems. It is commonly calculated as a sum-score, by summing the number of biomarkers that fall into a high-risk category. However, this method implies that biomarkers are equally informative for constructing the allostatic load index. Our goals were to 1) evaluate whether this is true, and 2) identify an abbreviated set of biomarkers that are most informative. We analyzed data from the 2015-2016 National Health Examination and Nutrition Survey, using twelve biomarkers that together measure the inflammatory, cardiovascular, lipid and glucose physiological systems. Using a 2 parameter logistic item response theory (IRT) model, we evaluated each biomarker's ability to discriminate across different levels of the latent trait. Body-mass-index and C-reactive protein provided the most discrimination. Although there was a general monotonic relationship between the sum-score and the IRT score, differences remained because the IRT score also depended on the biomarker characteristics (e.g. discrimination) and response pattern.

e-mail: shelley.liu@mountsinai.org

Subgroup-Specific Dose Finding in Phase I Clinical Trials Based on Time to Toxicity Allowing Adaptive Subgroup Combination

Andrew G. Chapple*, Louisiana State University
Peter F. Thall, University of Texas MD Anderson Cancer Center

A Bayesian design is presented that does precision dose finding based on time to toxicity in a phase I clinical trial with two or more patient subgroups. The design, called Sub-TITE, makes sequentially adaptive subgroup-specific dose-finding decisions while possibly combining subgroups that have similar estimated dose-toxicity curves. Decisions are made based on partial follow up information similar to the TITE-CRM. Spike-and-slab clustering priors are assumed for subgroup parameters, with latent subgroup combination variables included in the logistic model to allow different subgroups to be combined for dose finding if they are homogeneous. A simulation study shows that, when the dose-toxicity curves differ between all subgroups, Sub-TITE has superior performance compared with applying the TITE-CRM while ignoring subgroups and has slightly better performance than applying the TITE-CRM separately within subgroups. When two or more subgroups are truly homogeneous, the Sub-TITE design is substantially superior to running separate trials within all subgroups.

email: achapp@lsuhsc.edu

Evaluation of Continuous Monitoring Approach in Early Phase Oncology Trial

Suhyun Kang*, Eli Lilly and Company
Jingyi Liu, Eli Lilly and Company

In early phase oncology trials, a typical approach to interim analysis is reviewing patients' data at a few specified numbers of patients. This approach may prevent us from stopping development of futile treatment earlier or getting efficacious treatment to market faster. In fast-paced Early phase oncology trials, agility in decision making is key to utilize assets and more importantly to ensure patients' right and well-being. In this presentation, we evaluate a continuous monitoring approach as a guidance for decision making in early phase oncology trial. The continuous monitoring approach uses more frequent review of the data and it allows flexible interim schedule and quick go/no go decision. We perform simulation studies for various scenarios of efficacy (inefficacious with few responders, intermediate, and overwhelming efficacy) and quantify saving in the number of patients and time spent on clinical trials using the continuous monitoring approach. We assess agreement level between the interim decision based on the continuous monitoring approach and the final decision made at the end of clinical trial. Comparison studies with conventional two stage design will be conducted.

email: kang_suhyun@lilly.com

ABSTRACTS & POSTER PRESENTATIONS

PA-CRM: A Continuous Reassessment Method for Pediatric Phase I Trials with Concurrent Adult Trials

Yimei Li*, University of Pennsylvania
Ying Yuan, University of Texas MD Anderson Cancer Center

Pediatric phase I trials are usually carried out after the adult trial has started, but not completed yet. As the pediatric trial progresses, in light of the accrued interim data from the concurrent adult trial, the pediatric protocol often is amended to modify the original pediatric dose escalation design. This frequently is done in an ad hoc way, interrupting patient accrual and slowing down the trial. We develop a pediatric continuous reassessment method (PA-CRM) to streamline this process, providing a more efficient and rigorous method to find the MTD for pediatric phase I trials. We use a discounted joint likelihood of the adult and pediatric data, with a discount parameter controlling information borrowing between pediatric and adult trials. According to the interim adult and pediatric data, the discount parameter is adaptively updated using the Bayesian model averaging method. We examine the PA-CRM through simulations, and compare it with the two alternative approaches, which ignore adult data completely or simply pool it together with the pediatric data. The results demonstrate that the PA-CRM has good operating characteristics and is robust to various assumptions.

email: liy3@email.chop.edu

Two-Stage Enrichment Clinical Trial Design with Adjustment for Misclassification in Predictive Biomarkers

Yong Lin*, Rutgers University
Weichung Joe Shih, Rutgers University
Shou-En Lu, Rutgers University

A two-stage enrichment design is a type of adaptive design, which extends a stratified design with a futility analysis on the marker negative cohort at the first stage, and the second stage can be either a targeted design with only the marker positive stratum, or still the stratified design with both marker strata, depending on the result of the interim futility analysis. In this paper we consider the situation where the marker assay and the classification rule are possibly subject to error. We derive the sequential tests for the global hypothesis as well as the component tests for the overall cohort and the marker-positive cohort. We discuss the power analysis with the control of the type-I error rate and show the adverse impact of the misclassification on the powers. We also show the enhanced power of the two-stage enrichment over the one-stage design, and illustrate with examples of the recent successful development of immunotherapy in non-small-cell lung cancer.

e-mail: linyo@rutgers.edu

Incorporating Real-World Evidence or Historical Data to Improve Phase I Clinical Trial Designs

Yanhong Zhou*, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center
J. Jack Lee, University of Texas MD Anderson Cancer Center

Incorporating historical data or real-world evidence has great potential to improve the success of phase I clinical trials and accelerate drug development. We propose a general framework to achieve this goal, with a particular focus on the model-assisted designs, including mTPI, BOIN and keyboard designs. In contrast to model-based designs (such as the CRM), for which prior information can be conveniently incorporated through specifying a “skeleton”, i.e., the prior estimate of toxicity probability at each dose, little work has been done for incorporating available historical data into the model-assisted designs. The proposed approach takes a similar approach as the CRM skeleton, combined with the notion of effective sample size, to allow users to incorporate prior information into the model-assisted designs. Simulation study shows that the resulting designs, BOIN in particular, are capable of effectively incorporating prior information to improve the operating characteristics of the designs, yet maintain the simplicity of model-assisted designs. An interactive application has been developed to facilitate the use of the proposed methodology.

e-mail: yanzhou03@gmail.com

Density Estimation Based on Pooled Biomarkers using Dirichlet Process Mixtures

Zichen Ma*, University of South Carolina

Measuring biomarkers based on pooled specimens whereby individual specimens are combined and measurements taken using the pooled specimens is often not only more efficient than measuring each individual, but of practical necessity as well. A difficulty arises if one wants to estimate the distribution of the biomarkers at the individual level. Mathematically, a pooled biomarker is a c-fold convolution of individual biomarkers where c denotes the pool size. Hence estimating individual biomarker distribution based on pooled biomarkers is considered as a deconvolution problem. Following the seminal work on nonparametricdeconvolution by Fan (1991), most existing approaches estimate the latent individual density using kernel-based methods. In this work, we propose an alternative approach from the Bayesian perspective in which the prior over the distribution at the individual level is a Dirichlet process mixture.

email: zichen@email.sc.edu

ABSTRACTS & POSTER PRESENTATIONS

89. Electronic health records DATA ANALYSIS AND META-ANALYSIS

The Impact of Covariance Priors on Arm-Based Bayesian Network Meta-Analyses with Binary Outcomes

Zhenxun Wang*, University of Minnesota
Lifeng Lin, Florida State University
James S. Hodges, University of Minnesota
Haitao Chu, University of Minnesota

Bayesian analyses with the arm-based network meta-analysis (AB-NMA) model require researchers to specify a prior distribution for the covariance matrix of the study-specific treatment effects. The commonly-used conjugate prior for the covariance matrix, the inverse-Wishart (IW) distribution, has several limitations. For example, although the IW distribution is often described as weakly-informative, it may in fact provide strong information when some variance components are close to zero. Also, the IW prior generally leads to underestimation of correlations between treatments, which are critical for borrowing strength across treatment arms to produce less biased estimates. Alternatively, several separation strategies can be considered. To study the IW prior's impact on NMA results and compare it with separation strategies, we did simulation studies under different missing-treatment mechanisms. A separation strategy with appropriate priors for the correlation matrix and variances performs better than the IW prior. Finally, we re-analyzed three case studies and illustrated the importance of sensitivity analyses with different prior specifications when performing AB-NMA.

email: wang6795@umn.edu

A Bayesian Multivariate Meta-Analysis of Prevalence Data

Lianne Siegel*, University of Minnesota
Kyle Rudser, University of Minnesota
Siobhan Sutcliffe, Washington University School of Medicine
Alayne Markland, University of Alabama at the Birmingham VA Medical Center
Linda Brubaker, University of California, San Diego
Sheila Gahagan, University of California, San Diego
Ann E. Stapleton, University of Washington
Haitao Chu, University of Minnesota

When conducting a meta-analysis involving prevalence data for an outcome with several subtypes, each of them is typically analyzed separately using a univariate meta-analysis model. Recently, multivariate meta-analysis models have been shown to correspond to a decrease in bias and variance for multiple correlated outcomes compared to univariate meta-analysis, when some studies only report a subset of the outcomes. In this article, we propose a novel Bayesian multivariate random effects model to account for the natural constraint that the prevalence of any given subtype cannot be larger than that of the overall prevalence. Extensive simulation studies show that this new model can reduce bias and variance when estimating subtype prevalences in the presence of missing data, compared to standard univariate and multivariate random effects models. The data from a rapid review on occupation and lower urinary tract symptoms by the Prevention of Lower Urinary Tract Symptoms (PLUS) Research Consortium are analyzed as a case study to estimate the prevalence of urinary incontinence and several incontinence subtypes among women in suspected high risk work environments.

email: siege245@umn.edu

An Augmented Estimation Procedure for EHR-based Association Studies Accounting for Differential Misclassification

Jiayi Tong*, University of Pennsylvania
Jing Huang, University of Pennsylvania
Jessica Chubak, Kaiser Permanente Washington Health Research Institute
Xuan Wang, Zhejiang University
Jason H. Moore, University of Pennsylvania
Rebecca Hubbard, University of Pennsylvania
Yong Chen, University of Pennsylvania

The ability to identify novel risk factors for health outcomes is a key strength of EHR-based research. However, the validity of such studies is limited by error in EHR-derived phenotypes. The objective of this study was to develop a novel procedure for reducing bias in estimated associations between risk factors and phenotypes in EHR. The proposed method combines the strengths of a gold-standard phenotype obtained through manual chart review for a small subset of patients and an automatically-derived phenotype available for all patients but is potentially error-prone (hereafter referred to as algorithm-derived phenotype). An augmented estimator of associations is obtained by optimally combining these two phenotypes. We conducted simulation studies and analysis with data from Kaiser Permanente Washington to evaluate the performance of the augmented estimator. Compared to the estimator using validation data only, the augmented estimator has lower variance. Compared to the estimator using the algorithm-derived phenotypes, the augmented estimator has smaller bias. The proposed estimator can effectively improve analyses of risk factors using EHR data.

email: Jiayi.Tong@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Testing Calibration of Risk Prediction Models Using Positive-Only EHR Data

Lingjiao Zhang*, University of Pennsylvania
Yanyuan Ma, The Pennsylvania State University
Daniel Herman, University of Pennsylvania
Jinbo Chen, University of Pennsylvania

Conventional calibration and predictive performance assessment of risk prediction models rely on an annotated validation set. Annotating the validation set requires labor intensive manual chart review. Hereby we develop new methods for assessing model calibration and prediction accuracy using "positive-only" EHR data, which is a cohort sample that consists of a group of labeled cases and a large number of unlabeled patients. For the unlabeled patients, we first provide model-free estimate for the number of cases in each risk region, then construct a calibration statistic by aggregating differences between the model-free and model-based number of cases across risk regions. We show that the proposed statistic follows a Chi-squared distribution provided that the labeled cases are representative of all cases. We also propose consistent estimators for predictive accuracy measures and derive their asymptotic properties. We demonstrate performance of the proposed methods through extensive simulation studies and apply them to Penn Medicine EHRs to validate the preliminary risk prediction models for primary aldosteronism.

e-mail: lingjiao@pennmedicine.upenn.edu

Bias Reduction Methods for Propensity Scores Estimated from Mismeasured EHR-Derived Covariates

Joanna Grace Harton*, University of Pennsylvania
Rebecca A. Hubbard, University of Pennsylvania
Nandita Mitra, University of Pennsylvania

As use of electronic health records (EHR) to estimate treatment effects has become widespread, concern about bias due to error in EHR-derived covariates has also grown. While methods exist to address measurement error in individual covariates, little prior research has investigated the implications of using propensity scores for confounder control when propensity scores are constructed from both accurate and mismeasured covariates. We used simulation studies to compare alternative approaches of accounting for error in propensity scores, including regression calibration, efficient regression calibration, monte carlo regression calibration, two-stage calibration, and multiple imputation. We compared performance for varying scenarios for link function, validation sample and main sample size, strength of confounding, and error structures in the mismeasured covariate. We applied these approaches to a real-world EHR-based comparative effectiveness study of alternative treatments for urothelial cancer. Based on our comparison of measurement error correction methods for propensity scores, we provide recommendations for best practices for EHR-based studies with imperfect covariates.

e-mail: jograce@pennmedicine.upenn.edu

Bayesian Network Meta-Regression for Partially Collapsed Ordinal Outcomes: Latent Counts Approach

Yeongjin Gwon*, University of Nebraska Medical Center
Ming-Hui Chen, University of Connecticut
Mo May, Amgen Inc.
Xun Jiang, Amgen Inc.
Amy Xia, Amgen Inc.
Joseph Ibrahim, University of North Carolina, Chapel Hill

In this talk, we propose a Bayesian network-meta regression approach for modeling partially collapsed ordinal outcomes under logit link. Specifically, we develop regression model based on aggregate trial-level covariates for the variances of the random effects to capture heterogeneity across trials. We also use latent counts to account for uncertainty in the counts which specific values are unknown, while the bounds of these counts are known. An efficient Markov chain Monte Carlo sampling algorithm is developed to carry out Bayesian computation based on these latent counts. We further develop Bayesian model assessment measure to evaluate the goodness-of-fit. A case study demonstrating the usefulness of the proposed methodology is carried out using aggregate partially collapsed ordinal outcomes data from multiple clinical trials for treating Crohn's disease.

e-mail: yeongjin.gwon@unmc.edu

Efficient and Robust Methods for Causally Interpretable Meta-Analysis: Transporting Inferences From Multiple Randomized Trials to a Target Population

Issa J. Dahabreh*, Brown University
Jon A. Steingrimsson, Brown University
Sarah E. Robertson, Brown University
Lucia C. Petito, Northwestern University
Miguel A. Hernán, Harvard University

For many comparative effectiveness questions, evidence is available from multiple randomized trials. Synthesizing the trials' results with standard meta-analysis methods produces estimates that do not have a clear causal interpretation when each trial samples eligible individuals from a different underlying population and treatment effects vary across populations. We present identification results for causally interpretable meta-analyses that combine information from a collection of randomized trials to draw causal inferences for a target population from which experimental data are not available. We propose doubly robust estimators for potential outcome means and average treatment effects in the target population that use covariate, treatment, and outcome data from the collection of trials, but only covariate data from a sample of the target population. We study the large-sample properties of the estimators, examine their finite-sample properties in simulation studies, and demonstrate their implementation using data from a multi-center clinical trial.

e-mail: issa_dahabreh@brown.edu

ABSTRACTS & POSTER PRESENTATIONS

90. SMALL THINGS THAT MAKE A BIG DIFFERENCE: MICROBIOME ANALYSIS

Robust Inter-Taxa Dependency Estimation for High-Dimensional Microbiome Data

Arun A. Srinivasan*, The Pennsylvania State University
Danning Li, Jilin University
Lingzhou Xue, The Pennsylvania State University
Xiang Zhan, The Pennsylvania State University

A fundamental problem in microbiome analysis is studying the dependence relationships between microbial taxa in different cohorts. There is an ever-growing list of literature on the link between taxa and public health issues such as obesity and inflammatory bowel disease (IBD). Thus, it is paramount to model the microbial ecosystem. Microbiome data is littered with many statistical challenges such as high-dimensionality, compositional structure, and outliers. We robustly estimate the relationships between taxa in the presence of extreme values through the board class of elliptical distributions. We propose the Composition Adjusted Thresholding for Heavy Tailed Data (Heavy COAT) method to model the dependence structure. Heavy COAT first constructs a robust estimator of the dependence structure through the use of Tyler's M-estimator. The second step employs an l1-penalization method to induce sparsity to account for high-dimensionality. We evaluate our method through numerical studies. We apply Heavy COAT to mucosal microbiome data of individuals with IBD to investigate effect of antibiotics on microbial health and the relationship between bacterial phyla and IBD risk factors.

e-mail: uus91@psu.edu

Analysis of Compositions of Microbiomes with Bias Correction

Huang Lin*, University of Pittsburgh
Shyamal Das Peddada, University of Pittsburgh

Differential abundance (DA) analysis of microbiome data continues to be a challenging problem due to the complexity of the data. Numerous methods have been proposed in the literature and there have been misunderstandings and controversies, in part because there is a lack of clarity on what parameters are to be tested and what hypotheses a given method/statistic are really testing. A major hurdle in performing analyses of such data is the bias introduced by differences in the sampling fractions across samples. We propose a novel method called Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC), which estimates the unknown sampling fractions and corrects the bias induced by it. The resulting absolute abundance data are modeled using a linear regression framework. This formulation makes a fundamental advancement in the field because, unlike any of the existing methods, it (a) provides statistically valid test with appropriate p-values, (b) controls the False Discovery Rate (FDR), (c) maintains adequate power, (d) provides biologically meaningful results, and (e) is computationally simple to implement.

e-mail: HUL40@pitt.edu

Zero-Inflated Poisson Factor Model with Application to Microbiome Absolute Abundance Data

Tianchen Xu*, Columbia University
Ryan T. Demmer, University of Minnesota
Gen Li, Columbia University

Dimension reduction of high-dimensional microbiome data facilitates subsequent analysis such as regression and clustering. Most existing reduction methods cannot fully accommodate the special features of the data such as count-valued and excessive zero reads. We propose a zero-inflated Poisson factor analysis (ZIPFA) model in this article. The model assumes that microbiome absolute abundance data follow zero-inflated Poisson distributions with library size as offset and Poisson rates negatively related to the inflated zero occurrences. The latent parameters of the model form a low-rank matrix consisting of interpretable loadings and low-dimensional scores which can be used for further analyses. We develop an efficient and robust expectation-maximization (EM) algorithm for parameter estimation. We demonstrate the efficacy of the proposed method using comprehensive simulation studies. The application to the Oral Infections, Glucose Intolerance and Insulin Resistance Study (ORIGINS) provides valuable insights into the relation between subgingival microbiome and periodontal disease.

email: tx2155@columbia.edu

Zero-Inflated Topic Models for Human Microbiome Data

Rebecca A. Deek*, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

In recent years the human microbiome, and the role it plays in human health and disease, has increasingly become of scientific interest. Such data is known to be both compositional and zero-inflated making its analysis using currently available methods difficult. Topic modeling is an unsupervised classification technique commonly used in natural language processing to determine latent groupings, or topics, of text documents. When applied to microbiome data such latent topics can be viewed as potentially meaningful underlying microbial community structures. One such topic model is Latent Dirichlet Allocation (LDA). We show that in its current form the LDA model does not accurately capture the characteristic zero-inflation found in microbiome data and propose modifications to do so. We evaluate and compare model performance using real human microbiome data sets.

email: rebecca.deek@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Modeling of Microbiome Count Data for Network Analysis

Qiwei Li*, University of Texas, Dallas
Shuang Jiang, Southern Methodist University
Xiaowei Zhan, University of Texas Southwestern Medical Center

Constructing a microbial interaction network provides key insights into the complex biological mechanism of how the microbiome community functions relevant to human health. An enormous amount of metagenomic sequencing data made available by next generation sequencing (NGS) technology, which motivates the development of analytical tools to quantify the microbial interactions. In this work, we consider an association network represented by an undirected graph where vertices (i.e. microbial tax) are linked through edges (i.e. taxa-taxa interactions). We propose a Bayesian hierarchical model to identify a set of such interactions. The first level is a multivariate count generative model that links the raw counts in each sample to their normalized abundances and a sample-specific normalized factor. We use Dirichlet process as a flexible nonparametric mixing distribution to estimate those normalized factors. The second level is a Gaussian graphical model with a feature selection scheme for identifying those taxa-taxa interactions in a homogeneous disease and case group, respectively. A comprehensive comparative analysis based on synthetic data is conducted.

email: qiwei.li@utdallas.edu

Sparse Kernel RV for Identifying Genomic Features Related to Microbiome Community Composition

Nanxun Ma*, University of Washington
Anna Plantinga, Williams College
Michael C. Wu, Fred Hutchinson Cancer Research Center

Identification of genomic features related to microbiome composition can provide important biological insights into the relationship between microbiome, genomics, and diseases. Existing approaches for detecting associations between two sets of high-dimensional data may fail for microbiome data which incorporate phylogenetic or ecologically relevant structure. We propose a sparse kernel RV feature selection method to identify genomic features that are associated with overall microbiome compensation. This method is based on constructing KRV statistic by using kernels frameworks for both genomic and microbiome data. For genomic data, we construct modified kernels by adding a new vector of feature specific weights with a penalty so that KRV statistic is maximized under constraints, and for microbiome data, construct fixed, ecologically relevant kernels without breaking its structures. Results show that sparse KRV method is comparable to sparse correlation method with normal data, but also works on highly sparse, microbiome count data when adapting kernel structure based on Unifrac distances. We illustrate our approach on real data containing microbiome and omics data.

email: nanxunma@uw.edu

A Bayesian Semiparametric Approach to Wild-Type Distribution Estimation: Accounting for Contamination and Measurement Error (BayesACME)

Will A. Eagan*, Purdue University
Bruce A. Craig, Purdue University

Antimicrobial resistance is a major challenge to modern medicine and of grave concern to public health. To monitor this resistance, organizations analyze “drug/bug” collections of clinical assay results with the goal of estimating the distribution of susceptible (wild-type) strains and its prevalence. This estimation is challenging for two primary reasons. First, the collection of assay results is a mixture of susceptible and resistant (non-wild-type) strains. Second, the most commonly-used dilution assay produces interval-censored readings. Various methods have been proposed, each utilizing the assay readings below a threshold to estimate the wild-type prevalence and distribution. These methods; however, do not account for a third challenge, the inherent assay variability that has been shown to encompass a three-fold dilution range. To account for this, we propose a Bayesian semiparametric method that incorporates measurement error in the model and also extend allows estimation of prevalence over time. The feasibility of this approach and its improved precision is demonstrated through a simulation study and an application to a real data set.

email: weagan@purdue.edu

91. STATISTICAL GENETICS: SEQUENCING DATA ANALYSIS

lncDIFF: A Novel Quasi-Likelihood Method for Differential Expression Analysis of Non-Coding RNA

Qian Li*, University of South Florida
Xiaoqing Yu, Moffitt Cancer Center
Ritu Chaudhary, Moffitt Cancer Center
Robbert J. Slebos, Moffitt Cancer Center
Christine Chung, Moffitt Cancer Center
Xuefeng Wang, Moffitt Cancer Center

Long non-coding RNA (lncRNA) expression data have been increasingly used in finding diagnostic and prognostic biomarkers in cancer studies. Existing RNA-Seq differential analysis tools for do not effectively accommodate low abundance genes.We investigated the distribution pattern of low expression genes in lncRNAs and mRNAs, and proposed a new method implemented in R package lncDIFF to detect differentially expressed (DE) lncRNA genes. lncDIFF adopts the generalized linear model with zero-inflated Exponential quasi-likelihood to estimate group effect on normalized counts, and employs the likelihood ratio test to detect DE genes. Simulation results showed that lncDIFF was able to detect low abundance DE genes with more power and lower false discovery rate compared to other differential analysis tools. Analysis on a head and neck squamous cell carcinomas study showed that lncDIFF had higher sensitivity in identifying novel lncRNA genes with relatively large fold change and prognostic value. lncDIFF is a powerful differential analysis tool for low abundance lncRNA, available at https://github.com/qianli10000/lncDIFF.

email: qian.li@epi.usf.edu

ABSTRACTS & POSTER PRESENTATIONS

ASEP: Gene-based Detection of Allele-Specific Expression in a Population by RNA-seq

Jiaxin Fan*, University of Pennsylvania Perelman School of Medicine
Jian Hu, University of Pennsylvania Perelman School of Medicine
Chenyi Xue, Columbia University
Hanrui Zhang, Columbia University
Muredach P. Reilly, Columbia University
Rui Xiao, University of Pennsylvania Perelman School of Medicine
Mingyao Li, University of Pennsylvania Perelman School of Medicine

Allele-specific expression (ASE) analysis, which quantifies the relative expression of two alleles in a diploid individual, is powerful for identifying cis-regulated gene expression variations that underlie phenotypic differences among individuals. Existing gene-level ASE detection methods analyze one individual at a time, therefore wasting shared information across individuals. However, ASE detection across individuals is challenging because the data often include individuals that are either heterozygous or homozygous for the unobserved cis-regulatory SNP, leading to heterogeneity in ASE. To model multi-individual information, we developed ASEP, a mixture model with subject-specific random effect accounting for within gene multi-SNP correlations. ASEP is able to detect gene-level ASE under one condition and differential ASE between two conditions (e.g., pre- versus post-treatment). Extensive simulations have demonstrated ASEP's convincing performance under a variety of scenarios. We further applied ASEP to RNA-seq data of human macrophages, and identified genes showing differential ASE pre- versus post-stimulation.

email: jiaxinf@pennmedicine.upenn.edu

A Sparse Negative Binomial Classifier with Covariate Adjustment for RNA-seq Data

Md Tanbin Rahman*, University of Texas MD Anderson Cancer Center
Hsin-En Huang, National Tsing Hua University
An-Shun Tai, National Tsing Hua University
Wen-Ping Hsieh, National Tsing Hua University
George Tseng, University of Pittsburgh

Supervised machine learning methods have been increasingly used in biomedical research and in clinical practice. In transcriptomic applications, RNA-seq data have become dominating and have gradually replaced traditional microarray due to its reduced background noise and increased digital precision. Most existing machine learning methods are, however, designed for continuous intensities of microarray and are not suitable for RNA-seq count data. In this paper, we develop two negative binomial models via generalized linear model framework. We will first introduce the model sNBLDAGLM with regularization for genes and then its extension to the model sNBLDAGLM.sC with double regularization for gene and covariate sparsity to accommodate three key elements: adequate modeling of count data with overdispersion, gene selection and adjustment for covariate effect. The two proposed sparse negative binomial classifiers are evaluated in simulations and two real applications using cervical tumor miRNA-seq data and schizophrenia post-mortem brain tissue RNA-seq data to demonstrate its superior performance in prediction accuracy and feature selection.

email: MDR56@pitt.edu

A Functional Regression Based Approach for Gene-Based Association Testing of Quantitative Trait in Family Studies

Chi-Yang Chiu*, University of Tennessee Health Science Center

We propose a functional linear mixed models (FLMM) based approach for gene-based association testing of quantitative traits in family studies. The association between a quantitative trait and multiple genetic variants adjusted for covariates is modeled under the framework of linear mixed models. The effects of genetic variants are modeled through fixed effects and a variance component. The fixed effects of genetics variants is modeled by using functional data analysis (FDA) techniques. A variance component is used to model the correlation due to identity by state (IBS). One more variance component is used to model the correlation between family members that is usually called identity by decent (IBD). To identify the genetic association, the fixed genetic effects and the IBS variance component are testing simultaneously. The bootstrap method is used to compute the p-values for likelihood ratio (LR) and F-type test statistics. We evaluate the performance of the proposed method with extensive simulations and an analysis of a real data set.

email: chiuchiyang@gmail.com

A Systematic Evaluation of Single-Cell RNA-seq Imputation Methods

Wenpin Hou*, Johns Hopkins University
Zhicheng Ji, Johns Hopkins University
Hongkai Ji, Johns Hopkins University
Stephanie C. Hicks, Johns Hopkins University

The rapid development of single-cell RNA-sequencing (scRNA-seq) technology, with increased sparsity compared to bulk RNA-sequencing (RNA-seq), has led to the emergence of many methods for preprocessing, including imputation methods. Here, we systematically evaluated the performance of 17 state-of-the-art scRNA-seq imputation methods using cell line and tissue data measured across experimental platforms, including plate-based and UMI protocols. Specifically, we assessed the similarity of imputed cell profiles to bulk samples and investigated whether methods preserved biological signals (or introduced false signals) in three downstream analysis: clustering, differential analysis and pseudotime inference. Broadly, we found significant variability in the performance of the methods across evaluation settings. Specifically, some methods recover true biological signal in gene-level analysis, but they can also introduce false signals. However, most methods failed to improve results from clustering and trajectory analysis and thus should be used with caution. Of the methods considered, kNN-smoothing, scVI, MAGIC and SAVER outperformed the other methods most consistently.

e-mail: wp.hou3@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

A Comprehensive Evaluation of Preprocessing Methods for Single-Cell RNA Sequencing Data

Shih-Kai Chu*, Vanderbilt University Medical Center
Qi Liu, Vanderbilt University Medical Center
Yu Shyr, Vanderbilt University Medical Center

Normalization and batch correction are critical steps in the preprocessing single cell RNA sequencing (scRNA-seq) data. Although numerous methods have already been developed, there is no guidance for choosing the appropriate procedures in different scenarios. We benchmarked 23 procedures combining normalization and batch correction methods in multiple scenarios, which considered relative magnitude of batch effects compared to biological effects and imbalanced cell compositions. The performance was evaluated by the capabilities to reduce batch effects, as well as to recover biological effects of interest. Our results show that batch effects can be removed at least by 60% on most procedures when they are not confounded with biological effects and they are not the major contributor to the variations. When batch effects confound with biological effects, the performance depends on their underlying mathematical models of batch correction (i.e., using linear model, or nonlinear transformation). The performance assessment of popular scRNA-seq preprocessing procedures can serve as a guideline to help users select the best method in the different scenarios.

e-mail: shih-kai.chu@vumc.org

Fast Clustering for Single-Cell RNA-seq Data using Mini-Batch k-Means

Stephanie C. Hicks*, Johns Hopkins Bloomberg School of Public Health
Ruoxi Liu, Johns Hopkins University
Yuwei Ni, Weill Cornell Medical College
Elizabeth Purdom, University of California, Berkeley
Davide Risso, University of Padova

Single-cell RNA-Seq (scRNA-seq) is the most widely used high-throughput technology to measure genome-wide gene expression at the single-cell level. One of the most common analyses of scRNA-seq data detects distinct subpopulations of cells through the use of unsupervised clustering algorithms. However, recent advances in scRNA-seq technologies results in current datasets ranging from thousands to millions of cells. Popular clustering algorithms, such as k-means, typically require the data to be loaded entirely into memory and therefore can be slow or impossible to run with datasets of this size. To address this, we developed an open-source implementation of the mini-batch k-means algorithm, mbkmeans, which leverages in memory and on-disk data representations, such as HDF5 files. We illustrate the accuracy and advantages of our approach with significant savings in terms of speed and memory-usage. Finally, we demonstrate the performance of mbkmeans using large Human Cell Atlas datasets including a dataset with 1.3 million observations.

e-mail: shicks19@jhu.edu

92. ROBUST MODELING AND INFERENCE

Robust Estimation with Outcome Misclassification and Covariate Measurement Error in Logistic Regression

Sarah C. Lotspeich*, Vanderbilt University Medical Center
Bryan E. Shepherd, Vanderbilt University Medical Center
Gustavo G.C. Amorim, Vanderbilt University Medical Center
Pamela A. Shaw, University of Pennsylvania
Ran Tao, Vanderbilt University Medical Center

While electronic health records were first intended to support clinical care and billing, these databases are now used for clinical investigations aimed at preventing disease, improving patient care, and informing policymaking. However, outcomes and covariates can be captured with error and their errors correlated. While complete data validation is cost prohibitive, a two-phase design may be feasible for addressing errors. During Phase I error-prone variables are observed for all subjects, and this information is then used to select a Phase II validation subsample. Previous approaches to outcome misclassification using two-phase design data are limited to error-prone categorical predictors and make distributional assumptions about the errors. We propose a semiparametric approach to two-phase designs with a misclassified binary outcome and categorical or continuous error-prone predictors, allowing for dependent errors and arbitrary second-phase selection. The method is robust because it leaves the predictors' error mechanisms unspecified. Performance is compared to existing approaches through extensive simulation and use illustrated in an observational HIV study.

email: sarah.c.lotspeich@vanderbilt.edu

Implementing Interventions to Combat the Opioid Epidemic During a Rising Tide of Activities Aimed at Improving Patient Outcomes: Evaluating the Robustness of Parallel-Group and Stepped-Wedge Cluster Randomized Trials to Confounding from External Events

Lior Rennert*, Clemson University

In response to the opioid epidemic, the National Institute on Drug Abuse recently awarded four states roughly $100 million each to implement an integrated set of interventions in high-risk communities using a parallel-group cluster randomized trial (CRT) or a stepped-wedge design (SWD). The objective is to reduce opioid overdose deaths by 40% over a three-year period. However, over the past several years there has been a rising tide of activities aimed at improving opioid-related outcomes. State laws limiting initial opioid prescriptions were enacted in all four states between 2016 and 2017. In late 2017, the CDC launched an intensive media awareness campaign in three of these four states. External events such as policies and awareness campaigns, along with other interventions, are expected to impact opioid-related outcomes during the same time period. This has major implications for the proposed interventions, as these external events may confound the intervention effect estimate. Here we evaluate the robustness of the CRT and SWD to confounding from external events, and discuss modeling strategies to reduce the bias in the intervention effect estimate in these scenarios.

email: lior.rennert@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Robust Statistical Models for Impact Injury Risk Estimation

Anjishnu Banerjee*, Medical College of Wisconsin
Narayan Yoganandan, Medical College of Wisconsin

Injury probability curves form the primary basis for United States Federal Motor Vehicle Safety Standards evaluating crashworthiness and safety of vehicles sold in the USA. However, in spite of their importance, such injury risk curves remain largely understudied and current methods for their estimation lead to risk curves of poor quality. We present an approach that combines advanced statistical experimental design, mathematical modeling, and hierarchical Bayesian time-to-event data methods to create predictive risk curves. We also propose fundamental changes in the way risk curves are constructed from these experiments, moving beyond simple risk curve estimation to more informative and predictive risk surface estimation. Our proposed risk surface methodology would provide a holistic assessment of injury probability by accommodating multiple injury measures and multiple biomechanical metrics, borrowing information across all of them, in the same estimation model. A by-product of our method is a new way to evaluate quality of risk curves.

email: abanerjee@mcw.edu

Joint Testing of Donor/Recipient Genetic Matching Scores and Recipient Genotype has Robust Power for Finding Genes Associated with Transplant Outcomes

Victoria L. Arthur*, University of Pennsylvania Perelman School of Medicine
Sharon Browning, University of Washington
Bao-Li Chang, University of Pennsylvania
Brendan Keating, University of Pennsylvania
Jinbo Chen, University of Pennsylvania Perelman School of Medicine

Genetic matching between transplant donor and recipient pairs has traditionally focused on the HLA regions of the genome, but recent studies suggest that matching for non-HLA regions may be important as well. We propose three new genetic matching scores for use in association analyses of transplant outcomes. These scores describe genetic ancestry distance, using IBD or IBS, or genetic incompatibility of the two genomes and therefore may reflect underlying biological mechanisms for donor and recipient genes to influence transplant outcomes. Our simulation studies show that jointly testing these scores with the recipient genotype is a powerful method for preliminary screening and discovery of transplant outcome related SNPs and gene regions. Following these joint tests with marginal testing of the recipient genotype and matching score separately can lead to further understanding of the biological mechanisms behind transplant outcomes. In addition, we present results of a liver transplant data analysis that show joint testing can detect SNPs significantly associated with acute rejection in liver transplant.

email: vlynn@pennmedicine.upenn.edu

A Robust Bayesian Copas Selection Model for Detecting and Correcting Publication Bias

Ray Bai*, University of Pennsylvania
Yong Chen, University of Pennsylvania
Mary Regina Boland, University of Pennsylvania

The validity of conclusions from meta-analysis is potentially threatened by publication bias. Most existing procedures for handling this issue decouple testingfor the presence of publication bias from estimationof unknown parameters under publication bias. Most of these procedures also assume normality of the between-study random effects. This assumption may be invalid, especially if there are outliers in the studies included in the meta-analysis. Finally, there exist few procedures to quantify the magnitude of publication bias. In this paper, we simultaneously address all of these issues. First, we introduce the robust Bayesian Copas (RBC) selection model, which unifies inference and estimation and which offers robustness to strong assumptions about the distribution of the random effects. Second, we develop two new measures to quantify the magnitude of publication bias: one based on point estimates which can also be used for non-Bayesian methods, and one based on the posterior distribution. We illustrate our method through simulations and two case studies.

email: ray.bai@pennmedicine.upenn.edu

Estimation of Knots in Linear Spline Models

Guangyu Yang*, University of Michigan
Baqun Zhang, Shanghai University of Finance and Economics
Min Zhang, University of Michigan

Linear spline models are able to accommodate nonlinear effects while maintaining easy interpretation. It has significant applications in studying threshold effects and change-points. However, the lack of rigorously studied and computationally convenient method for estimating knots has limited its use in practice. A key difficulty in estimation of knots lies in the nondifferentiability. In contrast to previous methods which tackle the nondifferentiability by smoothing, we propose a novel and simple method to circumvent the difficulty by redefining derivatives at places that are not differentiable. As no smoothing parameter is involved, the method is computationally convenient. A two-step algorithm is used to take advantage of the analytic solution when knots are known, further improving stability. We show that the proposed estimator is consistent and asymptotically normal using the empirical process theory. Comprehensive simulation studies have shown our method performs well in terms of both statistical and computational properties and offers substantial improvement over existing methods.

email: yguangyu@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

93. HIGH DIMENSIONAL METHODS FOR MECHANISTIC INTEGRATION OF MULTI-TYPE OMICS

Integrating Heterogeneous Longitudinal Omics Data with Personalized Dynamic Network Analysis

Xing Qiu*, University of Rochester
Leqin Wu, Jinan University
Ya-xiang Yuan, Chinese Academy of Sciences
Hulin Wu, University of Texas Health Science Center at Houston

Dynamic network (DN) analysis has been successfully used in reconstructing gene regulatory networks, protein-protein interaction networks, and microbial interaction networks, etc. Often we can reconstruct subject-level DNs from longitudinal Omics data, and the results are a sample of networks that best reflect the heterogeneity among subjects. Network features can be extracted at the subject-level and used in integrative analysis based on statistical and machine learning methods. DN models are inherently complex (with ~p2 unknown parameters) and require advanced mathematical models such as ordinary differential equations (ODEs) or state space models (SSMs) to reconstruct. We developed an ODE-based DN reconstruction method based on similarity transformation and separable least squares, which is not only more accurate but also several magnitude of orders faster than the competing methods such as the direct least squares method and the two-stage method. As a result, we were able to reconstruct a large-scale DN with 1250 dimensions and 1,563,750 unknown parameters from the real data.

email: xing_qiu@urmc.rochester.edu

INFIMA Leverages Multi-Omic Model Organism Data to Identify Target Genes for Human GWAS Variants

Sunduz Keles*, University of Wisconsin, Madison
Chenyang Dong, University of Wisconsin, Madison

Genome-wide association studies (GWAS) revealed many variants that are statistically associated with disease risk, disease protection, or other traits. However, target susceptibility genes at most GWAS risk loci remain unknown. While transcriptome-wide association studies leverage reference transcriptomes to elucidate candidate genes, model organism studies largely remain as an untapped potential for unveiling susceptibility genes and functional investigation of findings from human GWAS. We developed a framework named, INFIMA for Integrative Fine Mapping with Model Organism Multi-Omics Data. INFIMA leverages multi-omics data from Diversity Outbred (DO) mice, derived from eight inbred mouse strains, and identifies candidate genes for human GWAS susceptibility loci. Our application to GWAS loci of 14 diabetic traits identified several novel susceptibility genes for these diabetic traits. We validated INFIMA fine mapping results with both mouse and human high throughput chromatin capture data from islet cells.

email: keles@stat.wisc.edu

Nonlinear Moderated Mediation Analysis with Genetical Genomics Data

Yuehua Cui*, Michigan State University
Bin Gao, Michigan State University
Xu Liu, Shanghai University of Finance and Economics

The central dogma tells us the causal chain from DNAs to RNAs and to proteins which further manifest into phenotypes. In this causal model, gene expressions can be viewed as mediators which mediate the relationship between genotypes and phenotypes. Here we propose a moderated mediation model considering the causal mediation effect of gene expressions on the relationship between genotypes and phenotypes, which could be nonlinearly moderated by environmental factors. The goal is to select important genetic and gene expression variables that can predict a phenotypic response under a high-dimensional setup. Given the fact that genes function in networks to fulfill their joint task, we incorporate gene network structures to further improve gene selection performance via a graph-constrained penalty. We establish the selection consistency property, and further perform simulation and real data analysis to show the utility of the method.

email: cuiy@msu.edu

High Dimensional Mediation Analysis for Causal Gene Selection

Qi Zhang*, University of Nebraska, Lincoln

Mediation analysis has been a popular framework for elucidating the mediating mechanism of the exposure effect on the outcome in many disciplines including genetic studies. Previous literature in causal mediation primarily focused on the classical settings with univariate exposure and univariate mediator, with recent growing interests in high dimensional mediator. I study the mediation model with high dimensional exposure and high dimensional mediator, and introduce two procedures for mediator selection, MedFix and MedMix. This study is motivated by the causal gene identification problem, where causal genes are defined as the genes that mediate the genetic effect. For this problem, the genetic variants are the high dimensional exposure, the gene expressions the high dimensional mediator, and the phenotype of interest the outcome. The proposed methods are evaluated on a mouse f2 dataset for diabetes study, and by extensive real data driven simulations. The results show that the mixed model based approach leads to higher accuracy in mediator selection, and is more reproducible across independent measurements of the response and more robust against model misspecification.

email: qi.zhang@unl.edu

ABSTRACTS & POSTER PRESENTATIONS

94. NEW WEIGHTING METHODS FOR CAUSAL INFERENCE

Propensity Score Weighting for Causal Inference with Multiple Treatments

Fan Li*, Yale School of Public Health
Fan Li, Duke University

Unconfounded comparisons of multiple groups are common in observational studies. We propose a unified framework, the balancing weights, for estimating causal effects with multiple treatments using propensity score weighting. The class of balancing weights include existing approaches such as inverse probability weights as special cases. Within this framework, we propose a class of target estimands and their corresponding nonparametric weighting estimators. We further develop the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weights correspond to the target population with the most overlap in covariates between treatments, similar to the population in equipoise in clinical trials. We show that the generalized overlap weights minimize the total asymptotic variance of the weighting estimators for the pairwise contrasts within the class of balancing weights. We illustrate these methods using a real comparative effectiveness dataset and further examine their properties by simulations.

email: fan.f.li@yale.edu

Methods for Balancing Covariates when Estimating Heterogeneous Treatment Effects in Observational Data

Laine Thomas*, Duke University
Fan Li, Duke University
Daniel Wojdyla, Duke Clinical Research Institute
Siyun Yang, Duke University

The COMPARE-UF registry evaluated myomectomy versus hysterectomy procedures for women with uterine fibroids using propensity score (PS) methods to adjust for pre-procedure differences, with pre-specified subgroup analysis (SGA). SGA is vulnerable to bias because standard PS methods do not achieve good covariate balance in subgroups. After inverse probability of treatment weighting (IPTW), younger women (age<40) receiving hysterectomy still have worse symptoms. We propose a novel PS method for SGA that takes advantage of the scientific knowledge encoded in pre-specified subgroups. That is, these subgroups are candidates for interaction in the PS model because physicians and patients think about them differently. This knowledge motivates an augmented candidate list of adjustment variables that includes interactions between all confounders and the subgroups of interest. We pair this with machine learning methods for variable selection, and the recently developed overlap weighting (OW). OW is more efficient than IPTW and may therefore mitigate the variance tradeoff that arises with a more complex PS model. We evaluate this approach by simulation and in COMPARE-UF.

email: laine.thomas@duke.edu

Flexible Regression Approach to Propensity Score Analysis and its Relationship with Matching and Weighting

Liang Li*, University of Texas MD Anderson Cancer Center
Huzhang Mao, Eli Lilly and Company

In propensity score analysis, the frequently used regression adjustment involves regressing the outcome on the estimated propensity score and treatment indicator. This approach can be highly efficient when model assumptions are valid, but can lead to biased results when the assumptions are violated. We extend the simple regression adjustment to a varying coefficient regression model that allows for nonlinear association between outcome and propensity score. We discuss its connection with some propensity score matching and weighting methods, and show that the proposed analytical framework unifies the four mainstream propensity score approaches (stratification, regression, matching and weighting) and handles commonly used causal estimands. We derive analytic point and variance estimators that properly take into account the sampling variability in the estimated propensity score. We illustrate this approach with simulations and a data application.

email: LLi15@mdanderson.org

Robust Inference when Combining Probability and Non-Probability Samples with High-Dimensional Data

Shu Yang*, North Carolina State University
Jae Kwang Kim, Iowa State University
Rui Song, North Carolina State University

Non-probability samples become increasingly popular in survey statistics but may suffer from selection bias that limits the generalizability of results to the target population. We consider integrating a non-probability sample with a carefully designed probability sample which provides the representative covariate information of the target population. We propose a two-step approach utilizing double estimating equations for variable selection and estimation. In the first step, we formulate the estimating equation for the sampling score by balancing the covariate information of the non-probability sample and the probability sample. We show that the penalized estimation equation approach enjoys the selection consistency property for general probability samples. In the second step, we focus on a doubly robust estimator of the finite population mean and re-estimate the model parameters by minimizing the asymptotic square bias of the doubly robust estimator. This estimating strategy mitigates the possible first-step selection error and renders the doubly robust estimator root-n consistent if either the sampling probability or the outcome model is correctly specified.

email: syang24@ncsu.edu

ABSTRACTS & POSTER PRESENTATIONS

95. USING MACHINE LEARNING TO ANALYZE RANDOMIZED TRIALS: VALID ESTIMATES AND CONFIDENCE INTERVALS WITHOUT MODEL ASSUMPTIONS

Performance Evaluation of Flexible Strategies for Estimating HIV Vaccine Efficacy

Alex Luedtke*, University of Washington

The HIV Vaccine Trials Network currently has five ongoing efficacy trials evaluating regimens of vaccine and monoclonal antibody products. The statistical analysis plans for these trials prespecify analyses that leverage supervised machine learning to improve the precision of efficacy estimates. Examples of planned analyses include covariate-adjusted estimators of the intent-to-treat efficacy, causally-motivated estimators of the per-protocol efficacy, and estimators of principally stratified treatment effects. In these analyses, covariate adjustment is expected to improve precision and/or reduce bias for the biological parameter that is of scientific interest. Though all of the estimation strategies that we plan to use have been proven to yield asymptotically valid confidence intervals under minimal conditions, there are currently no finite-sample performance guarantees available. In an effort to better understand the finite-sample behavior of these strategies, we have developed an adversarial Monte Carlo strategy that identifies data-generating mechanisms at which our estimators perform most poorly at a given sample size. I will conclude by describing this strategy.

email: aluedtke@uw.edu

Inference for Model-Light Machine Learning in Precision Medicine

Michael Kosorok*, University of North Carolina, Chapel Hill

Inference for machine-learning based dynamic treatment regime estimation can be challenging, especially for highly flexible models with large numbers of features. We explore some trade-offs in this context between precision and model flexibility.

email: kosorok@unc.edu

Synthetic Difference in Differences

David A. Hirshberg*, Stanford University
Dmitry Arkhangelsky, CEMFI, Madrid
Susan Athey, Stanford University
Guido Imbens, Stanford University
Stefan Wager, Stanford University

We propose a new estimator for the average treatment effect on the treated in panel data with simultaneous adoption of treatment. The estimator is a weighted version of the well-known difference in differences estimator. Like the synthetic control estimator, our estimator uses unit weights to improve the validity of the comparison between treated and control units. And like time-series forecasting methods based on linear regression, our estimator uses a weighted average of pre-treatment time periods that is predictive of the post-treatment period to improve the validity of pre/post comparisons. We find that this new Synthetic Difference in Differences estimator has attractive properties compared to synthetic control, linear forecasting, and difference-in-differences estimators. We show that our estimator is consistent and asymptotically normal under relatively weak assumptions and give a consistent estimator for its standard error.

email: davidahirshberg@stanford.edu

Machine Learning Versus Standard Methods for Covariate Adjustment: Performance Comparison Across 10 Completed Randomized Trials

Michael M. Rosenblum*, Johns Hopkins Bloomberg School of Public Health

Adjusting for prognostic baseline variables in estimating the average treatment effect in randomized trials (called covariate adjustment) has potential to substantially increase precision and power (and reduce the required sample size) in multiple disease areas. We compare estimators based on machine learning algorithms (such as random forest) versus simpler estimators (such as the unadjusted estimator and analysis of covariance) for estimating the average treatment effect in randomized trials. Our comparison of estimators is conducted using 10 completed randomized trial data sets. Some of the trials involve similar treatments and populations, which gives a unique opportunity to explore which methods (if any) consistently provide substantial precision gains.

email: mrosenbl@jhsph.edu

ABSTRACTS & POSTER PRESENTATIONS

96. RECENT DEVELOPMENTS IN SEMIPARAMETRIC TRANSFORMATION MODELS

Semiparametric Regression Models for Indirectly Observed Outcomes

Jan De Neve*, Ghent University
Heidelinde Dehaene, Ghent University

We propose a flexible framework to analyze outcomes that are indirectly observed via one or multiple proxies. Semiparametric transformation models, including Cox proportional hazards regression, turn out to be well suited to model the association between the covariates and the unobserved outcome. By coupling this regression model to a semiparametric measurement model, we can estimate these associations without requiring calibration data and without imposing strong assumptions on the relationship between the unobserved outcome and its proxy. When multiple proxies are available, we propose a data-driven aggregation resulting in an improved proxy with a superior quality than each of the proxies separately. We empirically validate the proposed methodology in a simulation study, revealing good finite sample properties, especially when multiple proxies are aggregated. The methods are demonstrated on a case study.

email: JanR.DeNeve@ugent.be

Addressing Outcome Detection Limits using Semiparametric Cumulative Probability Models

Bryan E. Shepherd*, Vanderbilt University
Yuqi Tian, Vanderbilt University

Left censoring due to assay detection limits is frequently encountered in biomedical research. Many approaches have been employed to address these types of data, but most make parametric assumptions that implicitly assume the distribution of the data below the detection limit can be predicted by models fit to data above the detection limit. We propose analyzing these data with semiparametric cumulative probability models (CPMs). CPMs belong to the class of semiparametric linear transformation models, and they are invariant to monotonic transformations. Fitting a CPM is equivalent to fitting an ordinal cumulative link model. The responses follow a mixture distribution with those observations below the detection limit being discrete ordinal and those above the detection limit being continuous. CPMs implicitly assign observations below the detection limit as having the lowest rank value. CPMs can also be used with minor modifications to address multiple detection limits, which may arise from different assays being used over time or across study sites. We illustrate the use of CPMs through simulations and real data examples from studies of HIV.

email: bryan.shepherd@vanderbilt.edu

Cumulative Probability Models for Big Data

Chun Li*, Case Western Reserve University

Transformation models are robust alternatives to traditional regression models. They can be used to nonparametrically estimate the outcome transformation necessary for regression analyses. Cumulative probability models (CPMs) (Liu et al. 2017, Stat in Med) were recently proposed as a way to fit semiparametric linear transformation models. CPMs are easy to fit using standard software for ordinal regression. However, the demand for computational resources increases quickly as the sample size increases (e.g., 100 GB memory may be needed when the sample size is 40,000). For big data, we implement a divide-and-conquer algorithm to fit CPMs: the data are partitioned into subsets, each is used to fit a CPM, and the results are aggregated to obtain the final estimates of the parameters and their variance-covariance matrix. We will demonstrate that the algorithm takes much less time and memory to run and it performs quite well. The algorithm makes it feasible to fit semiparametric linear transformation models for sample sizes in the millions.

email: lichun1668@gmail.com

97. INNOVATIONS IN STATISTICAL NEUROSCIENCE

A Study of Longitudinal Trends in Time-Frequency Transformations of EEG Data During a Learning Experiment

Damla Senturk*, University of California, Los Angeles
Joanna Boland, University of California, Los Angeles
Shafali Jeste, University of California, Los Angeles
Donatello Telesca, University of California, Los Angeles

EEG is a non-invasive and widely-available (low cost) brain imaging modality which records electrical activity in the brain. An event-related potential (ERP) is defined as the EEG wave- form measured in response to each stimulus in EEG experiments. We consider a time-frequency decomposition of the ERP data, targeting even richer information than is available in the single time or frequency domain analysis. We propose longitudinal time-frequency transformation of ERP (LTFT-ERP), where after signal-to-noise ratio is enhanced through MAP-ERP, a wavelet transformation is applied to the resulting ERPs from trials in a sliding window. LTFT-ERP, not only targets richer information in the signal through time-frequency transformations, it also allows modeling of longitudinal changes in the signal over trials throughout the learning experiment, adding an additional dimension for analysis (referred to as the longitudinal dimension). The proposed methodology is demonstrated through applications to a study on implicit learning of children with Autism Spectrum Disorder (ASD).

email: dsenturk@ucla.edu

ABSTRACTS & POSTER PRESENTATIONS

Improved Diagnostics and Prognostics using MRI in Multiple Sclerosis

Russell Shinohara*, University of Pennsylvania

Although multiple sclerosis (MS) is sometimes an easy disease to diagnose, there are many cases in which clinicians struggle with using the current diagnostic criteria. Furthermore, the hallmark white matter lesion in MS may appear homogenous and is generally treated as such, but histology tells us that there are different types of lesions in which competing biological processes are active. This is often due to the lack of specificity of classical MRI and biomarkers developed based on these modalities. In this presentation, we will review the state-of-the-art in measuring MS disease activity and progression using MRI, and discuss the challenges of using new, more sensitive and specific imaging modalities. We will further describe several statistical examinations of these imaging techniques, and new statistical approaches for analyzing these advanced MRI to develop diagnostic and prognostic biomarkers.

email: taki.shinohara@gmail.com

Intensity Warping for Multisite MRI Harmonization

Julia L. Wrobel*, Colorado School of Public Health
Melissa Martin, University of Pennsylvania
Taki Shinohara, University of Pennsylvania
Jeff Goldsmith, Columbia University

In multisite neuroimaging studies there is often unwanted technical variation across scanners and sites. These scanner effects can hinder detection of biological features of interest, produce inconsistent results, and lead to spurious associations. We assess scanner effects in two brain magnetic resonance imaging (MRI) studies where subjects were measured on multiple scanners within a short time frame, so that one could assume differences between images were due to technical rather than biological effects. We propose a tool to harmonize images by identifying and removing within-subject scanner effects. Our goals were to (1) establish a method that removes scanner effects by leveraging multiple scans collected on the same subject, and (2) develop a technique to quantify scanner effects in multisite trials so these can be reduced as a preprocessing step. We found that unharmonized images were highly variable across site and scanner type, and our method reduced variability by warping intensity distributions. We further studied the ability to predict intensity harmonization results for a scan taken on an existing subject at a new site using cross-validation.

email: julia.wrobel@cuanschutz.edu

Bayesian Approaches for Estimating Dynamic Functional Network Connectivity in fMRI Data

Michele Guindani*, University of California, Irvine

Dynamic functional connectivity, that is, the study of how interactions among brain regions change dynamically over the course of an fMRI experiment, has recently received wide interest in the neuroimaging literature. Current approaches for studying dynamic connectivity often rely on ad hoc approaches for inference, with the fMRI time courses segmented by a sequence of sliding windows. We discuss Bayesian approaches to dynamic functional connectivity, based on the estimation of time varying networks. Our methods utilize dynamic space-state models for classification of latent cognitive states, achieving estimation of the networks in an integrated framework that borrows strength over the entire time course of the experiment. Network structures and transitions can be inferred by prior information or available covariates from the experiment. We discuss the performance of our methods on simulated and real fMRI data.

email: mguindan@uci.edu

98. ARTIFICIAL INTELLIGENCE FOR PREDICTION OF HEALTH OUTCOMES

Distributed Learning from Multiple EHR Databases for Predicting Medical Events

Qi Long*, University of Pennsylvania
Ziyi Li, Emory University
Kirk Roberts, University of Texas Health Science Center at Houston
Xiaoqian Jiang, University of Texas Health Science Center at Houston

Electronic health records (EHRs) data offer great promises in personalized medicine. However, they also present analytical challenges due to their irregularity and complexity. For example, EHRs data include both structure and unstructured data. In addition, sharing EHRs data across multiple institutions/sites may be infeasible due to privacy concerns and regulatory hurdles. We propose a distributed learning method to build prediction model for undiagnosed conditions using data from multiple EHRs systems without sharing subject-level data across them. The proposed approach can use both structured and unstructured EHRs data collected over time. Our numerical studies demonstrate that the proposed method can build predictive models in a distributed fashion with privacy protection and the resulting models achieve comparable prediction accuracy compared with existing methods that need to use aggregated data pooled across all sites. Our method has the potential to enable diagnosis and treatment at an earlier stage of the natural history of a disease, which is known to be associated with better patient outcomes.

email: qlong@upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Deep Learning with Time-to-Event Outcomes

Jon Steingrimsson*, Brown University
Samantha Morrison, Brown University
Constantine Gatsonis, Brown University

Deep learning is a class of algorithms that uses multiple layers to create a risk prediction model. The layers involve an unknown weight vector that is estimated by minimizing a loss function. We extend the deep learning algorithms to handle censoring by replacing the loss function used in the absence of censoring by censoring unbiased loss functions. We discuss properties of these loss functions and practical issues related to implementation of the deep learning algorithms. The performance of the resulting algorithms is evaluated through simulation studies and by analyzing data on cancer patients.

email: jon.arni.steingrimsson@gmail.com

A Scalable Discrete-Time Survival Model for Neural Networks

Balasubramanian Narasimhan*, Stanford University

Neural networks are widely used for machine learning in medicine. Several neural-network methods have been proposed for fitting survival models where one has time to event data. We describe a discrete-time survival model (https://doi.org/10.7717/peerj.6257) called Nnet-survival that is designed to be used with neural networks. The model is flexible, so that the baseline hazard rate and the effect of the input data on hazard probability can vary with follow-up time. It can deal with scale with large datasets as the model is trained using minibatch stochastic gradient descent. Our implementation uses the Keras deep learning framework. We demonstrate the performance of the model on both simulated and real data and compare it to existing models such as Cox-nnet and Deepsurv.

email: naras@stanford.edu

Deep Learning for Dynamic Prediction of Cardiovascular Events

Lihui Zhao*, Northwestern University

Cardiovascular disease (CVD) is the leading cause of morbidity and mortality. CVD risk prediction plays a central role in clinical CVD prevention strategies, by aiding decision making for lifestyle modification and to match the intensity of therapy to the absolute risk of a given patient. Various CVD risk factors have been identified and used to construct multivariate risk prediction algorithms. However, these algorithms are generally based on the risk factors measured at a single time. Since risk factors like blood pressure (BP) are regularly collected in clinical practice, and electronic medical records are making longitudinal data on these risk factors available to clinicians, dynamic prediction of CVD risk on a real-time basis using the history of CV risk factors will likely improve the precision of personalized CVD risk prediction. We will present deep learning methods to build dynamic CVD risk prediction models using repeated measured risk factor levels. The pooled data from multiple community-based CVD cohorts will be used.

email: lihui.zhao@northwestern.edu

99. LATENT VARIABLES AND PROCESSES

Modeling the Effects of Multiple Exposures with Unknown Group Memberships: A Bayesian Latent Variable Approach

Alexis E. Zavez*, University of Rochester Medical Center
Emeir M. McSorley, Ulster University
Sally W. Thurston, University of Rochester Medical Center

We propose a Bayesian latent variable model to allow estimation of the covariate-adjusted relationships between an outcome and a small number of latent exposure variables, using data from multiple observed exposures. Each latent variable is assumed to be represented by multiple exposures, where membership of the observed exposures to latent groups is unknown. Our model assumes that one measured exposure variable can be considered as a sentinel marker for each latent variable, while membership of the other measured exposures is estimated using MCMC sampling based on a classical measurement error model framework. We illustrate our model using data on multiple cytokines and birth weight from the Seychelles Child Development Study, and evaluate the performance of our model in a simulation study. Classification of cytokines into Th1 and Th2 cytokine classes in the Seychelles study revealed some differences from standard Th1/Th2 classifications. In simulations, our model correctly classified measured exposures into latent groups, and posterior estimates, absolute bias, and coverage were similar to those from the oracle model.

email: alexis_zavez@urmc.rochester.edu

A Time-Dependent Structural Model Between Latent Classes and Competing Risks Outcomes

Teng Fei*, Emory University
John Hanfelt, Emory University
Limin Peng, Emory University

Latent class analysis is an intuitive tool to characterize disease phenotype heterogeneity. With data more frequently collected on multiple phenotypes in chronic disease studies, it is of rising interest to investigate how the latent classes embedded in one phenotype are related to another phenotype. Motivated by a cohort with mild cognitive impairment (MCI) from the Uniform Data Set (UDS), we propose a time-dependent structural model between latent classes and competing risk outcomes. We develop a two-step estimation procedure which circumvents latent class assignment and is rigorously justified for accounting for the uncertainty in classifying latent classes. The new method also properly addresses the random censoring to the competing risks and the missing failure types of competing risks. The asymptotic properties of the resulting estimator are established. We develop sample-based inference procedures whereas standard bootstrapping inference is infeasible. Simulation studies demonstrate the advantages of the new method over benchmark tools. An application to the MCI data from UDS uncovers detailed pictures of the neuropathological relevance of the baseline MCI subgroups.

email: tfei@emory.edu

ABSTRACTS & POSTER PRESENTATIONS

Dirichlet Depths for Point Process

Kai Qi*, Florida State University
Yang Chen, Florida State University
Wei Wu, Florida State University

Statistical depths have been well studied for multivariate and functional data over the past few decades, but remain under-explored for point processes. A first attempt on the notion of point process depth was conducted recently where the depth was defined as a weighted product of two terms: (1) the probability of the number of events in each process (2) the depth of the event times conditioned on the number of events by using a Mahalanobis depth. We point out that multivariate depths such as the Mahalanobis depth may not be directly used because they often neglect the important ordered property in the point process events. To deal with this problem, we propose a model-based approach for point processes systematically. We examine the mathematical properties of the new depths and conduct the asymptotic analysis. In addition, we illustrate the new methods using various simulated and real experiment data. It is found that the proposed framework provides a proper center-outward rank and the new methods have superior decoding performance to previous methods in a neural spike train dataset.

email: kq15b@my.fsu.edu

Acknowledging the Dilution Effect in Group Testing Regression: A New Approach

Stefani C. Mokalled*, Clemson University
Christopher S. McMahan, Clemson University
Derek A. Brown, Clemson University
Joshua M. Tebbs, University of South Carolina
Christopher R. Bilder, University of Nebraska, Lincoln

From screening for infectious diseases to drug discovery, group testing has proven to be a cost efficient alternative to individual level testing. Cost savings are realized through testing pooled biospecimens formed by amalgamating individual samples. A common concern that arises due to pooling is the “dilution” effect; i.e., the signal from a positive individual’s specimen might be diluted past an assay’s threshold of detection by pooling it with multiple negative samples. To account for this effect, we propose a new statistical framework for group testing data that merges the areas of classification and estimation. The proposed approach analyzes continuous biomarker levels observed from assaying pooled samples to estimate a regression function describing the covariate dependent probability of infection for each individual and the distributions of the biomarker levels of the cases and controls. We illustrate how the estimates of the individual level biomarker distributions can be used to identify pool-specific diagnostic thresholds. The methodologies are evaluated through numerical studies and are illustrated using Hepatitis B virus data on Irish prisoners.

email: smokall@g.clemson.edu

Modeling Brain Waves as a Mixture of Latent Processes

Guillermo Cuauhtemoctzin Granados Garcia*, King Abdullah University of Science and Technology
Hernando Ombao, King Abdullah University of Science and Technology
Mark Fiecas, University of Minnesota
Babak Shahbaba, University of California, Irvine

Brain electrical activity is measured by electroencephalograms (EEG) or event-related potentials (LFP). We propose a new methodology to characterize a single brain wave by decomposing it as a discrete mixture of parametric processes. We based the inference on the standardized Spectral Density Function (SDF) as a Dirichlet Process mixture of kernels each derived from a latent second-order auto-regressive process defined by a location parameter as a frequency peak and scale parameter as the spread of the peak. We present an algorithm for computing the posterior distribution of the parameters of each mixture component, the number of components and the mixing weights using a Metropolis-Hastings within Gibbs algorithm. The advantage of the proposed model is the peaks of the brain activity and the complexity of the mixture are determined from the data and the estimator has a simple interpretation in the time and frequency domain. A simulation study is carried out to estimate the SDF of two ARMA models. Finally, the model is used to study the hippocampal neural mechanism underlying the memory for a sequence of events using rat LFP data.

email: guillermo.granadosgarcia@kaust.edu.sa

A Method to Flexibly Incorporate Covariates in Latent Class Analysis with Application to Mild Cognitive Impairment

Grace Kim*, Emory University Rollins School of Public Health
John Hanfelt, Emory University Rollins School of Public Health

Mild Cognitive Impairment (MCI) is a neurocognitive disorder with heterogeneous subclinical entities, necessitating the assessment of numerous features for accurate classification. Latent class regression, an important extension of the latent class analysis framework, can be used to incorporate covariates as risk factors for MCI subtype class membership. We explore an application of latent class regression where covariates unintentionally alter clinical interpretation of latent classes of MCI. We introduce the concept of a covariate activity governor, which provides a flexible method of covariate incorporation without extensively distorting the clinical interpretation of latent classes. We consider a scenario with covariates strongly related to comorbidity and less so to the disease of interest, and show that without the activity governor, latent class regression yields classes more closely related to comorbidity than to the disease. By applying the activity governor to a sample of 1655 participants in the Uniform Data Set of the National Alzheimer’s Coordinating Center, we identify a clinically interpretable model of the structure of MCI in the presence of vascular covariates.

email: gskim@emory.edu

ABSTRACTS & POSTER PRESENTATIONS

Exploration of Misspecification in Latent Class Trajectory Analysis (LCTA) and Growth Mixture Modeling (GMM): Error Structure Matters

Megan L. Neely*, Duke University
Jane Pendergast, Duke University
Bida Gu, Duke University
Natasha Dmitreava, Duke University Medical Center
Carl Pieper, Duke University

Modeling longitudinal trajectories and identifying latent classes of those trajectories is common in biomedical research. Software to identify latent classes is readily available, leading to increased use of LTCA and GMM. LCTA assumes observations are independent conditional on class membership. However, within-person correlation is often non-negligible and can be accounted for in GMM using random effects. Despite differences in assumptions, we found LCTA (via PROC TRAJ in SAS) is widely used. In this work, we investigated if LCTA is robust to the presence of within-person correlations, how correlation misspecification impacts findings from LCTA and GMM, and availability of software for implementing LCTA and GMM. Using simulation, we varied correlation structures and strength. We found, even in the presence of weak correlation, LCTA performed poorly in class enumeration resulting in biased estimation of class trajectories, incorporation of the correct correlation structure was crucial for GMM in class enumeration, and as expected, under correlation misspecification both LCTA and GMM gave unbiased estimates of class trajectories when the number of classes was correctly specified.

email: megan.neely@duke.edu

100. TIME-TO-EVENT DATA ANALYSIS: SURVIVAL OF THE FITTEST

Survival Analysis under the Cox Proportional Hazards Model with Pooled Covariates

Paramita Saha Chaudhuri*, McGill University
Lamin Juwara, McGill University

For a time-to-event outcome and an expensive-to-measure exposure, we develop a pooling design and propose a likelihood-based approach to estimate the hazard ratios (HRs) of a Cox Proportional Hazards (PH) model. To consistently estimate HRs with a continuous time-to-event outcome (individually observed) and a continuous exposure that is subject to pooling, we first focus on a risk set-matched nested case-control (NCC) subcohort of the original cohort. Recasting the original problem within an NCC subcohort enables us to form risk set-matched strata, randomly form pooling groups with the matched strata and pool the exposures correspondingly without explicitly considering the event times in the estimation. Then, the pooled exposure levels can be used within a standard maximum likelihood framework with a matched case-control design to consistently estimate the HRs and obtain the model-based standard error. Our simulation results and analysis of the SMART study indicate that the HRs estimated using this pooled design are comparable to the estimates obtained from the NCC subcohort.

email: paramita.sahachaudhuri.work@gmail.com

Quantile Association Regression on Bivariate Survival Data

Ling-Wan Chen*, National Institute of Environmental Health Sciences, National Institutes of Health
Yu Cheng, University of Pittsburgh
Ying Ding, University of Pittsburgh
Ruosha Li, University of Texas Health Science Center at Houston

The association between two event times is of scientific importance in various fields. The local association measures capture the dynamic pattern of association over time, and it is desirable to examine the degree to which local association depends on different characteristics of the population. In this work, we adopt a novel quantile-based local association measure, and propose a conditional quantile association regression model to allow covariate effects in the local association analysis for bivariate survival data. An estimating equation for the quantile association coefficients is constructed on the basis of the relationship between this quantile association measure and the conditional copula. The asymptotic properties for the resulting estimators are rigorously derived. To avoid estimating density functions, we extend the induced smoothing idea to our proposed estimators in obtaining the covariance matrix. The proposed estimators and inference procedure are evaluated through simulations, and applied to an age-related macular degeneration (AMD) dataset, where we explore the association between AMD progression times in the two eyes of the same patient.

email: lingwan.chen@gmail.com

Restricted Mean Survival Time as a Function of Restriction Time

Yingchao Zhong*, University of Michigan
Douglas E. Schaubel, University of Pennsylvania

Restricted mean survival time (RMST) is a clinically interpretable and meaningful survival metric that has gained popularity in recent years. Although various estimators of RMST have differed with respect to assumptions and estimation methods, all have depended heavily on a single pre-selected value of the truncation time, L. In many practical settings, no obvious choice for L exists, such that the truncation time is specified arbitrarily. For settings in which several values of L were potentially of interest, separate analyses would have to be performed for each unique value of L. To avoid the need to pre-select a single value of L, we propose an inference framework for directly modeling RMST as a continuous function of L. Large-sample properties are derived. Simulation studies are performed to evaluate the performance of the methods in finite sample sizes. The proposed framework is applied to kidney transplant data obtained from the Scientific Registry of Transplant Recipients (SRTR).

email: zhongych@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Quantile Regression on Cause-Specific Inactivity Time

Yichen Jia*, University of Pittsburgh
Jong-Hyeon Jeong, University of Pittsburgh

In time-to-event analysis, the traditional summary measures have been based on the hazard function, survival function, quantile event time and residual lifetime. Under competing risks, furthermore, typical summary measures have been cause-specific hazard function and cumulative incidence function. Recently inactivity time has recaptured attention in the literature, being interpreted as life lost. In this paper, we propose a quantile regression model to associate the inactivity time with potential predictors under competing risks. We define the proper cumulative distribution function of the inactivity time distribution for each specific event type among those subjects who experience the same type of events during a follow-up period. A score function-type estimating equation is developed and asymptotic properties of the regression coefficient estimators are derived. A computationally efficient perturbation method is adopted to infer the regression coefficients. Simulation results show that our proposed method works well under the assumed finite sample settings. The proposed method is illustrated with a real data from a breast cancer study.

e-mail: yij22@pitt.edu

Relaxing the Independence Assumption in Relative Survival: A Parametric Approach

Reuben Adatorwovor*, University of North Carolina, Chapel Hill
Jason Fine, University of North Carolina, Chapel Hill
Aurelien Latouche, Conservatoire National des Arts et Métiers and Institut Curie, St-Cloud, France

With known cause of death (CoD), competing risk survival methods are applicable in estimating disease-specific survival. Relative survival analysis maybe used to estimate disease-specific survival in a population with unreliable CoD. This method is popular for cancer registry data regardless of CoD information. The standard estimator is the ratio of all-cause survival in the disease cohort to the expected survival from a reference population. The CoD due to disease competes with other competing mortality, inducing dependence among the CoD. This estimate is valid when deaths from disease and other causes are independent. To relax this assumption, we formulate such dependence using a copula-based likelihood to fit parametric model to the distribution of disease-specific death using registry data, with the copula assumed known and the distribution of other CoD derived from the reference population. Since dependence is nonidentifiable and unverifiable from the observed data, we propose a sensitivity analysis, in which survival is estimated across a range of dependence structures. We demonstrate the practical utility through simulations and an application to breast cancer data.

email: reubenadat@gmail.com

Estimation of Effect Measures in Survival Analysis that Allow Causal Interpretation

Kjetil Røysland*, University of Oslo

It has recently been emphasized, see (Hernan 2010), that the common interpretation of hazards, as risk of death during an infinitesimal interval for an individual is often not true. Hazards have a built-in bias since when conditioning on recent survival, we actually condition on a collider that is likely to open a non-causal pathway from the exposure to the event of interest. It does not even matter if the underlying model is causal. Aalen et.al 2015 showed that this is likely to be a problem in even RCTs. In an attempt to deal with this problem, we suggest a quite general method to estimate many other parameters in survival analysis that do not have the built the built-in bias as the hazards have, and will often allow interpretations that are more intuitive to clinicians. Our parameters are solutions of ordinary differential equations driven by cumulative hazards. These equations translate into recursively defined estimators that can be easily implemented on a computer. By using existing theory for stochastic differential equations, we are able to show that our method is both consistent and efficient.

email: kjetil.roysland@medisin.uio.no

101. RISKY BUSINESS: DIAGNOSTICS, ROC, AND PREDICTION

NMADiagT: An R package for Network Meta-Analysis of Multiple Diagnostic Tests

Boyang Lu*, University of Minnesota
Qinshu Lian, Genentech
James S. Hodges, University of Minnesota
Haitao Chu, University of Minnesota

Network meta-analysis is a commonly used approach for analyzing accuracy of diagnostic tests. At present, methods and corresponding R packages for evaluating the accuracy of diagnostic tests mostly focus on one diagnostic test with all studies either having or not having a gold standard test. However, it is more efficient to include multiple studies, and for the studies to include different diagnostic tests with or without a gold standard and using various designs. Recently, a Bayesian hierarchical model and a Bayesian hierarchical summary receiver operating characteristic model were extended to network meta-analysis of diagnostic tests to simultaneously compare multiple tests within a missing data framework. Despite the significance of these methods, the complexity of coding them hinders their application. This paper introduces an R package, NMADiagT , which provides user friendly functions for general users. This package evaluates the accuracy of multiple diagnostic tests and gives graphical representation of the results.

email: lu000083@umn.edu

ABSTRACTS & POSTER PRESENTATIONS

Informative Back-End Screening

Michael R. Stutz*, University of South Carolina
Joshua M. Tebbs, University of South Carolina

Group testing is a cost efficient method of disease screening, whereby individual specimens are pooled together and tested as a whole for the presence of disease. Traditional pooling algorithms are historically thought to be less sensitive than individual testing, likely etiologic to their underuse in disease surveillance. In this article, we propose an informative back-end screening procedure in which regression methods are used to identify specimens that have been potentially misclassified. We develop new parametric and nonparametric regression methods using the expectation-maximization algorithm which allow our procedure to be implemented with any group testing algorithm. In addition, our algorithms are the first within the group testing literature to integrate machine learning techniques. We demonstrate that with the addition of our back-end screening procedure, group testing can be both more cost efficient and more accurate than individual testing. We apply our regression methods to chlamydia screening data collected at the University of Iowa, and we develop the process whereby lab technicians can perform back-end screening using our open source R code.

email: stutzm@email.sc.edu

Patient-Reported Outcome (PRO) Assessment in Diagnostic Devices: A Novel Approach

Saryet Kucukemiroglu*, U.S. Food and Drug Administration
Manasi Sheth, U.S. Food and Drug Administration

In a public health regulatory setting, it is important for patients to have access to high-quality, safe, and effective medical devices. It is necessary to partner with patients by incorporating the patient perspective as evidence in the decision-making process, including both patient preference information (PPI) and patient-reported outcomes (PROs). PROs are often relevant in assessing diagnostic evaluations and can be used to capture a patient's everyday experience with a medical device, including experience outside of the clinician's office and the effects of treatment on a patient's activities of daily living. Furthermore, in some cases, PRO measures enable us to measure important health status information that cannot yet be detected by other measures, such as pain. To be useful to patients, researchers, and decision makers, PROs must undergo a validation process to support the accuracy and reliability of measurements from a device. Here, we present a novel approach for analyzing PROs obtained from two examples of diagnostic medical devices.

email: Saryet.Kucukemiroglu@fda.hhs.gov

A Placement-Value Based Approach to Concave ROC Curves

Soutik Ghosal*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Zhen Chen, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

The receiver operating characteristic (ROC) curve has emerged as an important statistical tool for assessing the performance of classifiers in a wide range of disciplines. While theoretical consideration aspires concave ROC curves, common models in literature do not guarantee such a feature. For practical purposes, the area under a ROC curve (AUC) can be interpreted easily when it is above 0.5. We propose a placement value-based approach to ensure the concavity of ROC curves and also extend it to allow direct covariate effect modeling. Simulation studies are provided to assess the performance of the proposed model under various scenarios, and a real data analysis is presented to investigate how estimated fetal weight (EFW) during pregnancy predicts large newborns and whether this prediction varies by the gestational age at which EFW was measured.

email: soutikghosal@gmail.com

Inference in ROC Curves for Two-Phase Nested Case-Control Biomarker Studies

Leonidas E. Bantis*, University of Kansas Medical Center
Ziding Feng, Fred Hutchinson Cancer Research Center

Two-phase nested case-control sampling designs are common in biomarker studies. In phase I, a large sample size that is representative of the target population is available and is referenced mainly for the clinical characteristics of the participants. In phase II, measurements are taken for a subsample of phase I participants due to limited resources. Since this subsampling typically uses some matching criteria, inherent bias is present. This biased sampling needs to be taken into account when clinical questions refer to the target population. While many researchers focus on the area under the ROC curve to assess the accuracy of a biomarker, such a measure does not provide an appealing clinical interpretation. Clinicians are most often interested in the performance of a biomarker at high levels of sensitivity or specificity to avoid underdiagnosis or overdiagnosis, respectively. This is driven by the seriousness of a false negative or the invasiveness of the work-up required to identify a false positive. We develop new methods for making inferences around the sensitivity at a given specificity, while also accounting for the aforementioned biased sampling.

e-mail: leobantis@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Diagnostic Evaluation of Quantitative Features of Functional Markers

Jeong Hoon Jang*, Indiana University
Amita K. Manatunga, Emory University

With advancement in data collection technology, more and more diagnostic markers are being collected as functional data. The unit of observation for each functional marker is a smooth curve defined on a time or space continuum, and its flexible and dynamic structure contains rich clinical information. In many clinical practices, it is standard to describe and diagnose a disease using "quantitative features" that characterize various interpretable patterns of a functional marker, such as area under the curve, maximum value and time to maximum. Here, we present a novel statistical framework for evaluating the diagnostic accuracy of quantitative features based on the area under the ROC curve (AUC). Using summary functionals that represent various quantitative features, we develop a non-parametric AUC estimator that addresses discreteness and noise in functional data and establish its asymptotic properties. To describe the heterogeneity of AUC in different subpopulations, we propose a sensible adaptation of a semi-parametric regression model, whose parameters are estimated by the proposed estimating equations. The proposed methods are illustrated using a renal study.

email: jeojang@iu.edu

Evaluation of Multiple Diagnostic Tests using Multi-Institutional Data with Missing Components

Jiasheng Shi*, Children's Hospital of Philadelphia
Jing Huang, University of Pennsylvania and The Children's Hospital of Philadelphia
Yong Chen, University of Pennsylvania

Multiple tests are often available for diagnosis of a certain disease as the advancement of biomedical research. However, the accuracies and costs of tests could be different. It is important to quantify and compare accuracies of tests in order to improve the disease assessment. In this talk, we will discuss a developed meta-analysis methods to improve the accuracy of estimation within a feasible computational cost by combining data from multiple studies or institution and applying composite likelihood, while a possible data missing situation may occur across all diagnostic tests from a single study or institution.

email: jiashengshi036@gmail.com

102. INTERVAL-CENSORED AND MULTIVARIATE SURVIVAL DATA

A Divide-and-Combine Approach of Multivariate Survival Analysis in Big Data

Wei Wang*, Rutgers University
Shou-En Lu, Rutgers University
Jerry Q. Cheng, New York Institute of Technology

Multivariate failure time data are frequently analyzed using the marginal proportional hazards (PH) model and the frailty model approaches. When the sample size is extraordinarily large, using either approach could face computational challenges. In this paper, we focus on the marginal model and propose a divide-and-combine (DC) approach to analyze multivariate failure time data. Specifically, we randomly divide the full data into S subsets and propose a weighted method to combine the S estimators, each from an individual subset. Under mild conditions, we show that the combined estimator is asymptotically equivalent to the estimator obtained from the full data as if the data were analyzed all at once. In addition, we propose a confidence distribution approach to perform regularized estimation. Theoretical properties, such as consistency, oracle property, and asymptotic equivalence between the full data approach and the DC approach are studied. Performance of the proposed methods, including savings in computation time, is investigated using simulation studies. A real data analysis is provided to illustrate the proposed methods.

email: ww249@sph.rutgers.edu

Nonparametric Inference for Nonhomogeneous Multi-State Processes Based on Clustered Observations

Giorgos Bakoyannis*, Indiana University

Frequently, clinical trials and observational studies involve complex event history data with multiple events. When the observations are independent, the analysis of such studies can be based on standard methods for multi-state models. However, the independence assumption is often violated, such as in multicenter studies and cluster randomized trials, which makes the use of standard methods improper. In this work we propose nonparametric estimators and two-sample tests for transition and state occupation probabilities under general multi-state models based on clustered observations. The proposed methods do not impose Markov assumptions or assumptions regarding the within-cluster dependence, and are applicable to situations with informative cluster size. The asymptotic properties of the proposed methods are rigorously established. Simulation studies show that the performance of the proposed methods is good, and that methods that ignore the within-cluster dependence can lead to invalid inferences. Finally, the methods are applied to data from a multicenter randomized controlled trial.

email: gbakogia@iu.edu

ABSTRACTS & POSTER PRESENTATIONS

Flexible, Unified Approach for Analyzing Arbitrarily-Censored and/or Left-Truncated Interval-Censored Data

Prabhashi Withana Gamage*, James Madison University
Christopher McMahan, Clemson University
Lianming Wang, University of South Carolina

A prominent feature of the survival data is that the event time is generally unavailable but is known relative to observation times. Some event times, however, can be fully observed or censored; i.e., arbitrarily-censored. Further, some studies consider people who have not experienced the event before the enrollment; i.e., left-truncated. To handle these difficulties, this work focuses on developing a unified model, under the proportional hazards assumption, which can be used to analyze such survival data. To obtain the modeling flexibility, a monotone spline representation was used to model the cumulative baseline hazard function. Through a novel data augmenting process, we developed an expectation-maximization (EM) algorithm to estimate all the parameters. The proposed EM algorithm is easy to implement and has quick convergence. The performance of the proposed methodology is evaluated through a simulation study. Moreover, the proposed method is illustrated using the data obtained from a Demographic and Health Surveys in Nigeria about child mortality and a clinical trial involves in prostate cancer study.

email: withanpw@jmu.edu

Potential Intransitivity of Win-Ratio Preferences: Is it a Problem and What Do We Do About It?

David Oakes*, University of Rochester

Pairwise preferences assigned by methods involving prioritized comparisons including the win-ratio statistic are not always transitive. Intransitivity tends to occur when length of follow-up varies substantially between individuals and rankings based of primary events disagree with those based on secondary events. The problem may be avoided by the use of the "curtailed win-ratio" statistic introduced by Oakes (2016) but at the cost of some loss of information. These issues will be discussed in the context of cardiovascular clinical trials.

e-mail: david_oakes@urmc.rochester.edu

Bayesian Analysis of Multivariate Survival Data Based on Vine Copulas

Guanyu Hu*, University of Connecticut
Dooti Roy, Boehringer Ingelheim
Dipak Dey, University of Connecticut

This talk introduces a novel copula based methodology to analyze right censored multivariate survival data. In practice, implementation of existing methodologies for analyzing multivariate survival data often leads to challenges with respect to evaluation of the likelihood and other computational issues. Using a vine copula structure, we propose a computationally tractable Bayesian modeling approach for the analysis of multivariate survival data. Extensive simulation studies show the effectness of our proposed methods. Finally, we illustrate the practical merit of our proposed methods based on real data applications.

e-mail: guanyu.hu@uconn.edu

Non-parametric estimation in an illness-death model with component-wise censoring

Anne Eaton*, University of Minnesota

In disease settings where patients are at risk for death and a serious non-fatal event, composite endpoints defined as the time until the earliest of death or the non-fatal event are often used as the primary endpoint in clinical trials. In practice, if the non-fatal event can only be detected at clinic visits and the death time is known exactly, the resulting composite endpoint exhibits "component-wise censoring". The method recommended by the FDA to estimate event-free survival for this type of data fails to account for component-wise censoring. We apply a kernel method previously proposed for a marker process in a novel way to produce a non-parametric estimator that accounts for component-wise censoring. The key insight that allows us to apply this kernel method is thinking of non-fatal event status as an intermittently observed, binary marker variable rather than thinking of time to the non-fatal event as interval censored. We also obtain estimates of the probability of being alive with the non-fatal event, and the restricted mean time patients spend in disease states. The method can be used in the setting of reversible or irreversible non-fatal events. We perform a simulation study to compare our method to existing multistate survival methods and apply the methods on data from a large randomized trial studying interventions for reducing the risk of coronary heart disease in high-risk men.

e-mail: eato0055@umn.edu

---

103. GRAPHICAL MODELS AND APPLICATIONS

Inference of Large Modified Poisson-Type Graphical Models: Application to RNA-seq Data in Childhood Atopic Asthma Studies

Rong Zhang*, University of Pittsburgh
Juan C. Celedon, UPMC Children's Hospital of Pittsburgh
Wei Chen, UPMC Children's Hospital of Pittsburgh
Zhao Ren, University of Pittsburgh

The discreteness and the high dimensions of NGS data have posed great challenges in biological network analysis. Although estimation theories for four high-dimensional modified Poisson-type graphical models have been proposed to tailor the network analysis of count-valued data, the statistical inference of these models is still largely unknown. We herein propose a novel two-step procedure in both edge-wise and global statistical inference of these modified Poisson-type graphical models using a cutting-edge generalized low-dimensional projection approach for bias correction. An extensive simulation study illustrates asymptotic normality of edge-wise inference and more accurate inferential results in multiple testing compared to the sole estimation. We have also applied our method to a novel RNA-seq gene expression data set in childhood atopic asthma in Puerto Ricans. Compared to the existing method of sole estimation or continuizing discrete values for further analysis which to some extent removes intrinsically useful information in data, our method provides more biologically meaningful results.

e-mail: roz16@pitt.edu

ABSTRACTS & POSTER PRESENTATIONS

Assisted Estimation of Gene Expression Graphical Models

Huangdi Yi*, Yale School of Public Health
Yifan Sun, Renmin University of China
Qingzhao Zhang, Xiamen University
Yang Li, Renmin University of China
Shuangge Ma, Yale School of Public Health

In the study of gene expression (GE) data, network analysis has played a uniquely important role. To accommodate the high dimensionality and low sample size and generate interpretable results, regularized estimation is usually conducted in the construction of GE Gaussian Graphical Models (GGMs). To better decipher the interconnections among GEs, conditional GGMs (cGGMs), which accommodate GEs and their regulators, have been constructed. In practical data analysis, the construction of both GGMs and cGGMs is often unsatisfactory, mainly caused by the large number of model parameters and limited sample size. In this study, we recognize that, with the regulation between GEs and regulators, the sparsity structures of the GGMs and cGGMs satisfy a hierarchy. Accordingly, we propose a joint estimation which reinforces the hierarchical structure and use GGMs to assist the construction of cGGMs and vice versa. Consistency properties are rigorously established, and an effective computational algorithm is developed. The assisted model outperforms the separate estimation of GGMs and cGGMs. Two TCGA datasets are analyzed, leading to findings different from the direct competitors.

e-mail: huangdi.yi@yale.edu

Directed Acyclic Graph Assisted Methods for Estimating Average Treatment Effect

Jingchao Sun*, University of Louisville
Maiying Kong, University of Louisville
Scott Davis Duncan, University of Louisville
Subhadip Pal, University of Louisville

Observational data can be very useful to examine average treatment effects (ATE). Propensity score based inverse probability weighting (IPW) method has been very powerful to estimate ATE if the assumptions on exchangeability and positivity hold. Directed acyclic graph (DAG) provide a feasible way to check the exchangeability, that is, the treatment and the potential outcome are independent given the variables, which block the back-door path from treatment to the potential outcome. That is, we only need to adjust a set of variables, which block all back-door paths from treatment to the potential outcomes, rather than all confounding variables. We carry out the simulation to examine the performance of the propensity score based IPW method in estimating ATE when the minimal set and maximum set of confounding variables are included in the propensity score estimation. The simulation results indicate that the performance of ATE estimation based on the minimal set of confounding variables is comparable with that with maximal set of confounding variables. We applied the method to examine if tracheostomy is a cause of death for infants based on the 2016 HCUP database.

e-mail: sunjingchao2012@hotmail.com

Gene Network Analysis Based on Single Cell RNA Sequencing Data

Meichen Dong*, University of North Carolina, Chapel Hill
Fei Zou, University of North Carolina, Chapel Hill

Integrating biological knowledge with gene regulatory networks and pathways can shed light on therapeutic targets of complex human diseases. Cutting edge single-cell sequencing data enables researchers to study networks and pathways for complex tissues and under different conditions. While constructing networks, current methods treat single cells equally distant from each other under a condition without exploiting the inter-cell relationships. Here, we propose a framework to construct gene networks based on fused-lasso regression, where single cells are first ordered after dimension reduction, and then the ordering information is used to induce constraints on the covariance structures among the single cells to construct more robust and efficient networks. The proposed method is first illustrated on synthetically simulated data and then applied to a medulloblastoma scRNA-seq dataset.

e-mail: meichen@live.unc.edu

Selection and Estimation of Conditional Graphical Models

Stephen Salerno*, University of Michigan
Yi Li, University of Michigan

Graphical models have grown increasingly popular for the analysis of high-dimensional network data. A prime example is in multi-platform genetics studies where researchers examine the relationship between gene expression and an individual's DNA profile. However, current conditional graphical models do not account for differences in gene co-regulation networks that exist due to the inherent variability between individual-specific DNA profiles. We propose a new class of Gaussian graphical models which not only accommodate the conditional mean structures of responses on a set of associated predictors, but also quantify the dependence of edges on high-dimensional DNA profiles. We present a KKT optimization algorithm for the selection and estimation of our proposed graph and explore several algorithmic considerations necessary to make this method scalable and computationally feasible. In simulation, we test our method on several biologically plausible graphs, including hub-based networks and those following a scale-free degree distribution. We also apply our methodology to a large-scale cohort study containing a wealth of cross-platform genomic data.

e-mail: salernos@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Joint Estimation of the Two-Level Gaussian Graphical Models across Multiple Classes

Inyoung Kim*, Virginia Tech
Liang Shan, University of Alabama at Birmingham

The Gaussian graphical model has been a popular tool for investigating the conditional dependency structure between random variables by estimating sparse precision matrices. However, the ability to investigate the conditional dependency structure when a two-level structure exists among the variables is still limited. Some variables are considered as higher-level variables while others are nested in these higher-level variables - the latter are called lower-level variables. Higher-level variables are not isolated; instead, they work together to accomplish certain tasks. Therefore, our main interest is to simultaneously explore conditional dependency structures among higher-level variables and among lower-level variables. Given two-level data from heterogeneous classes, we propose a method to jointly estimate the two-level Gaussian graphical models across multiple classes, so that common structures in terms of the two-level conditional dependency is shared during the estimation procedure, yet unique structures for each class are retained as well. We also demonstrate the advantages of our approach using breast cancer patient data.

e-mail: inyoungk@vt.edu

104. SUPPORT VECTOR MACHINEs, NEURAL NETWORKs AND DEEP LEARNING

ForgeNet: A Graph Deep Neural Network Model Using Tree-Based Ensemble Classifiers for Feature Graph Construction

Yunchuan Kong*, Emory University
Tianwei Yu, Emory University

A unique challenge in predictive modeling for omics data has been the small number of samples (n) versus the large number of features (p). This property brings difficulties for disease outcome classification using deep learning. Sparse learning by incorporating known functional relations between the biological units, such as the graph-embedded deep feedforward network (GEDFN), has been a solution to this issue. However, such methods require an existing feature graph, and potential misspecification of the feature graph can be harmful on classification and feature selection. To address this limitation and develop a robust classifier without relying on external knowledge, we propose a forest graph-embedded deep feedforward network (forgeNet) model, to integrate the GEDFN architecture with a forest feature graph extractor, so that the feature graph can be learned in a supervised manner and specifically constructed for a given prediction task. The resulting high classification accuracy of the simulation and the real data experiments suggests that the method is a valuable addition to sparse deep learning models for omics data.

e-mail: yunchuan.kong@emory.edu

GWAS-Based Deep Learning for Survival Prediction

Tao Sun*, University of Pittsburgh
Wei Chen, University of Pittsburgh
Ying Ding, University of Pittsburgh

Survival prediction is crucial for understanding the dynamic risks of disease progression, which enhances personalized prevention and clinical management. The massive genetics data provide unique opportunities for developing accurate survival models. Recent advances in deep learning have made remarkable progress. However, applications of deep learning in survival prediction are limited. Motivated by developing powerful prediction models for the progression of Age-related Macular Degeneration (AMD), we develop a multi-layer deep neural network (DNN) survival model to effectively extract features. Simulations are performed to compare DNN with several other survival models. Finally, using GWAS data from two large-scale clinical trials of AMD, we show that DNN not only achieves high prediction accuracy but also successfully detects clinically meaningful subgroups. Moreover, we obtain a subject-specific importance measure for each predictor from the DNN survival model. This is the first time that a DNN survival model is successfully used under both large n (>7800) and large p (millions).

e-mail: suntaojj@gmail.com

An Inferential Framework for Individualized Minimal Clinically Importance Difference with a Linear Structure

Zehua Zhou*, State University of New York at Buffalo
Jiwei Zhao, State University of New York at Buffalo

In recent years, the minimal clinically important difference (MCID) has gained a lot of attention as a useful tool to support clinical relevance in fields such as orthopedics and rheumatology. Various methods have been developed to estimate MCID, but most of them lack theoretical justification. Besides, few studies focused on the estimation of the MCID at the individual level (iMCID), which facilitates the investigation of the population heterogeneity. In this paper, we propose a general surrogate loss family to estimate the iMCID and explore the asymptotic behavior of the estimated iMCID in a linear structure. Based on the asymptotic normality, we can construct an interval estimation for linear iMCID, making the estimation more informative. The outperformance of our proposed method and the asymptotic behavior of estimated iMCID are validated through the comprehensive simulation studies and the real data analysis of a randomized controlled trial.

e-mail: zehuazho@buffalo.edu

ABSTRACTS & POSTER PRESENTATIONS

Deep Neural Networks for Survival Analysis Using Pseudo Values

Lili Zhao*, University of Michigan
Feng Dai, Merck & Co., Inc.

There has been increasing interest in modelling survival data using deep learning methods in medical research. Current approaches have focused on designing special cost functions to handle censored survival data. We propose a very different method with two steps. In the first step, we transform each subject's survival time into a series of jackknife pseudo conditional survival probabilities and then use these pseudo probabilities as a quantitative response variable in the deep neural network model. By using the pseudo values, we reduce a complex survival analysis to a standard regression problem, which greatly simplifies the neural network construction. Our two-step approach is simple, yet very flexible in making risk predictions for survival data, which is very appealing from the practice point of view.

e-mail: zhaolili@umich.edu

Neural Networks for Clustered and Longitudinal Data using Mixed Effects Models

Francesca Mandel*, University of Pennsylvania
Ian Barnett, University of Pennsylvania

Mobile health data affords new opportunities for predicting future health status by leveraging an individual's behavioral history alongside data from similar patients. Methods that incorporate both individual-level and sample-level effects are critical to using this data to its full predictive capacity. Neural networks are powerful tools for prediction, but many assume input observations are independent even when they are clustered or correlated in some way, such as in longitudinal data. Generalized linear mixed models (GLMM) provide a flexible framework for modeling longitudinal data but have poor predictive power particularly when the data is highly nonlinear. We propose a generalized neural network mixed model (GNMM) that replaces the linear fixed effect in a GLMM with the output of a feed-forward neural network. The model simultaneously accounts for the correlation structure and complex nonlinear relationship between input variables and outcomes, and it utilizes the predictive power of neural networks. We apply this approach to predict depression and anxiety levels of schizophrenic patients using longitudinal data collected from passive smartphone sensor data.

e-mail: francesca.mandel@pennmedicine.upenn.edu

105. ADVANCES IN STATISTICAL MODELING FOR MULTI-OMICS DATA INTEGRATION

Gene-Set Integrative Omics Analysis Using Tensor-Based Association Tests

Jung-Ying Tzeng*, North Carolina State University
Meng Yang, The SAS Institute
Wenbin Lu, North Carolina State University
Jeff Miecznikowski, University of Buffalo
Sheng-Mao Chang, National Cheng-Kung University

Integrative multiomics analyses integrate complementary level of information from different molecular events and have great potentials to detect novel disease genes and elucidate disease mechanisms. One major focus of integrative analysis has been on identifying gene-sets associated with clinical outcomes, and a common strategy is to regress clinical outcomes on all genomic variables in a gene set. However, such joint modeling methods encounter the challenges of high-dimensional inference, especially the sample size is usually moderate either due to research resources or missing data. In this work, we consider a tensor-based framework to enhance model efficiency for variable-wise inference. The tensor framework reduces the number of parameters by accounting for the inherent matrix structure of an individual's multiomics data and naturally incorporates the relationship among omics variables. We study the variable-specific testing procedure under tensor regression framework; we evaluate the performance of the tensor-based test using simulations and real data application on the Uterine Corpus Endometrial Carcinoma dataset from the Cancer Genome Atlas (TCGA).

e-mail: jytzeng@stat.ncsu.edu

Radiogenomic Analysis of Lower Grade Gliomas Incorporating Tumor Heterogeneity in Imaging Through Densities

Shariq Mohammed*, University of Michigan
Sebastian Kurtek, The Ohio State University
Karthik Bharath, University of Nottingham
Arvind Rao, University of Michigan
Veerabhadran Baladandayuthapani, University of Michigan

Recent technological advancements have enabled detailed studies of associations between molecular signatures of cancer and tumor heterogeneity through multi-platform data integration of both genomic and radiomic types. We will present a method to integrate and harness imaging and genomic data in patients with lower grade gliomas (LGG), a type of brain cancer, in order to develop a formal regression framework for modelling association between them. Imaging data is represented through voxel intensity probability density functions of tumor sub-regions obtained from multimodal magnetic resonance imaging (MRI), and genomic data through molecular signatures in the form of enrichment scores corresponding to their gene expression profiles. Employing a Riemannian-geometric framework we construct density-based predictors to include in a Bayesian regression model with pathway enrichment score as the response. Variable selection compatible with the grouping structure amongst the predictors, induced through the tumor sub-regions, is carried out under a group spike-and-slab prior.

e-mail: shariqm@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Regression and Clustering Models to Incorporate Multi-Layer Overlapping Group Structure in Multi-Omics Applications

George Tseng*, University of Pittsburgh

In this talk, we consider prior knowledge of a hierarchical overlapping group structure to improve variable selection in regression and clustering setting of multi-omics data analysis. For instance, a biological pathway contains tens to hundreds of genes and a gene can be mapped to multiple experimentally measured features (such as its mRNA expression, copy number variation and methylation levels of possibly multiple sites). In addition to the hierarchical structure, the groups at the same level may overlap (e.g. two pathways can share common genes). We propose Bayesian indicator models for such applications. Theoretically, we can show selection consistency and asymptotic normality for the soft-thresholding estimator of the posterior median estimates in regression setting. We demonstrate improved performance in simulations and real applications. If time allows, I will cover some recent work in outcome-guided disease subtyping model for precision medicine and its extension to multi-omics data.

e-mail: ctseng@pitt.edu

Graphical Models for Data Integration and Mediation Analysis

Min Jin Ha*, University of Texas MD Anderson Cancer Center
Veera Baladandayuthapani, University of Michigan

Integrative network modeling of data arising from multiple genomic platforms provides insight into the holistic picture of the interactive system, as well as the flow of information across many disease domains. The basic data structure consists of a sequence of hierarchically ordered datasets for each individual subject, which facilitates integration of diverse inputs, such as genomic, transcriptomic, and proteomic data. A primary analytical task in such contexts is to model the layered architecture of networks where the vertices can be naturally partitioned into ordered layers, dictated by multiple platforms, and exhibit both undirected and directed relationships. We propose a multi-layered Gaussian graphical model (mlGGM) to investigate conditional independence structures in such multi-level genomic networks. We use a Bayesian node-wise selection approach that coherently accounts for the multiple types of dependencies in mlGGM, that is used for finding causal factors for outcome variables via mediation analysis.

e-mail: mjha@mdanderson.org

106. CAUSAL INFERENCE AND NETWORK DEPENDENCE: FROM PEER EFFECTS TO THE REPLICATION CRISIS IN EPIDEMIOLOGY

Social Network Dependence, the Replication Crisis, and (in)valid Inference

Elizabeth L. Ogburn*, Johns Hopkins University

We show that social network structure can result in a new kind of structural confounding, confounding by network structure, potentially contributing to replication crises across the health and social sciences. Researchers in these fields frequently sample subjects from one or a small number of communities, schools, hospitals, etc., and while many of the limitations of such convenience samples are well-known, the issue of statistical dependence due to social network ties has not previously been addressed. A paradigmatic example of this is the Framingham Heart Study (FHS). Using a statistic that we adapted to measure network dependence, we test for network dependence and for possible confounding by network structure in several of the thousands of influential papers published using FHS data. Results suggest that some of the many decades of research on coronary heart disease, other health outcomes, and peer influence using FHS data may suffer from spurious associations and anticonservative inference due to unacknowledged network structure.

e-mail: eogburn@jhsph.edu

Nonparametric Identification of Causal Intervention Effects Under Contagion

Forrest W. Crawford*, Yale School of Public Health
Xiaoxuan Cai, Yale School of Public Health
Wen Wei Loh, University of Ghent

Defining and identifying causal intervention effects for transmissible infectious disease outcomes is challenging because a treatment – such as a vaccine – given to one individual may affect the infection outcomes of others. Epidemiologists have proposed causal estimands to quantify effects of interventions under contagion using a two-person partnership model of infection transmission. However, these simple partnership models require structural assumptions that preclude realistic infectious disease transmission dynamics, limiting their conceptual usefulness in defining and identifying causal treatment effects in empirical intervention trials. In this presentation, we propose causal intervention effects under arbitrary infectious disease transmission dynamics, and give nonparametric identification results showing how effects can be estimated in empirical trials using time-to-infection data or binary outcome data. The key insight is that contagion is a causal phenomenon that induces conditional independencies of infection outcomes that can be exploited for the identification of clinically meaningful causal estimands.

e-mail: forest.crawford@yale.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Auto-g-Computation of Network Causal Effects: Incarceration and Infection in a High Risk Network

Isabel R. Fulcher*, Harvard Medical School
Eric J. Tchetgen Tchetgen, University of Pennsylvania
Ilya Shpitser, Johns Hopkins University

The Networks, Norms, and HIV/STI Risk Among Youth (NNAHRAY) study consists of observational data on a network composed of interconnected sexual and injection drug use partnerships. Like many interpersonal networks, two persons who are not engaged in a partnership may still be associated via transitive connections. Thus, when estimating causal effects of an intervention, statistical methods should account for both (1) long-range outcome dependence between two persons on the network and (2) arbitrary forms of interference whereby one person's outcome is affected by other persons' exposures. In recent work, we developed the auto-g-computation algorithm for causal inference on a single realization of a network of connected units. This algorithm relied on certain coding estimators, which are generally inefficient by virtue of censoring observations and may be unstable even in moderately dense networks. To address this, we develop a Bayesian auto-g-computation algorithm which incorporates data on the entire network. We then evaluate the effect of prior incarceration on HIV, STI, and Hepatitis C prevalence on the NNAHRAY network.

e-mail: isabel_fulcher@hms.harvard.edu

Heterogeneous Causal Effects under Network Interference

Laura Forastiere*, Yale University
Costanza Tortú, IMT Lucca, Italy
Falco Bargagli-Stoffi, IMT Lucca, Italy

Spillovers are a crucial component in understanding the full impact of interventions at the population-level. Information about spillovers of health interventions would support decisions about how best to deliver interventions and can be used to guide public funds allocation. In fact, policy makers can gain from understanding the heterogeneity of spillover effects to identify the most contagious or influential individuals and those who are more susceptible. Social network targeting shows great promise in behavioral change interventions and policy makers are in need of guidance on how best to design their programs so as to use social networks to maximize adoption of healthy behaviors for improving community health.Under a causal inference framework, we develop machine learning methods to assess the heterogeneity of treatment and spillover effects in a two-stage randomized experiment with clustered networks.

e-mail: laura.forastiere@yale.edu

---

107. FLEXIBLE SPATIO-TEMPORAL MODELS FOR ENVIRONMENTAL AND ECOLOGICAL PROCESSES

Evaluating Proxy Influence in Assimilated Paleoclimate Reconstructions - Testing the Exchangeability of Two Ensembles of Spatial Processes

Bo Li*, University of Illinois at Urbana-Champaign
Trevor Harris, University of Illinois at Urbana-Champaign
Nathan Steiger, Columbia University
Jason Smerdon, Columbia University
Naveen Narisetty, University of Illinois at Urbana-Champaign
J. Derek Tucker, Sandia National Lab

Climate field reconstructions (CFR) attempt to estimate spatiotemporal fields of climate variables in the past using climate proxies. While many different CFR products and methods exist, Data Assimilation (DA) methods are a recent and promising new means of deriving CFRs that optimally fuse large collections of proxies with climate model information. Despite the growing application of DA-based CFRs, little is understood about how much the assimilated proxies change the statistical properties of the climate model data. We propose a robust and computationally efficient method, based on functional data depth, to evaluate the differences in the distributions of two spatiotemporal processes. We apply our test to study global and regional proxy influence in DA-based CFRs by comparing the background and analysis states. We find that the analysis states are significantly altered from the climate-model-based background states due to the assimilation of paleoclimate proxies. Moreover, the difference between the analysis and background states increases as the number of assimilated prox- ies increases, even in regions far beyond proxy collection sites.

e-mail: libo@illinois.edu

ABSTRACTS & POSTER PRESENTATIONS

Fusing Multiple Existing Space-Time Categorical Land Cover Datasets

Amanda S. Hering*, Baylor University
Nicolás Rodríguez-Jeangros, Colorado School of Mines
John E. McCray, Colorado School of Mines

Land cover (LC) products are derived primarily from satellite spectral imagery and are essential inputs for environmental studies. However, existing LC products have different temporal and spatial resolutions and different LC classes that rarely provide enough detail. We develop a method for fusing multiple existing LC products to produce a single LC record for a large spatial-temporal grid, referred to as spatiotemporal categorical map fusion (SCaMF). We first reconcile the LC classes across products and then present a probabilistic weighted nearest neighbor estimator of LC class. This estimator depends on three unknown parameters that are estimated using numerical optimization to maximize an agreement criterion that we define. We illustrate the method using six LC products over the Rocky Mountains and show the improvement gained by using data-driven information describing the spatial-temporal behavior of each LC class. Additional flexibility to adapt to local nonstationarities and to produce more detailed classes is incorporated to produce a yearly product from 1983-2012 on 30 and 50 m grids.

e-mail: mandy_hering@baylor.edu

Inverse Reinforcement Learning for Animal Behavior from Environmental Cues

Toryn L.J. Schafer*, University of Missouri
Christopher K. Wikle, University of Missouri

Animal movement is a complex spatio-temporal process due to interactions of environmental and social cues. Agent-based methods allow for defining simple rules that generate complex group behaviors. The governing rules of such models are typically set a priori and parameters are learned from observed animal trajectories. Instead of making simplifying assumptions across all anticipated scenarios, inverse reinforcement learning provides inference on the short-term (local) rules governing long term behavior policies by using properties of a Markov decision process. Learning agent rewards is enhanced through the use of deep models.

e-mail: toryn.27@gmail.com

HIGH-DIMENSIONAL MULTIVARIATE GEOSTATISTICS: A BAYESIAN MATRIX-NORMAL APPROACH

Lu Zhang*, UCLA-Fielding School of Public Health

A key challenge facing statisticians is the modeling for massive spatial datasets collected over a large number of locations, likely into the tens of millions, and a considerably large number of spatially oriented outcomes or dependent variables. Spatial process models in such settings require multivariate spatial processes constructed on high-dimensions, where dimension refers to both the number of locations and the number of outcomes. We develop Bayesian multivariate geostatistical models through a Matrix-Normal approach, showing how to fit them to large-scale multivariate spatial datasets and how to efficiently obtain Bayesian inference over a high-dimensional parameter space.

e-mail: Lu.Zhang@ucla.edu

---

108. RECENT ADVANCES IN NEUROIMAGING ANALYTICS

Covariance Regression in Brain Imaging

Brian S. Caffo*, Johns Hopkins University
Yi Zhao, Indiana University Purdue University Indianapolis
Bingkai Wang, Johns Hopkins University
Xi (Rossi) Luo, University of Texas Health Science Center at Houston

In this talk, we cover methodology for jointly analyzing a collection of covariance or correlation matrices that depend on other variables. This covariance-as-an-outcome regression problem arises commonly in the study of brain imaging, where the covariance matrix in question is an estimate of functional or structural connectivity. Two main approaches to covariance regression exists: outer product models and joint diagonalization approaches. We investigate joint diagonalization approaches and discuss the benefits and costs of this solution. We distinguish between diagonalization approaches where the eigenvectors are selected in the absence of covariate information and those that chose the eigenvectors so that the result regression model holds best. The methods are applied to resting state functional magnetic resonance imaging data in a study of aphasia and potential interventions.

e-mail: bcaffo@jhsph.edu

Bayesian Modeling of Multiple Structural Connectivity Networks during the Progression of Alzheimer's Disease

Christine B. Peterson*, University of Texas MD Anderson Cancer Center

In this talk, I will discuss a novel approach for inference of multiple networks with related edge values across groups. Specifically, we propose learning a Gaussian graphical model for each group within a joint framework, where we rely on Bayesian hierarchical priors to link the precision matrix entries across groups. Our proposal differs from existing approaches in that it flexibly learns which groups have the most similar edge values, and accounts for the strength of connection (rather than only edge presence or absence) when sharing information across groups. We apply this method to infer structural changes in brain connectivity resulting from the progression of Alzheimer's Disease. Our results identify key network alterations which may reflect disruptions to the healthy brain. We also illustrate the proposed method through simulations, where we demonstrate its performance in structure learning and precision matrix estimation with respect to alternative approaches.

e-mail: cbpeterson@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Modeling Lead-Lag Dynamics in High Dimensional Time Series

Hernando Ombao*, King Abdullah University of Science and Technology
Chee-Ming Ting, King Abdullah University of Science and Technology
Marco Pinto, Oslo Metropolitan University

In this talk, we tackle one of the fundamental problems in neuroimaging which is characterizing and conducting formal statistical inference on lead-lag dynamics in brain signals. We focus on brain signals with high temporal resolution (e.g., electroencephalograms and local field potentials). The challenges to analyzing these brain signals are data size, high dimensionality, low signal-to-noise ratio and inherent complexity. We will first discuss new dependence measures for describing the non-linear dynamics of the oscillatory activity. These measures capture the delicate time-varying relationship between phases and amplitudes. However, it is challenging to fully implement these measures under the high-dimensional setting. We explore some biologically-guided dimension reduction techniques and propose multi-state models under which we can formally define non-linear dependence and rigorously conduct statistical inference. The proposed method and model will be applied to electroencephalograms recorded in an experiment to study the synchronicity between dyads (two human participants). The goal is to determine if there is synchronicity between two brains.

e-mail: hernando.ombao@kaust.edu.sa

Modeling Positive Definite Matrices in Diffusion Tensor Imaging

Dipankar Bandyopadhyay*, Virginia Commonwealth University
Zhou Lan, The Pennsylvania State University
Brian J. Reich, North Carolina State University

Diffusion tensor imaging (DTI), a neuroimaging technique, produces voxel-level spatially referenced positive definite (p.d) matrices as responses, which available geostatistical modeling tools are unable to handle efficiently. In this talk, we propose a matrix-variate semiparametric mixture model under a Bayesian paradigm, where the p.d. matrices are distributed as a mixture of inverse Wishart distributions, with the spatial dependence captured by a Markov model for the mixture component labels. The nice conjugacy and double Metropolis-Hastings algorithm result in a fast and elegant Bayesian computing. Simulation results show that our method is powerful and robust, with improved performances compared to univariate alternatives. Furthermore, we apply this method to investigate the effect of cocaine use on brain structure.

e-mail: bandyopd@gmail.com

109. NOVEL TENSOR METHODS FOR COMPLEX BIOMEDICAL DATA

Generalized Tensor Regression with Covariates on Multiple Modes

Miaoyan Wang*, University of Wisconsin, Madison
Zhuoyan Xu, University of Wisconsin, Madison
Jiaxin Hu, University of Wisconsin, Madison

We consider the problem of tensor-response regression given covariates on multiple modes. Such data problems arise frequently in applications such as neuroimaging, network analysis, and spatial-temporal modeling. We propose a new family of tensor response regression models that incorporate covariates, and establish the theoretical accuracy guarantees. Unlike earlier methods, our method allows the contribution of covariates from multiple modes, whenever available. An efficient alternating updating algorithm is further developed. Our proposal handles a broad range of data types, including continuous, count, and binary observations. The simulation study demonstrates the outperformance of our appraoch over classical multivariate regression. We apply the method to diffusion tensor imaging (DTI) data from human connection project. Our method is able to identify the global connectivity network pattern and to pinpoint the local regions that are strongly affected by age/gender.

e-mail: miaoyan.wang@wisc.edu

Co-Manifold Learning on Tensors

Eric Chi*, North Carolina State University

We introduce a new method for performing joint dimension reduction, or manifold learning, along the modes of a data tensor, in order to re-order the fibers along each mode so that the resulting permuted tensor is smooth. Our approach leverages recent work on a convex formulation of the co-clustering problem to construct multi scale metrics along each tensor mode. We illustrate how our method can identify the coupled intrinsic geometries in simulated and real data examples.

e-mail: eric_chi@ncsu.edu

ABSTRACTS & POSTER PRESENTATIONS

Nonparametric Regression for Brain Imaging Data Analysis

Weining Shen*, University of California, Irvine

With the rapid growth of neuroimaging technologies, a great effort has been dedicated recently to investigate the dynamic changes in brain activity. Examples include time course calcium imaging and dynamic brain functional connectivity. In this talk, I will discuss a novel nonparametric matrix response regression model to characterize the nonlinear association between 2D image outcomes and predictors such as time and patient information. The estimation procedure can be formulated as a nuclear norm regularization problem, which can capture the underlying low-rank structure of the dynamic 2D images. We present a computationally efficient algorithm, derive the asymptotic theory and show that the method outperforms other existing approaches in simulations. We then apply the proposed method to a calcium imaging study for estimating the change of fluorescent intensities of neurons, and an electroencephalography study for a comparison in the dynamic connectivity covariance matrices between alcoholic and control individuals.

e-mail: weinings@uci.edu

Brain Regions Identified as Being Associated with Verbal Reasoning through the Use of Imaging Regression via Internal Variation

Xuan Bi*, University of Minnesota
Long Feng, Yale University
Heping Zhang, Yale University

Brain-imaging data have been increasingly used to understand intellectual disabilities. Despite significant progress in biomedical research, the neurological mechanisms remain unknown. We investigate verbal reasoning, a reliable measure of individuals' general intellectual abilities, and develop a class of high-order imaging regression models to identify relevant brain subregions. A key novelty is to take advantage of spatial brain structures, and specifically the piecewise smooth nature of most imaging coefficients in the form of high-order tensors. Our approach provides an effective method for identifying brain subregions potentially underlying certain intellectual disabilities. The idea is a carefully constructed concept called Internal Variation. We present our results from the analysis of the Philadelphia Neurodevelopmental Cohort for which we preprocessed magnetic resonance images from 978 individuals. Our analysis identified a subregion across the cingulate cortex and the corpus callosum as being associated with individuals' verbal reasoning ability, which, to the best of our knowledge, is a novel region that has not been reported in the literature.

e-mail: xbi@umn.edu

110. INTEGRATIVE ANALYSIS OF CLINICAL TRIALS AND REAL-WORLD EVIDENCE STUDIES

On Using Electronic Health Records to Improve Optimal Treatment Rules in Randomized Trials

Peng Wu*, Columbia University and Visa Inc.
Donglin Zeng, University of North Carolina, Chapel Hill
Haoda Fu, Eli Lilly and Company
Yuanjia Wang, Columbia University

Data from randomized controlled trials (RCTs) are used to infer valid individualized treatment rules (ITRs) using statistical learning methods. However, RCTs are usually conducted under specific criteria, thus limiting their generalizability to a broader population in real world. Because electronic health records (EHRs) document treatment prescriptions in the real world, transferring information in EHRs to RCTs, if done appropriately, could potentially improve the performance of ITRs, in terms of precision and generalizability. In this work, we propose a new domain adaptation method to learn ITRs by incorporating information from EHRs. We first pre-train "super" features from EHRs that summarize physician treatment decisions and patient observed benefits. We then augment the feature space of the RCT and learn the optimal ITRs by stratifying by super features using subjects enrolled in RCT. We adopt Q-learning and a modified matched-learning algorithm for estimation. We present heuristic justification of our method and conduct simulation studies to demonstrate the performance of super features.

e-mail: pengwu2394@gmail.com

Making Use of Information Contained in Existing Black-Box-Type Risk Calculators

Peisong Han*, University of Michigan
Jeremy M.G. Taylor, University of Michigan
Bhramar Mukherjee, University of Michigan

Consider the setting where (i) individual-level data are collected to build a regression model for the association between observing an event of interest and certain covariates, and (ii) some risk calculators predicting the risk of the event using less detailed covariates are available, possibly as black boxes with little information available about how they were built. We propose a general empirical-likelihood-based framework to integrate the rich auxiliary information contained in the calculators into fitting the regression model in order to improve the efficiency for the estimation of regression parameters. Both theoretical and numerical investigations show that the calculator information can help substantially reduce the variance of regression parameter estimation. As an application, we study the dependence of the risk of high grade prostate cancer on both conventional risk factors and newly identified biomarkers by integrating information from the Prostate Biopsy Collaborative Group (PBCG) risk calculator, which was built based on conventional risk factors alone.

e-mail: peisong@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

Integrative Analysis of Randomized Clinical Trials with Real World Evidence Studies

Lin Dong*, Wells Fargo Bank
Shu Yang, North Carolina State University

In this paper, we leverage the complementing features of randomized clinical trials (RCT) and real-world evidence (RWE) to estimate the average treatment effect of the target population. We propose a calibration weighting estimator that enforces the covariate balance between the RCT and RWE study. We further propose a doubly robust augmented calibration weighting estimator that can be applied in the event that treatment and outcome information is also available from the RWE study. This estimator achieves the semiparametric efficiency bound. We establish asymptotic results under mild regularity conditions, and examine the finite sample performances of the proposed estimators by simulation experiments. We apply our proposed methods to estimate the effect of adjuvant chemotherapy in early-stage resected non–small-cell lung cancer integrating data from a RCT and a sample from the National Cancer Database.

e-mail: ldong7@ncsu.edu

Risk Projection for Time-to-Event Outcome Leveraging External Summary Statistics with Source Individual-Level Data

Jiayin Zheng*, Fred Hutchinson Cancer Research Center
Li Hsu, Fred Hutchinson Cancer Research Center
Yingye Zheng, Fred Hutchinson Cancer Research Center

Risk stratification based on prediction models for chronic diseases has become increasingly important in medical practice. When a prediction model developed in an existing source study (cohort/case-control study/trial) is applied to a targeted population, due to potential discrepancy in baseline disease incidence and shift in patient composition, the model-based absolute risk may under- or over-estimates the observed risk in the new population/cohort. Remedy of such a poorly calibrated prediction is needed for proper medical decision-making. In this article, assuming the relative risks of predictors same between the two populations, we propose a novel weighted estimating equation approach to re-calibrate the projected risk for the targeted population through updating the baseline risk, leveraging known overall disease-free survival probability and summary information of risk factors from the targeted population. By solving the weighted estimating equation, we obtain an estimator that is more efficient when the risk factor distributions are same between the source and targeted populations, and more robust when they differ.

e-mail: statzjy@gmail.com

111. CLUSTERED DATA METHODS

Modeling Tooth-Loss using Inverse Probability Censoring Weights in Longitudinal Clustered Data with Informative Cluster Size

Aya A. Mitani*, Harvard T. H. Chan School of Public Health
Elizabeth K. Kaye, Boston University Henry M. Goldman School of Dental Medicine
Kerrie P. Nelson, Boston University School of Public Health

Periodontal disease is a serious gum infection that may lead to loss of teeth. Using standard marginal models to study the association between person-level predictors and tooth-level outcomes can lead to biased estimates because the independence assumption between the outcome (periodontal disease) and cluster size (number of teeth per person) is violated. Specifically, the baseline number of teeth of a patient is informative. A cluster-weighted generalized estimating equations (CWGEE) approach can be used to obtain unbiased marginal inference from data with informative cluster size (ICS). However, in many longitudinal studies of dental health, the rate of tooth-loss over time is also informative, creating a missing at random data mechanism. We propose a novel modeling approach that incorporates inverse probability censoring weights into CWGEE with binary outcomes to account for ICS and informative tooth-loss over time. In an extensive simulation study, we demonstrate that results obtained from our proposed method yield lower bias and excellent coverage probability compared to those obtained from traditional methods which do not account for ICS or informative drop-out.

e-mail: amitani@hsph.harvard.edu

Partially Pooled Propensity Score Models for Average Treatment Effect Estimation with Multilevel Data

Youjin Lee*, University of Pennsylvania
Trang Nguyen, Johns Hopkins Bloomberg School of Public Health
Elizabeth Stuart, Johns Hopkins Bloomberg School of Public Health

Causal inference analyses often use existing observational data, which in many cases has some clustering of individuals. In this paper we discuss propensity score weighting methods in a multilevel setting where within clusters individuals share unmeasured variables that are related to treatment assignment and the potential outcomes. We focus in particular on settings where multilevel modeling approaches are either not feasible or not useful due to the presence of a large number of small clusters. We found, both through numerical experiments and theoretical derivations, that a strategy of grouping clusters with similar treatment prevalence and estimating propensity scores within such cluster groups is effective in reducing bias from unmeasured cluster-level covariates. We apply our proposed method in evaluating the effectiveness of center-based pre-school program participation on children's achievement at kindergarten, using the Early Childhood Longitudinal Study, Kindergarten data.

e-mail: youjin.lee@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

Outcome-Guided Disease Subtyping for High-Dimensional Omics Data

Peng Liu*, University of Pittsburgh
Lu Tang, University of Pittsburgh
George Tseng, University of Pittsburgh

High-throughput technologies have been used to identify disease subtypes that could not be observed otherwise. The classical unsupervised clustering strategy concerns primarily the identification of subpopulations with similar feature patterns. However, as confounders (e.g. gender, age) related features may dominate clustering, the resulting clusters may fail to capture clinically meaningful subtypes. Therefore, an outcome guided subtyping procedure is necessary. Existing methods such as supervised clustering apply a two-stage approach which depends on an arbitrary screening of outcome associated features. In this paper, we propose a unified latent generative model to perform outcome-guided disease subtyping, which guarantees the resulting subtypes to concern the disease of interest. The new method performs feature selection, latent subtype characterization and outcome prediction simultaneously. Simulations and application to a complex lung dataset with transcriptomic and phenomic data demonstrate advantages of the new method suitable to explore toward precision medicine.

e-mail: pel67@pitt.edu

The Impact of Sample Size Re-Estimation using Baseline ICC in Cluster Randomized Trials: A Simulation Study

Kaleab Z. Abebe*, University of Pittsburgh
Kelley A. Jones, University of Pittsburgh
Taylor Paglisotti, University of Pittsburgh
Elizabeth Miller, University of Pittsburgh
Daniel J. Tancredi, University of California, Davis

Power and sample size for cluster randomized trials (CRTs) depend heavily on assumed intra-class correlation coefficients (ICCs), a measure of closeness of the outcomes from individuals that make up each cluster, but often assumed ICCs are uncertain. Few CRTs (approx. 35%) report ICCs at the end of study to inform future trial design, and fewer still report uncertainty intervals for ICC estimates. Inaccurate ICCs lead to CRTs that are inappropriately powered. Sample size re-estimation (SSR) using an internal pilot study is one way to mitigate the problem of inaccurate ICC estimates. Yet, this method is more common in adaptive design literature, and the optimal size of the internal pilot is unclear. Few, if any CRTs that include a baseline phase have used the ICC estimates from baseline data in a blinded SSR. Using simulation, we addressed two aims: to 1) characterize the association between baseline and end-of-study ICCs; and 2) estimate the increase in power of a blinded SSR based on baseline ICCs. Additionally, we use data from a recently completed CRT of a gender violence prevention program to see how this proposed SSR design would have changed the primary results.

e-mail: kza3@pitt.edu

Hypothesis Testing for Community Detection in Network Data

Chetkar Jha*, University of Pennsylvania
Mingyao Li, University of Pennsylvania
Ian Barnett, University of Pennsylvania

Community detection is an important problem in network analysis. Recently, Lei (2016) proposed a sequential hypothesis testing approach for detecting the true number of communities in Stochastic Block Models (SBM) and Degree Corrected Stochastic Block Models (DCSBM). Their approach assumes that the underlying graph model is dense and all the blocks are balanced. However, these assumptions are restrictive in the sense that the real-life graphs tend to be sparse and contains imbalanced blocks. We study Lei (2016)'s approach when the assumptions of denseness of graph and balanced size is relaxed. Specifically, we derive conditions on composition of balanced and imbalanced blocks for which we can guarantee the consistency of Lei(2016)'s approach. We demonstrate the usefulness of our approach through simulation and real data data applications.

e-mail: Chetkar.Jha@Pennmedicine.upenn.edu

On the Interplay Between Exposure Misclassification and Informative Cluster Size

Glen McGee*, Harvard University
Marianthi-Anna Kioumourtzoglou, Columbia University
Marc G. Weisskopf, Harvard University
Sebastien Haneuse, Harvard University
Brent A. Coull, Harvard University

We study the impact of exposure misclassification when cluster size is potentially informative (i.e. related to outcomes) and when misclassification is differential by cluster size. First, we show that misclassification in an exposure related to cluster size can induce informativeness when cluster size would otherwise be non-informative. Second, we show that misclassification that is differential by informative cluster size can not only attenuate estimates of exposure effects but even inflate or reverse the sign of estimates. To correct for bias in estimating marginal parameters, we propose: (i) an observed likelihood approach for joint marginalized models of cluster size and outcomes and (ii) an expected estimating equations approach. Although we focus on estimating marginal parameters, a corollary is that the observed likelihood approach permits valid inference for conditional parameters as well. Using data from the Nurses Health Study II, we compare the results of the proposed methods when applied to motivating data on the multigenerational effect of in-utero diethylstilbestrol exposure on attention-deficit/hyperactivity disorder.

e-mail: glen.w.mcgee@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

An Alternative to the Logistic GLMM with Normal Random Effects for Estimating Dose Response in the Presence of Extreme Between Subject Heterogeneity

Joe Bible*, Clemson University
Christopher McMahan, Clemson University

We propose an alternative to the logistic GLMM with normal random effects for modeling binary outcomes for data that exhibit substantial between subject heterogeneity for modelling dose response. Our method is motivated by and applied to a meta-study of six opiate cessation trials where there is considerable variation in relapse rates between individuals within a given trial as well as between individuals in different trials.

e-mail: jbible831@gmail.com

112. SUBGROUP ANALYSIS

Inference on Selected Subgroups in Clinical Trials

Xinzhou Guo*, Harvard University
Xuming He, University of Michigan

When existing clinical trial data suggest a promising subgroup, we must address the question of how good the selected subgroup really is. The usual statistical inference applied to the selected subgroup, assuming that the subgroup is chosen independent of the data, will lead to overly optimistic evaluation of the selected subgroup. In this paper, we address the issue of bias and develop a bootstrap-based inference procedure for the best selected subgroup effect. The proposed inference procedure is model-free, easy to compute, and asymptotically sharp. We demonstrate the merit of our proposed method by re-analyzing the MONET1 trial and show that how the subgroup is selected post hoc should play an important role in any statistical analysis.

e-mail: xinzhoug@umich.edu

A Simultaneous Inference Procedure to Identify Subgroups in Targeted Therapy Development with Time-to-Event Outcomes

Yue Wei*, University of Pittsburgh
Jason Hsu, The Ohio State University
Ying Ding, University of Pittsburgh

The uptake of targeted therapies has largely changed the field of medicine. Instead of the “one-fits-all” approach, one aspect is to develop drugs that target a subgroup of patients. Usually many markers need to be tested and within each one, it is necessary to infer treatment effect in marker-defined groups and their combinations. In this research, we develop a simultaneous inference procedure to identify subgroups with enhanced treatment efficacy in clinical trials with time-to-event outcomes. Specifically, we provide simultaneous confidence intervals based on a logic-respecting efficacy measure, which appropriately adjust within- and between-marker multiplicities for comparing groups. Realistic simulations are conducted using true genotype data and various efficacy scenarios to evaluate method performance. We recommend practically useful rules for selecting candidate markers and subgroups for targeting. Finally we apply the method to an Age-related macular degeneration (AMD) study and successfully identify subgroups that exhibit enhanced efficacy in delaying AMD progression for the antioxidant supplements treatment.

e-mail: yuw95@pitt.edu

Cross-Platform Omics Prediction (CPOP) Procedure Enables Precision Medicine

Kevin Y.X. Wang*, The University of Sydney
Varsha Tembe, Melanoma Institute Australia and The University of Sydney
Gullietta Pupo, Melanoma Institute Australia and The University of Sydney
Garth Tarr, The University of Sydney
Samuel Mueller, The University of Sydney
Graham Mann, Melanoma Institute Australia and The University of Sydney
Jean Y.H. Yang, The University of Sydney

Risk models separately constructed on two independent omics data for prediction of the same biological outcome are typically not “transferable” as the same features in independent omics data rarely share the same scale. Models that do not take between-data variations into account typically exhibit poor prediction power across datasets and platforms. Classical approaches including data normalisation and refitting model parameters are impractical if the data or model are placed in lock-down as is typical in clinical implementation. To this end, we propose a new procedure, Cross-Platform Omics Prediction (CPOP) for building clinically implementable models using omics data. CPOP preferentially selects the most statistically stable ratio-based features across multiple omics datasets with model estimate stabilisation. Application of CPOP on a collection of cross-platform data and a prospective experiment demonstrates that CPOP can select features that are stable across multiple data without model re-training and data re-normalisation. Together, we show that CPOP can construct reproducible models that ultimately strengthens precision medicine research.

e-mail: kevin.wang@sydney.edu.au

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Subgroup Analysis in Regression using Mixture Models

Yunju Im*, University of Iowa
Aixin Tan, University of Iowa

Heterogeneity occurs in many regression problems, where members from different latent subgroups respond differently to the covariates of interest (e.g., treatments) even after adjusting for other covariates. To identify such subgroups, our work adopts a Bayesian model called the Mixture of Finite Mixtures (MFM), for which the number of subgroups needs not be specified a priori and is modeled as a random variable. The Bayesian MFM model was not commonly used in earlier applications largely due to computational difficulties. Instead, an alternative Bayesian model, the Dirichlet Process Mixture Model (DPMM) has been widely used for clustering although it is a misspecified model for many applications. The popularity of DPMM is partly due to its mathematical properties that enable efficient computing algorithms. We propose a class of conditional MFMs tailored to regression setups and solve the computing problem by extending the results in Miller and Harrison (2018). Using simulated and real data, we show the benefits of our conditional MFM, compared to that of existing frequentist methods, the DPMM, and the original MFM models in various setup.

e-mail: yunju-im@uiowa.edu

Adaptive Subgroup Identification in Phase I-II Clinical Trials

Alexandra M. Curtis*, University of Iowa
Brian J. Smith, University of Iowa
Andrew G. Chapple, Louisiana State University School of Public Health

In most models and algorithms for dose-finding clinical trials, trial participants are assumed to be homogeneous – the optimal dose is the same for all those who qualify for the trial. However, if there are heterogeneous populations that may benefit from the same treatment, it is inefficient to conduct dose-finding separately for each group, and assuming homogeneity across all sub-populations may lead to identification of the incorrect dose for some (or all) subgroups. To accommodate heterogeneity in dose-finding trials when both efficacy and toxicity outcomes must be used to identify the optimal dose (as in immunotherapeutic oncology treatments), we propose an adaptive Bayesian clustering method which builds on the Sub-TITE clustering model described by Chapple and Thall (2018) for phase I trials. We provide a comparison of operating characteristics between our method, Bayesian hierarchical models (Cunanan and Koopmeiners 2018), and the method of fitting separate models for each subgroup in a variety of relevant scenarios, as well as randomly generated scenarios.

e-mail: alexandra-curtis@uiowa.edu

Identifying Effect Modifiers and Subgroups that May Benefit from Treatment when the Number of Covariates is Large

John A. Craycroft*, University of Louisville
Maiying Kong, University of Louisville
Subhadip Pal, University of Louisville

Observational studies differ from experimental studies in that assignment of subjects to treatments is not randomized but rather occurs due to natural mechanisms, which are usually hidden from the researchers. Yet objectives of the two studies are frequently the same: identify the average treatment effect (ATE) of some exposure on a population. Furthermore, in both types of studies it is frequently of interest to learn whether ATE differs across particular subgroups of subjects. While these objectives can be achieved directly in an experimental context due to the design imposed on the study, in an observational study special care must be taken to avoid confounding bias in ATE estimates, particularly when the number of covariates is large. This research focuses on avoiding confounding bias in estimation of ATE, with special focus on identifying effect modifiers. We present a method which efficiently selects effect modifiers from the set of covariates & computes unbiased estimates for subgroups of interest. The goal is to deliver more targeted advice describing circumstances where a treatment may be more beneficial for one/some groups vs. others.

e-mail: john.craycroft@louisville.edu

113. FUNCTIONAL DATA ANALYSIS: BELOW THE SURFACE

Imaging Genetics: Where the Statistics of fMRI and Genome-Wide Association Studies Collide

Kristen N. Knight*, University of Georgia

The goal of the emerging field of imaging genetics is to combine the neuroimaging power of fMRI and the sequencing of the complex human genome to produce a unified approach to understanding the acquisition and progression of psychiatric illnesses. Despite the work towards discovery in the past decade, the groundwork of statistical methodology for imaging genetics remains in its infancy. Numerous challenges exist for this BIG data problem, such as easing the computational burden, minimizing false positives, all while accounting for human heterogeneity. I will outline the current direction of imaging genetics research and highlight the short-comings of past studies through simulation and replication studies. Then, I will introduce my own approach by providing a applicable framework for imaging genetics analyzes. Emphasis will be placed on dimension reduction and candidate-gene selection for genome-wide association studies with a neuroimaging phenotype. Data gleaned from the Alzheimers Disease Neuroimaging Initiative (ADNI) will be analyzed with this new approach. Results will be compared among healthy, mild cognitive impaired and Alzheimer's Disease diagnosed subjects.

email: knk84226@uga.edu

ABSTRACTS & POSTER PRESENTATIONS

Bayesian Quantile Monotone Single-Index Model for Bounded Response Using Functional and Scalar Predictors

Bradley B. Hupf*, Florida State University
Debajyoti Sinha, Florida State University
Eric Chicken, Florida State University
Greg Hajcak, Florida State University

Functional predictors with bounded responses have become increasingly common in recent mental health studies. We propose a novel quantile estimator using a single-index model and a monotone link function within a Bayesian framework. Compared to existing methods, our proposed model offers clear practical advantages: a clinically interpretable index with a non-decreasing relationship with a pre-specified quantile of the bounded response, accommodation of unknown non-linear effects and interaction effects of predictors, and an objective data-driven identification of a biomarker for dysphoria (called reward positivity) derived from the functional predictor. Our proposed method outperforms competitors based on unrestricted single-index models and is therefore the more practical choice for analysis. We examine the properties of our estimator through a simulation study and on our motivating application, a clinical study of adolescent dysphoria (a bounded response) and a functional neural covariate.

e-mail: bradley.hupf@stat.fsu.edu

Sparse Log-Contrast Regression with Functional Compositional Predictors: Linking Gut Microbiome Trajectory in Early Postnatal Period to Neurobehavioral Development of Preterm Infants

Zhe Sun*, University of Connecticut
Wanli Xu, University of Connecticut
Xiaomei Cong, University of Connecticut
Gen Li, Columbia University
Kun Chen, University of Connecticut

It is hypothesized that stressful early life experience of preterm infant is imprinting gut microbiome and hence certain microbiome markers are predictive of later infant neurodevelopment. A preterm infant study was conducted by collecting infant fecal samples during the infants' first month of postnatal age, resulting in functional compositional microbiome data. To identify microbiome markers and estimate how the gut microbiome compositions during early postnatal stage impact later neurobehavioral outcomes of the preterm infants, we innovate a sparse log-contrast regression with functional compositional predictors. The functional simplex structure is strictly preserved, and the functional compositional predictors are allowed to have sparse, smoothly varying, and accumulating effects on the outcome through time. We develop an efficient algorithm and obtain theoretical performance guarantees. Our approach yields insightful results in that the identified microbiome markers and the estimated time dynamics of their impact on the neurobehavioral outcome shed lights on the linkage between stress accumulation in early postnatal stage and neurodevelpomental process of infants.

e-mail: zhe.sun@uconn.edu

Principle ERP Reduction and Analysis

Emilie Campos*, University of California, Los Angeles
Chad Hazlett, University of California, Los Angeles
Patricia Tan, University of California, Los Angeles
Holly Truong, University of California, Los Angeles
Sandra Loo, University of California, Los Angeles
Charlotte DiStefano, University of California, Los Angeles
Shafali Jeste, University of California, Los Angeles
Damla Senturk, University of California, Los Angeles

Event-related potentials (ERP) waveforms are the summation of many overlapping signals. Changes in the peak or mean amplitude of a waveform over a given time period, therefore, cannot reliably be attributed to a particular ERP component of ex ante interest. Though this problem is widely recognized, it is not well addressed in practice. Our approach begins by presuming that any observed ERP waveform — at any electrode, for any trial type, and for any participant — is approximately a weighted combination of signals from an underlying set of what we refer to as principle ERPs, or pERPs. First, we propose the principle ERP reduction (pERP-RED) algorithm for investigators to estimate a suitable set of pERPs from their data, which may span multiple tasks. Next, we provide tools and illustrations of pERP-space analysis, whereby observed ERPs are decomposed into the amplitudes of the contributing pERPs, which can be contrasted across conditions or groups. We demonstrate this suite of tools through simulations and on real data collected from multiple experiments on participants diagnosed with Autism Spectrum Disorder and Attention Deficit Hyperactivity Disorder.

e-mail: emjcampos00@gmail.com

Approaches for Extending Multiple Imputation to Handle Scalar and Functional Data

Adam Ciarleglio*, The George Washington University

Missing data are a common problem in biomedical research. Valid approaches for addressing this problem have been proposed and are regularly implemented in applications where the data are exclusively scalar-valued. However, with advances in technology and data storage, biomedical studies are beginning to collect both scalar and functional data, both of which may be subject to missingness. We propose extensions of multiple imputation with predictive mean matching and imputation by local residual draws as two approaches for handling missing scalar and functional data. The two methods are compared via a simulation study and applied to data from a study of subjects with major depressive disorder for which both clinical (scalar) and imaging (functional) data are available.

e-mail: aciarleglio@gwu.edu

ABSTRACTS & POSTER PRESENTATIONS

Statistical Analysis of Heart Rate Variability from Electrocardiogram Data

Andrada E. Ivanescu*, Montclair State University
Naresh Punjabi, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University

There are patterns of heart rate variability that emerge during sleep. Statistical analysis of heart rate variability from electrocardiogram data during sleep includes the study of sleep stages and the influence on heart rate variability. We discuss several algorithmic steps to identify patterns of heart rate variability during sleep.

e-mail: ivanescua@montclair.edu

Interpretable Principal Components Analysis for Multilevel Multivariate Functional Data, with Application to EEG Experiments

Jun Zhang*, University of Pittsburgh
Greg J. Siegle, University of Pittsburgh
Wendy D' Andrea, New School for Social Research
Robert T. Krafty, University of Pittsburgh

Many studies collect functional data from multiple subjects that have both multilevel and multivariate structures. An example of such data comes from neuroscience experiments where participants' brain activity is recorded using modalities such as EEG and summarized as power within multiple time-varying frequency bands within multiple electrodes, or brain regions. This article introduces a novel approach to conducting interpretable principal components analysis on multilevel multivariate functional data that decomposes total variation into subject-level and replicate-within-subject-level (i.e. electrode-level) variation and provides interpretable components that can be both sparse among variates (e.g. frequency bands) and have localized support over time within each frequency band. The sparsity and localization of components is achieved by solving an innovative rank-one based convex optimization problem with block Frobenius and matrix L1-norm based penalties. The method is used to analyze data from a study to better understand reactions to emotional information in individuals with the symptom of dissociation, revealing new neurophysiological insights into blunted affect.

e-mail: juz30@pitt.edu

114. HIV, INFECTIOUS DISEASE AND MORE

A Hybrid Compartment/Agent-Based Model for Infectious Disease Modeling

Shannon Gallagher*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
William Eddy, Carnegie Mellon University

In infectious disease modeling, compartment models (CM) and agent-based models (AM) are two classes of models used to infer information about a disease and to explore hypothetical scenarios about what could have happened or could happen in the future. CMs describe the progression through a disease for groups of individuals whereas AMs describe the progression through a disease for individuals (or agents) by way of interaction among the agents and their environment. We present a method to create a hybrid CM-AM where the model is selected and parameters are estimated in the CM-paradigm and then incorporated into the AM-paradigm. Our method allows for estimates and models to be verified using existing CM diagnostics. As a consequence, we can be more confident in the results from our AM when exploring hypothetical scenarios. We demonstrate our method with an application to a measles outbreak and show that isolating infectious children could have produced a significantly weaker outbreak.

e-mail: skgallagher19@gmail.com

Analysis of Two-Phase Studies using Generalized Method of Moments

Prosenjit Kundu*, Johns Hopkins Bloomberg School of Public Health
Nilanjan Chatterjee, Johns Hopkins Bloomberg School of Public Health and Johns Hopkins University School of Medicine

Two-phase design can reduce the cost of epidemiological studies by limiting the ascertainment of expensive variables to an efficiently selected subset(phase-II) of a larger(phase-I) study. Efficient analysis of the resulting data combining disparate information from phase-I and phase-II, however, can be complex. Most of the existing methods including semi-parametric maximum-likelihood, require phase-I data to be summarized into a fixed number of strata. We propose a method where phase-I data is summarized by parameters in a reduced logistic regression model of a binary outcome on available covariates. We setup estimating equations for parameters in the desired extended logistic regression model using information on the reduced model parameters and complete phase-II data after adjusting for non-random sampling at phase-II. We generalized method of moments to solve the overly identified equations and derive asymptotic theory for the proposed estimator. Both, simulation studies and analysis of the US National Wilms Tumor data, show that the use of reduced parametric models, as opposed to summarizing data into strata, can lead to more efficient utilization of phase-I data.

e-mail: pkundu3@jhu.edu

ABSTRACTS & POSTER PRESENTATIONS

Bias and Efficiency in Group Testing Estimation for Infectious Disease Surveillance

Katherine M. Bindbeutel*, Radford University
Md S. Warasi, Radford University

Group testing has long been used as a cost-effective procedure in biomedical applications for the screening and surveillance of infectious diseases. In such settings, a set of individual samples such as blood and urine are pooled together and tested simultaneously to yield a positive or negative result. Next, individuals of the negative pools are diagnosed as negative for the disease, and individuals of the positive pools are tested one by one to identify the positive case(s). Ideally, a pool tests negative when all individuals in the pool are truly negative, and a pool tests positive when at least one individual in the pool is truly positive. However, due to the errors in diagnostic assays, testing responses can be misclassified. When the test outcomes are used in group testing models without an adjustment for testing errors, estimates of the disease prevalence can be severely biased. In this project, we will study the properties of group testing estimates, quantify estimates' bias and efficiency, and develop optimal pooling strategies. This will provide methods to improve the accuracy and efficiency of conducting disease surveillance in public health applications.

e-mail: kbindbeutel@radford.edu

Mediation Effect Sizes for Latent Outcome Models using Explained Variance Decomposition

Yue Jiang*, University of North Carolina, Chapel Hill
Shanshan Zhao, National Institute of Environmental Health Sciences, National Institutes of Health
Jason Peter Fine, University of North Carolina, Chapel Hill

Mediation analyses explore underlying processes by which exposures affect outcomes. Relatively little research has focused on developing mediation effect size measures, especially for categorical and censored time-to-event outcomes. We propose a variance-decomposition approach that investigates variance explained by the direct and indirect pathways on an underlying latent outcome, and define an effect measure based on the relative proportion of the explained variance in the latent outcome attributable to the indirect effect. To deal with correlated direct and indirect effects in the multiple exposure setting, we further propose constrained estimates based on simple projections of unconstrained estimators that demonstrate computational advantages compared to constrained maximum likelihood estimates. Under regularity conditions, proposed estimators are consistent and asymptotically normal, and under further conditions demonstrate robustness against misspecification of the assumed latent variable error distribution. Large sample properties are supported through numerical simulation. Analyses of two real-world datasets demonstrate the practicality of our proposed methods.

e-mail: yuejiang@live.unc.edu

Toward Evaluation of Disseminated Effects of Non-Randomized HIV Prevention Interventions Among Observed Networks of People who Inject Drugs

Ashley Buchanan*, University of Rhode Island
Natallia Katenka, University of Rhode Island
TingFang Lee, University of Rhode Island
M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington
Samuel Friedman, New York University
Georgios Nikolopoulos, University of Cyprus

People who inject drugs are embedded in social networks or communities and exert biological and social influence on the members of their social networks. The direct effect of an intervention is the effect on the index participants (i.e., participants who received the intervention) and the disseminated effect is the effect on the participants who shared a network with the index participant. We analyzed a network of people who inject drugs from the Transmission Reduction Intervention Project (TRIP) from 2013 to 2015 in Athens, Greece, where links were defined by shared drug use and social behaviors. In our setting, the study design is an observed network with a nonrandomized intervention or exposure, where information is available on each participant and their connections with other participants. We assumed that smaller groupings or neighborhoods for each individual can be identified in the data. We used a group-level inverse probability weighted approach to quantify the direct and disseminated effects of nonrandomized interventions on subsequent HIV-related health outcomes. We employ several approaches for a bootstrap procedure to quantify uncertainty for the estimators.

e-mail: buchanan@uri.edu

Joint Model of Adherence to Dapivirine-containing Vaginal Ring and HIV-1 Risk

Qi Dong*, University of Washington
Elizabeth R. Brown, Fred Hutchinson Cancer Research Center
Jingyang Zhang, Fred Hutchinson Cancer Research Center

In clinical trials of dapivirine vaginal rings (DVR) for HIV-1 prevention in women, participants' self-reported adherence is often unreliable. Therefore, the classification of adherence relies on two objective measures: the amount of dapivirine remaining in returned rings and the dapivirine levels detected in plasma samples. However, assessing adherence based on these two measures is challenging because the dapivirine assay in rings suffers high measurement variability, and the plasma dapivirine levels cannot accurately identify the non-adherence in participants who remove the DVR after a study visit and re-insert it prior to the next visit. In this paper, we propose a Bayesian hidden Markov model (HMM) that jointly models the HIV infection status and the adherence measures. Specifically, our method uses both the dapivirine measurements in returned rings and in plasma samples to inform the latent adherence states, which serve as a time-varying covariate in the HIV infection model. We apply our model to analyze the data collected during the ASPIRE study conducted by the Microbicide Trials Network between 2012 and 2015.

e-mail: qd8@uw.edu

ABSTRACTS & POSTER PRESENTATIONS

The Mechanistic Analysis of Founder Virus Data in Challenge Models

Ana Maria Ortega-Villa*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Dean A. Follmann, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Repeated low-dose (RLD) challenge studies provide valuable information when evaluating candidate HIV vaccines. Current technology allows use of the number of infecting virions or founder viruses as a readout of infection. This work provides methods to characterize candidate vaccine's protective effect using the number of founder viruses, by determining the vaccine's action model (VAM). The VAMs we consider are a null model (no protection), a Leaky model in which the probability of infection is reduced by some factor in vaccinated subjects, the All-or-none model in which the vaccine either offers complete protection or no protection in vaccinated subjects, and a Combination model with both Leaky and All-or-none mechanisms. We consider two competing models involving maximum likelihood methods for a discrete survival model. These models assume that the founder virus population follows a Poisson distribution with either a fixed (Poisson model), or random (Negative Binomial Model) mean parameter. We illustrate the performance of these methodologies with a data example of SIV on non-human primates and a simulation study.

e-mail: ana.ortega-villa@nih.gov

115. CLINICAL TRIAL DESIGN AND ANALYSIS

Bayesian Design of Clinical Trials for Joint Models of Longitudinal and Time-to-Event Data

Jiawei Xu*, University of North Carolina, Chapel Hill
Matthew A. Psioda, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Joint models of longitudinal and time-to-event data are increasingly used for the analysis of clinical trials data. However, few methods have been proposed for designing clinical trials using these models. In this paper, we develop a Bayesian clinical trial design methodology focused on evaluating the treatment's effect on the time-to-event endpoint using a flexible trajectory joint model. By incorporating the longitudinal outcome trajectory into the hazard model for the time-to-event endpoint, the joint modeling framework allows for non-proportional hazards. Inference for the time-to-event endpoint is based on an average of a time-varying hazard ratio which can be decomposed according to the treatment's direct effect, and its indirect effect mediated through the longitudinal outcome. We propose an approach for sample size determination such that the design has a high power and a well-controlled type I error rate with both defined from a Bayesian perspective. We demonstrate the methodology by designing a breast cancer clinical trial with a primary time-to-event endpoint and where predictive longitudinal outcome measures are also collected periodically.

e-mail: jiawei@live.unc.edu

Statistical Support for Designing Non-Inferiority Trials: An Application to Rheumatoid Arthritis

Rebecca Rothwell*, U.S. Food and Drug Administration
Gregory Levin, U.S. Food and Drug Administration

Exposing subjects to known ineffective treatment for prolonged periods of time in a setting with proven effective therapies raises ethical concerns. Alternative options, such as relying on short-term placebo-controlled periods, can limit the interpretability of long-term results. Conducting a non-inferiority (NI) study can circumvent these issues by indirectly showing a treatment effect without relying on a placebo arm. Active-controlled trials can also provide long-term, reliable, controlled safety data and relevant information for treatment decisions in clinical practice. Successfully designing informative NI trials requires statistical involvement. In this project, we will explore one such case in rheumatoid arthritis. After considering the available historical data and using techniques such as Bayesian hierarchical models to estimate the effects of drugs in a particular class, we proposed margins and evaluated the feasibility in this therapeutic setting. We will discuss the methods, recommendations, and the important impact of statistical participation in collaboration with an FDA multidisciplinary team in this setting.

e-mail: Rebecca.Rothwell@fda.hhs.gov

Determining Mental Health Condition Patterns in Veterans with a Lifetime PTSD Diagnosis

Ilaria Domenicano*, Department of Veterans Affairs Cooperative Studies Program and Yale School of Public Health
Lori L. Davis, Tuscaloosa Veterans Affairs Medical Center and University of Alabama School of Medicine
Lisa Mueller, Edith Nourse Rogers Memorial Veterans Hospital
Tassos Constantino Kyriakides, Department of Veterans Affairs Cooperative Studies Program and Yale School of Public Health

Post-traumatic stress disorder (PTSD) is a mental disorder with symptoms that affect social and work situations, employment, and income. A randomized, controlled, multisite clinical trial was carried out by the VA Cooperative Study Program to compare two different interventions aimed at helping Veterans obtain and maintain competitive employment: the Individual Placement and Support (IPS) intervention, and the Transitional Work (TW) program. The IPS intervention included individual job counseling while the TW program provided temporary employment within VA facilities. We performed a latent profile analysis, searching for classes of patients with distinct mental health condition patterns. Classes were determined by 4 scores derived from Veterans' baseline assessments which measured: severity of PTSD Symptoms, quality of life, level of disability, and self-esteem. We studied the relationships between the classes and the probability of achieving permanent employment. We also compared the cumulative earnings of Veterans across classes. The identification of aspects that increase the chances of gaining a permanent job could guide more individualized interventions.

e-mail: ilaria.domenicano@yale.edu

ABSTRACTS & POSTER PRESENTATIONS

Estimation of Ascertainment Bias and its Effect on Power in Clinical Trials with Time-to-Event Outcomes

Erich J. Greene*, Yale Center for Analytical Sciences
Peter Peduzzi, Yale Center for Analytical Sciences
James Dziura, Yale Center for Analytical Sciences
Can Meng, Yale Center for Analytical Sciences
Denise Esserman, Yale Center for Analytical Sciences

While the gold standard for clinical trials is to blind all parties — participants, researchers, and evaluators — to treatment assignment, this is not always a possibility. When some or all of the above individuals know the treatment assignment, this leaves the study open to the introduction of post-randomization biases. In the Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE) trial, we were presented with the potential for the clinicians administering the treatment, as well as the individuals enrolled in the study, to introduce ascertainment bias. We present ways in which you can estimate the ascertainment bias for a time-to-event outcome and discuss its impact on the overall power of a trial versus changing of the outcome definition to a more stringent, unbiased definition. We found that for the majority of situations, it is better to revise the definition to a more stringent definition, even though fewer events would be observed.

e-mail: erich.greene@yale.edu

Design and Analysis Considerations for Utilizing a Tailoring Function in a snSMART with Continuous Outcomes

Holly E. Hartman*, University of Michigan
Roy N. Tamura, University of South Florida
Matthew J. Schipper, University of Michigan
Kelley Kidwell, University of Michigan

Small sample, sequential, multiple assignment, randomized trials (snSMARTs) are multistage trial designs to identify the best overall treatment. In snSMARTs, binary response/nonresponse outcomes are measured at intermediate and final timepoints. If the patient is responding at the the end of the first stage timepoint, they continue on the same treatment. Otherwise, they are re-randomized to one of the remaining treatments. Here we propose a modification to the snSMART design to allow for continuous outcomes. The probability of staying on the same treatment is a function of the first stage outcome thus eliminating the need for a categorical tailoring variable defining response/nonresponse. This re-randomization scheme allows for trials to continue without requiring a dichotomous variable. Additionally, we show that this method reduces bias and increases efficiency relative to a dichotomization cut off re-randomization scheme. We also show that patient outcomes are similar using this design relative to a traditional snSMART design.

e-mail: holhart@umich.edu

Two-Part Proportional Mixed Effects Model for Clinical Trials in Alzheimer's Disease

Guoqiao Wang*, Washington University in St. Louis
Yan Li, Washington University in St. Louis
Chengjie Xiong, Washington University in St. Louis
Lei Liu, Washington University in St. Louis
Andrew Aschenbrenner, Washington University in St. Louis
Jason Hassenstab, Washington University in St. Louis
Eric McDade, Washington University in St. Louis
Randall Bateman, Washington University in St. Louis

In clinical trials for Alzheimer's disease, participants are often a mixture of difference disease stages, e.g. symptomatic participants and asymptomatic ones. Participants from different disease stages often have different rates of change and different variability in the primary outcome. However, the primary analysis model typically lumps all these participants together and analyze the data using a mixed effects model, leading to some loss of power. We proposed the two-part proportional mixed effects model to simultaneously model the sub-cohorts (e.g. symptomatic vs asymptomatic) in one model with a shared proportional treatment effect. This two-part proportional model can be linear mixed effects models (time as continuous) or mixed effects models for repeated measures (time as categorical). Simulations showed that the two-part model can lead to over 30% power increase compared to lumping the two sub-cohorts in one model. The two-part model can be easily extended to the multiple-part latent class model where the sub-cohorts are classified by the model itself. The two-part proportional model has the same assumptions as the typical primary analysis models.

e-mail: guoqiao@wustl.edu

ABSTRACTS & POSTER PRESENTATIONS

116. MULTIVARIATE AND HIGH-DIMENSIONAL DATA ANALYSIS

On Genetic Correlation Estimation with Summary Statistics from Genome-Wide Association Studies

Bingxin Zhao*, University of North Carolina, Chapel Hill
Hongtu Zhu, University of North Carolina, Chapel Hill

Cross-trait polygenic risk score (PRS) method has gained popularity for assessing the genetic correlation of complex traits using summary statistics from biobank-scale genome-wide association studies (GWAS). However, empirical evidence has shown a common bias phenomenon that highly significant cross-trait PRS can only account for a very small amount of genetic variance (R-squared can be <1%) in independent testing GWAS. The aim of this paper is to investigate and address the bias phenomenon of cross-trait PRS. We theoretically show that the estimated genetic correlation can be asymptotically biased towards zero. We propose a consistent cross-trait PRS estimator to correct such asymptotic bias. We also study the variance of cross-trait PRS and explain why the estimator can still be significant even it is heavily biased towards zero. Our results may help demystify and tackle the puzzling "missing genetic overlap'" phenomenon of cross-trait PRS for dissecting the genetic similarity of closely related heritable traits. We illustrate our results by assessing the genetic correlation between human brain volume and reaction time.

e-mail: bingxin@live.unc.edu

Multivariate Association Analysis with Correlated Traits in Related Individuals

Souvik Seal*, University of Minnesota

Genome-wide association studies have reported variants that affect multiple correlated traits, demonstrating evidence of pleiotropy or shared genetic basis. Joint analysis of such traits can be challenging in the current mixed-model framework of GWA studies. It becomes even harder in a data-set with families or distantly related individuals due to an additional mode of dependency, coined as the "genetic similarity". The traditional modeling technique involves a lot of covariance parameters making it incredibly tough to implement when either the number of traits or the number of individuals is high. We propose a rapid test based on Linear Mixed Modeling and Seemingly Unrelated Regression. The test shares a strong theoretical connection with the traditional approach but is much faster and hence, suitable on a genome-wide scale. Through simulation studies, we have shown that the proposed approach outperforms the traditional and several other existing approaches in most of the cases. We have also analyzed the full sibling pairs from the UK Biobank data with four anthropometric traits and have detected some interesting associations.

e-mail: sealx017@umn.edu

Grafted and Vanishing Random Subspaces

Matthew Corsetti*, University of Rochester
Tanzy Love, University of Rochester

The Random Subspace Method (RSM) constructs an ensemble of learners, such as regression trees, using randomly chosen feature subsets. This reduces correlation among learners resulting in a stronger ensemble. RSM has a notable drawback. A randomly chosen feature subspace may lack the information needed to produce a reasonable learner. In this setting, the learner can be damaging to the ensemble. We present Grafted (GRS) and Vanishing Random Subspaces (VRS), two novel procedures for constructing ensembles of trees that do not suffer from the aforementioned drawback. As trees are grown, important variables are either guaranteed inclusion into (grafting) or exclusion from (vanishing) the randomly chosen feature subsets for a number of successive trees. GRS recycles a promising feature from one subset across several subsequent subsets while VRS prevents the important feature from being included in the successive subsets thus forcing new trees to explore different feature spaces. Results show improved predictive performance of GRS and VRS over RSM, Gradient Boosting and Random Forests for several datasets from the UCI Machine Learning Repository as well as simulated datasets.

e-mail: matthew_corsetti@urmc.rochester.edu

Modeling Repeated Multivariate Data to Estimate Individuals' Trajectories with Application to Scleroderma

Ji Soo Kim*, Johns Hopkins University
Ami Shah, Johns Hopkins University School of Medicine
Laura Hummers, Johns Hopkins University School of Medicine
Scott L. Zeger, Johns Hopkins University

For many chronic diseases, patients' health state is reflected in longitudinal clinical measures and/or occurrences of clinical events over their course of disease. The estimation of a patient's true disease trajectory is necessary as it generates real-time, actionable predictions to guide clinical care. Optimally defining latent health trajectory that characterizes overall health state fully utilizing information in multiple longitudinal markers requires fitting a joint model of all markers as opposed to marker-specific models. We develop a Bayesian model to estimate population, clinical subgroups, and individual trajectories and their dependence on baseline and time-varying covariates and examine their relative merits compared to marker-specific models. We demonstrate this approach on data from a study of autoimmune disease scleroderma, estimating latent health trajectory using two markers from heart and lung respectively and a marker for skin. The fitted model describes the effect of autoantibodies and other baseline clinical variables on trajectories and correlations among markers from different organs over time.

e-mail: jkim478@jhu.edu

ABSTRACTS & POSTER PRESENTATIONS

Nonignorable Item Nonresponse in Multivariate Outcomes

Sijing Li*, University of Wisconsin, Madison
Jun Shao, University of Wisconsin, Madison

To estimate unknown population parameters based on multivariate data having nonignorable item nonresponse, we propose an innovative data grouping approach according to the number of observed components in the multivariate study variable y when the joint distribution of y and covariate x is nonparametric and the nonresponse probability conditional on y and x has a parametric form. The likelihood based on observed data may not be identifiable even when the joint distribution of y and x is parametric. We solve this problem by utilizing a nonresponse instrument z, an auxiliary variable related to y but not related to the nonresponse probability conditional on y and x. Under some conditions we apply a modified generalized method of moments (GMM) to obtain estimators of the parameters in the nonresponse probability and the nonparametric joint distribution of y and x. Consistency and asymptotic normality of the proposed estimators are established. Simulation and real data results are presented.

e-mail: sli394@wisc.edu

Multivariate Association Analysis with Somatic Mutation Data

Chad He*, Fred Hutchinson Cancer Research Center
Yang Liu, Wright State University
Ulrike Peters, Fred Hutchinson Cancer Research Center
Li Hsu, Fred Hutchinson Cancer Research Center

Somatic mutations are the driving forces for tumor development, and recent advances in cancer genome sequencing have made it feasible to evaluate the association between somatic mutations and cancer-related traits. However, it is challenging to conduct statistical analysis for somatic mutations because of their low frequencies. Furthermore, cancer is a complex disease and it is often accompanied by multiple traits that reflect various aspects of cancer; how to combine the information of these traits to identify important somatic mutations poses additional challenges. We introduce a statistical approach, named as SOMAT, for detecting somatic mutations associated with multiple cancer-related traits. Our approach provides a flexible framework for analyzing multiple traits, and a data-adaptive procedure for effectively combining test statistics is proposed. Simulations show that the proposed approach works well in the considered situations. We also apply our approach to an exome-sequencing tumor dataset for illustration.

e-mail: qhe@fhcrc.org

117. ASYMMETRICAL STATISTICAL LEARNING FOR BINARY CLASSIFICATION

Introduction to Neyman-Pearson Classification

Jingyi Jessica Li*, University of California, Los Angeles
Xin Tong, University of Southern California
Yang Feng, Columbia University

In many binary classification applications such as disease diagnosis, practitioners commonly face the need to control type I error (i.e., the conditional probability of misclassifying a class 0 observation as class 1) so that it remains below a desired threshold. To address this need, the Neyman-Pearson (NP) classification paradigm is a natural choice; it minimizes type II error (i.e., the conditional probability of misclassifying a class 1 observation as class 0) while enforcing an upper bound, alpha, on the type I error. Although the NP paradigm has a century-long history in hypothesis testing, it has not been well recognized and implemented in classification schemes. Common practices that directly control the empirical type I error to no more than alpha do not satisfy the type I error control objective because the resulting classifiers are still likely to have type I errors much larger than alpha. In this talk, I will introduce an unbrella algorithm, which implements the NP paradigm for scoring-type classification methods, and the NP-ROC bands as a graphical tool for evaluating and comparing classification methods under the NP paradigm.

e-mail: jli@stat.ucla.edu

A Unified View of Asymmetric Binary Classification

Wei Vivian Li*, Rutgers, The State University of New Jersey
Xin Tong, University of Southern California
Jingyi Jessica Li, University of California, Los Angeles

Cost-sensitive (CS) classification methods account for asymmetric misclassification costs and are widely applied in real-world problems such as medical diagnosis, transaction monitoring, and fraud detection. The current approaches to binary CS learning usually assign different weights to the two classes in non-unified ways, and the three main ways include rebalancing the sample before training, changing the objective function for training, and adjusting the estimated posterior class probabilities after training. Moreover, existing CS learning work has only focused on improving empirical classification errors or costs incurred by the assigned weights while overlooking the changes in population classification errors. We propose an umbrella algorithm to estimate the population type I error control achieved by multiple binary CS learning approaches. Our algorithm for the first time establishes a connection between CS learning and the Neyman-Pearson classification paradigm, which minimizes the population type II error while enforcing an upper bound on the population type I error.

e-mail: vivian.li@rutgers.edu

ABSTRACTS & POSTER PRESENTATIONS

Neyman-Pearson Classification: Parametrics and Sample Size Requirement

Yang Feng*, New York University

The Neyman-Pearson (NP) paradigm in binary classification seeks classifiers that achieve a minimal type II error while enforcing the prioritized type I error controlled under some user-specified level α. This paradigm serves naturally in applications such as severe disease diagnosis and spam detection, where people have clear priorities among the two error types. Recently, Tong, Feng and Li (2018) proposed a nonparametric order statistics based umbrella algorithm that adapts all scoring-type classification methods (e.g., logistic regression, support vector machines, random forest) to respect the given type I error upper bound α with high probability, without specific distributional assumptions on the features and response. Universal the umbrella algorithm is, it demands an explicit minimum sample size requirement on class 0, which is usually the more scarce class. In this work, we employ the parametric linear discriminant analysis (LDA) model and propose a new parametric thresholding algorithm, which does not need the minimum sample size requirements on class 0 observations and thus is applicable to small sample applications such as rare disease diagnosis.

e-mail: yang.feng@nyu.edu

Intentional Control of Type I Error over Unconscious Data Distortion: A Neyman-Pearson Approach to Text Classification

Xin Tong*, University of Southern California
Lucy Xia, Hong Kong University of Science and Technology
Richard Zhao, The Pennsylvania State University
Yanhui Wu, University of Southern California

Digital texts have become an increasingly important source of data for social studies. However, textual data from open platforms are vulnerable to manipulation, often leading to bias in subsequent empirical analysis. This paper investigates the problem of data distortion in text classification when controlling type I error (a relevant textual message is classified as irrelevant) is the priority. The default classical classification paradigm that minimizes the overall classification error can yield an undesirably large type I error, and data distortion exacerbates this situation. As a solution, we propose the Neyman-Pearson (NP) classification paradigm which minimizes type II error under a user-specified type I error constraint. Theoretically, we show that while the classical oracle cannot be recovered under unknown data distortion even if one has the entire post-distortion population, the NP oracle is unaffected by data distortion and can be recovered under the same condition. Empirically, we illustrate the advantage of NP classification methods in a case study that classifies posts about strikes published on a leading Chinese blogging platform.

e-mail: xint@marshall.usc.edu

118. RECENT ADVANCES AND OPPORTUNITIES IN LARGE SCALE & MULTI-OMIC SINGLE-CELL DATA ANALYSIS

Statistical Analysis of Coupled Single-Cell RNA-seq and Immune Profiling Data

Hongkai Ji*, Johns Hopkins Bloomberg School of Public Health
Zhicheng Ji, Johns Hopkins Bloomberg School of Public Health

We present an analytical framework for analyzing coupled single-cell transcriptome (scRNA-seq) and T cell receptor sequencing (scTCR-seq) data. The framework provides key functions for preprocessing, aligning cells from different samples, detecting differential gene expression across biological conditions, analyzing sequence features in T cell repertoire, and linking sequence features to gene expression signatures. We demonstrate this framework by analyzing single-cell data both from public databases and from a neoadjuvant immunotherapy clinical trial for non-small cell lung cancer.

e-mail: hji@jhsph.edu

Assessing Consistency of Single Cell Unsupervised Multi-Omics Methods

Michael I. Love*, University of North Carolina, Chapel Hill

Recently, we have examined the consistency and overfitting of unsupervised multi-omics, or multi-modal, methods on bulk assays, e.g. RNA-seq, methylation, proteomics, where covariation across samples is considered. We proposed a cross-validation framework to determine if the projections identified by unsupervised methods, which should maximize shared variation across data modalities, in fact generalize to out-of-fold samples. Here, we discuss the application of the cross-validation framework to multi-omics single cell datasets, using existing methods and newly proposed methods specifically designed for single cell assays. As with samples in bulk assays, here out-of-fold cells can be used to determine the extent of consistency and overfitting of methods, and to compare across methods without resorting to creation of complex simulation datasets.

e-mail: michaelisaiahlove@gmail.com

Statistical Methods for Identifying and Characterizing Cell Populations using High-Dimensional Single-Cell Data

Raphael Gottardo*, Fred Hutchinson Cancer Research Center

New single-cell technologies including single-cell RNA-seq and CyTOF enable the unprecedented interrogation of single-cell phenotypes (and functions) under various biological conditions. A common statistical problem is the inference of such cell phenotypes from single-cell data. During this talk, I will present new statistical methodology to cluster high-dimensional single-cell data to define and characterize cell populations in an unsupervised way. For each cluster, I will show how our proposed approach can be used to define biologically meaningful cell phenotypes based on the markers (genes or proteins) identified as significant drivers of that cluster. Finally, I will illustrate this novel approach using several datasets that we have recently generated to characterize immune responses in the context of vaccination and immunotherapy.

e-mail: rgottard@fredhutch.org

ABSTRACTS & POSTER PRESENTATIONS

119. NOVEL STATISTICAL METHODS FOR COMPLEX INTERVAL-CENSORED SURVIVAL DATA

Semiparametric Regression Analysis of Multiple Censored Events in Family Studies

Donglin Zeng*, University of North Carolina, Chapel Hill
Fei Gao, Fred Hutchinson Cancer Research Center
Yuanjia Wang, Columbia University

Family history of disease is a major risk factor for health outcomes and assessing disease familial aggregation and genetic risk is essential for implementing precision medicine in the clinical setting. Cost-effective designs based on probands and their families have been proposed to collect family history data. However, the exact time of disease onset may not be available and is usually subject to interval censoring. Furthermore, in many studies, history of multiple events in relatives are collected. We propose a semiparametric regression model for the family history data that assumes a family-specific random effect and individual random effects to account for the dependence due to shared environmental exposures and unobserved genetic relatedness, respectively. To incorporate multiple events, we jointly model the onset of the primary disease of interest and a secondary disease outcome. We propose nonparametric maximum likelihood estimation for inference and develop a stable EM algorithm for computation. We examine the performance of the proposed methods through simulation studies and application to a large cohort study of Alzheimer's disease.

e-mail: dzeng@email.unc.edu

Modeling Interval Censored Time to Event Outcomes with Inflation of Zeros, with Application to Pediatric HIV Studies

Raji Balasubramanian*, University of Massachusetts, Amherst

We evaluate the properties of parametric mixture models for settings in which the study cohort characterized by interval censored observations of a time to event outcome in the presence of inflation of events at time zero. Motivated by applications in pediatric HIV studies where it is of interest to model the time to first positive diagnostic test among HIV infected infants, we evaluate parametric mixture models that can account for inflation of zeros in the presence of interval censored outcomes. We propose models that relax the proportional hazards assumption and evaluate the magnitude of bias when the mixture component is ignored in analysis. The proposed models will be evaluated in simulation studies and applied to data generated from pediatric HIV clinical trials.

e-mail: rbalasub@schoolph.umass.edu

Case-Cohort Studies with Multiple Interval-Censored Disease Outcomes

Qingning Zhou*, University of North Carolina, Charlotte
Jianwen Cai, University of North Carolina, Chapel Hill
Haibo Zhou, University of North Carolina, Chapel Hill

Interval-censored failure time data commonly arise in epidemiological and biomedical studies where the occurrence of an event or a disease is determined via periodic examinations. Subject to interval-censoring, available information on the failure time can be quite limited. Cost-effective sampling designs are desirable to enhance the study power, especially when the disease rate is low and the covariates are expensive to obtain. In this work, we formulate the case-cohort design with multiple interval-censored disease outcomes and generalize it to nonrare diseases where only a portion of diseased subjects are sampled. We develop a marginal sieve weighted likelihood approach, which assumes that the failure times marginally follow the proportional hazards model. We consider two types of weights to account for the sampling bias, and adopt a sieve method with Bernstein polynomials to handle the unknown baseline functions. We employ a weighted bootstrap procedure to obtain a variance estimate that is robust to the dependence structure between failure times. The proposed method is examined via simulations and illustrated with the ARIC data.

e-mail: qzhou8@uncc.edu

Adjusting for Covariate Measurement Error in Survival Analysis under Competing Risks

Sharon Xiangwen Xie*, University of Pennsylvania
Carrie Caswell, University of Pennsylvania

Time-to-event data in the presence of competing risks has been well studied in recent years. One popular approach to this problem is to model the subdistribution of competing risks with assumptions of proportional hazards and covariates measured without error. The estimator resulting from this model does not perform as expected when the covariates are measured with error, which is often the case in biomarker research. We propose a novel method which combines the intuition of the subdistribution model with risk set regression calibration, which corrects for measurement error in Cox regression by recalibrating at each failure time. We perform simulations to assess under which conditions the subdistribution hazard ratio estimator incurs bias in regression coefficients, and demonstrate that our new estimator reduces this bias. We show that the estimator is asymptotically normally distributed and provide a consistent variance estimator. This method is applied to Alzheimer's Disease Neuroimaging Initiative data, which examine the association between measurement-error prone cerebrospinal fluid biomarkers and risk of conversion to Alzheimer's disease.

e-mail: sxie@pennmedicine.upenn.edu

ABSTRACTS & POSTER PRESENTATIONS

120. MODERN GRAPHICAL MODELING OF COMPLEX BIOMEDICAL SYSTEMS

A Tripartite Latent Graph for Phenotype Discovery in EHR Data

Peter Mueller*, University of Texas, Austin
Yang Ni, Texas A&M University
Yuan Ji, The University of Chicago

We set up a tri-partite graph to infer underlying diseases in EHR data. A latent central layer in the graph represents unobserved diseases that explain symptoms that are recorded on patients, with patients and symptoms defininig the other two layers of the model. THe graphical model is mapped to a family of probability models that can be characterized as feature allocation with features representing latent diseases, and one feature specific parameter that links a set of symptoms to each disease. The model can alternatively be described as a sparse factor model, or categorical matrix factorization. The representation as a graphical model is mainly useful for a graphical summary of the inferred structure. Using a Bayesian approach, available prior information on known diseases greatly improves identifiability of latent diseases. We validate the proposed approach by simulation studies including mis-specified models and comparison with sparse latent factor models. In an application to Chinese electronic health records (EHR) data, we find results that agree with related clinical and medical knowledge.

e-mail: pmueller@math.utexas.edu

The Reduced PC-Algorithm: Improved Causal Structure Learning in Large Random Networks

Ali Shojaie*, University of Washington

We consider the task of estimating a high-dimensional directed acyclic graph, given observations from a linear structural equation model with arbitrary noise distribution. By exploiting properties of common random graphs, we develop a new algorithm that requires conditioning only on small sets of variables. The proposed algorithm offers significant gains in both computational and sample complexities. In particular, it results in more efficient and accurate estimation in large networks containing hub nodes, which are common in biological systems. We prove the consistency of the proposed algorithm, and show that it also requires a less stringent faithfulness assumption than the PC-Algorithm. Simulations in low and high-dimensional settings are used to illustrate these findings. An application to gene expression data suggests that the proposed algorithm can identify a greater number of clinically relevant genes than current methods. Time permitting, we will discuss the extension of the algorithm for causal structure learning in the presence of confounders and selection variables.

e-mail: ashojaie@uw.edu

Latent Network Estimation and variable selection for compositional data via variational em

Nathan Osborne*, Rice University
Christine B. Peterson, University of Texas MD Anderson Cancer Center
Marina Vannucci, Rice University

Network estimation and variable selection have been extensively studied in the statistical literature, but only recently have those two questions been addressed simultaneously. In this paper, we seek to develop a novel method to simultaneously estimate network interactions and associations to relevant covariates for count data, and specifically for compositional data, which have a fixed sum constraint. We use a hierarchical Bayesian model with latent layers and employ spike and slab priors for both edge and covariate selection. For posterior inference, we develop a variational inference scheme with an expectation maximization step, to enable efficient estimation. Through simulation studies, we demonstrate that the proposed model outperforms existing methods in its accuracy of network recovery. We show the practical utility of our model via an application to microbiome data from the Multi-Omic Microbiome Study-Pregnancy Initiative (MOMS-PI) study, to estimate the interaction between microbes in the vagina, as well as the interplay between vaginal cytokines and microbial abundances.

e-mail: nathan.osborne@rice.edu

Personalized Integrated Network Estimation

Veera Baladandayuthapani*, University of Michigan
Min Jin Ha, University of Texas MD Anderson Cancer Center
Yang Ni, Texas A&M University
Francesco C. Stingo, University of Florence, Italy

Personalized (patient-specific) approaches have recently emerged with a precision medicine paradigm that acknowledges the fact that molecular pathway structures and activity might be considerably different within and across patient populations. In the context of cancer, the functional cancer genome and proteome provide rich sources of information to identify patient-specific variations in signaling pathways and activities within and across tumors; however, current analytic methods lack the ability to exploit the diverse and multi-layered architecture of these complex biological networks. We consider the problem of modeling conditional independence structures in heterogenous data using Bayesian graphical regression techniques that allows patient-specific network estimation and inferences. We propose a novel specification of a conditional (in)dependence function of patient-specific covariates—which allows the structure of a directed graph to vary flexibly with the covariates; imposes sparsity in both edge and covariate selection; produces both subject-specific and predictive graphs; and is computationally tractable.

e-mail: veerab@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

121. HIGHLY EFFICIENT DESIGNS AND VALID ANALYSES FOR RESOURCE CONSTRAINED STUDIES

Semiparametric Generalized Linear Models for Analysis of Longitudinal Data with Biased Observation-Level Sampling

Paul Rathouz*, University of Texas, Austin

Rathouz and Gao (2009) proposed a novel class of generalized linear models indexed by a linear predictor and a link function for the mean of (Y|X). In this class, the distribution of (Y|X) is left unspecified and estimated from the data via exponential tilting of a reference distribution, yielding a semiparametric response model that is a member of the natural exponential family. We have since developed generalized case-control sampling designs for univariate data in this class of models. In this talk, we show how these designs extend to longitudinal studies with time-point specific sampling plans, where sampling may depend on earlier observations and/or on an auxiliary variable.

e-mail: paul.rathouz@austin.utexas.edu

Cluster-Based Outcome-Dependent Sampling in Resource-Limited Settings: Inference in Small-Samples

Sara M. Sauer*, Harvard T.H. Chan School of Public Health
Bethany Hedt-Gauthier, Harvard Medical School
Claudia Rivera-Rodriguez, University of Auckland
Sebastien Haneuse, Harvard T.H. Chan School of Public Health

Outcome-dependent sampling is an indispensable tool for carrying out cost-efficient research in resource-limited settings. One such sampling scheme is a cluster-based design where clusters of individuals are selected on the basis of the outcome rate of the individuals. For a given dataset collected via a cluster-based outcome-dependent sampling scheme, estimation for a marginal model can proceed using inverse-probability-weighted generalized estimating equations, where the cluster-specific weights are the inverse probability of the clinic's inclusion in the sample. We provide a detailed treatment of the asymptotic properties of this estimator, together with an explicit expression for the asymptotic variance and a corresponding estimator. Furthermore, motivated by a study we conducted in Rwanda, we provide expressions for small-sample bias corrections to both the point estimates and the standard error estimates. The proposed methods are illustrated through simulation and with data from 18 health centers in Rwanda, collected via a cluster-based outcome-dependent sampling scheme, with the goal of examining risk factors for low birthweight.

e-mail: ssauer@g.harvard.edu

Optimal Designs of Two-Phase Studies

Ran Tao*, Vanderbilt University Medical Center
Donglin Zeng, University of North Carolina, Chapel Hill
Danyu Lin, University of North Carolina, Chapel Hill

The two-phase design is a cost-effective sampling strategy to evaluate the effects of covariates on an outcome when certain covariates are too expensive to be measured on all study subjects. Under such a design, the outcome and inexpensive covariates are measured on all subjects in the first phase and the first-phase information is used to select subjects for measurements of expensive covariates in the second phase. Previous research on two-phase studies has focused largely on the inference procedures rather than the design aspects. We investigate the design efficiency of the two-phase study, as measured by the semiparametric efficiency bound for estimating the regression coefficients of expensive covariates. We consider general two-phase studies, where the outcome variable can be continuous, discrete, or censored, and the second-phase sampling can depend on the first-phase data in any manner. We develop optimal or approximately optimal two-phase designs, which can be substantially more efficient than the existing designs. We demonstrate the improvements of the new designs over the existing ones through extensive simulation studies and two large medical studies.

e-mail: r.tao@vanderbilt.edu

Predictive Case Control Designs for Modification Learning

Patrick James Heagerty*, University of Washington
Katherine Tan, Flatiron Health

Prediction models for clinical outcomes, originally developed using a source dataset, may additionally be applied to a new cohort. Prior to accurate application, the source model needs to be updated through external validation procedures such as modification learning. Model modification involves the dual goals of recalibrating overall predictions as well as revising individual feature effects, and generally requires the collection of an adequate sample of true outcome labels from the new setting. Unfortunately, outcome label collection is frequently an expensive and time-consuming process involving abstraction by human clinical experts. To reduce abstraction burden for such new data collection, we propose a class of designs based on original model scores and their associated outcome predictions. We provide mathematical justification that the general predictive score sampling class results in valid samples for analysis. Then, we focus attention specifically on a stratified sampling procedure that we call predictive case control (PCC) sampling, which allows the dual modification learning goals to be achieved at a smaller sample size compared to simple random sampling.

e-mail: heagerty@uw.edu

ABSTRACTS & POSTER PRESENTATIONS

122. STATISTICAL ANALYSIS OF TRACKING DATA FROM PERSONAL WEARABLE DEVICES

Smartphone-Based Estimation of Sleep

Ian J. Barnett*, University of Pennsylvania
Melissa Martin, University of Pennsylvania

Reliably estimating a person's sleep duration and quality with smartphone sensor data such as screen state and accelerometer can be complicated by sparse and missing data. Current methods of sleep estimation can fail when this data is sufficiently sparse or missing. We propose a sleep estimation model that borrows information from a person's previous nights of follow-up through the use of random effects so as to provide reasonable estimates of sleep even for nights where insufficient data is collected. We compare our approach with competing methods for sleep estimation in a cohort of children with borderline personality disorder where sleep esimates from actigraphy watches are used as the gold standard for the smartphone sensor-based comparisons.

e-mail: ibarnett@pennmedicine.upenn.edu

Quantifying Mortality Risks using Accelerometry Data Collected According to the Complex Survey Weighted Design

Ekaterina Smirnova*, Virginia Commonwealth University
Andrew Leroux, Johns Hopkins University
Lucia Tabacu, Old Dominion University
Ciprian Crainiceanu, Johns Hopkins University

Systematic assessment of population mortality risks and evaluation of behavioral patterns associated with mortality is critical for public health and preventive care. A growing number of population level studies (UK Biobank, National Health Examination Survey) uses accelerometers to associate daily physical activity patterns to health outcomes. Participants in such large population studies are often chosen according to the complex survey weighted design, which needs to be considered at modeling stage to assure accurate inferences. The raw data is typically summarized into a minute-level accelerometry count measure, which leads to functional data collected over 1440 minutes per each subject day. To build predictive models, this data can be either further summarized (e.g. total activity counts, time spent in sedentary activity) or considered as functional data. We discuss the challenges associated with predictive modeling of the survey weighted design physical activity data and proposed solutions. We illustrate these challenges on the example of building the all-cause mortality prediction model using National Health Examination Study (NHANES) data.

e-mail: ekaterina.smirnova@vcuhealth.org

Circadian Rhythm for Physical Activity of Infants Under 1-year Old

Jiawei Bai*, Johns Hopkins University
Sara Benjamin-Neelon, Johns Hopkins University
Vadim Zipunnikov, Johns Hopkins University

Accelerometers have been widely used to provide objectively measured physical activity in various population. Their non-invasive nature and ever improving battery life has enabled them to be deployed in typically hard-to-reach population, including elderly people or infants. Our motivating dataset is from a multi-visit observational study, Nurture, which collected 24-hour accelerometry data for at least 4 consecutive days during each visit on infants under 1 year old. Using a combination of methods, we developed a model to describe the circadian pattern of the infants, which showed a clearer and clearer day-night pattern during their first-year of life. We will also discuss additional data management, visualization and analysis challenges introduced by this new type of data.

e-mail: javybai@gmail.com

123. META-ANALYSIS METHODS

A Three-Groups Bayesian Approach for Identifying Genetic Modifiers from Disparate Data Sources, with Application to Parkinson's Disease

Daisy Philtron*, The Pennsylvania State University
Benjamin Shaby, Colorado State University
Vivian Cheng, The Pennsylvania State University

With the advent of new technologies such as RNA- and whole genome sequencing, genetic research into the cause and progression of neurodegenerative diseases has resulted in many valuable discoveries. Much remains unknown, however, and sharing information for the same genetic targets across multiple data types may offer an improvement in discrimination relative to performing multiple parallel analyses. In this work we propose a three-groups modeling approach that allows for information sharing between multiple data types and technologies using a fully Bayesian framework that classifies each signal as null, deleterious, or beneficial. This framework is flexible enough to incorporate a wide variety of data types. We employ the three-groups framework to jointly model RNA-seq and whole-genome sequencing data. We include the results from simulation studies and from an application of the three-groups framework to Parkinson's disease.

e-mail: dlp245@psu.edu

ABSTRACTS & POSTER PRESENTATIONS

Multi-Trait Analysis of Rare-Variant Association Summary Statistics using MTAR

Lan Luo*, University of Wisconsin, Madison
Judong Shen, Merck & Co., Inc.
Hong Zhang, Merck & Co., Inc.
Aparna Chhibber, Merck & Co., Inc.
Devan V. Mehrotra, Merck & Co., Inc.
Zheng-zheng Tang, Wisconsin Institute for Discovery at University of Wisconsin, Madison

There is a growing interest in integrating association evidence across multiple traits to improve the power of gene discovery and reveal pleiotropy. The majority of multitrait analysis methods focus on individual common variants in genome-wide association studies (GWAS). We introduce multitrait analysis of rare-variant associations (MTAR), a framework for joint analysis of association summary statistics between multiple rare variants and different traits, possibly from overlapping samples. MTAR tests accommodate a wide variety of association patterns across traits and variants and enrich their associations. We apply MTAR to rare-variant summary statistics for three lipid traits in the Global Lipids Genetics Consortium (GLGC, N ≈ 300,000). As compared to the 99 genome-wide significant genes identified in the single-trait-based tests, MTAR increases the number of associated genes to 139. Among the 11 novel lipid-associated genes exclusively discovered by MTAR, seven are replicated in an independent UK Biobank GWAS data. Our study demonstrates that MTAR is substantially more powerful than single-trait-based tests and highlights the value of MTAR for novel gene discovery.

e-mail: lluo24@wisc.edu

Empirical Bayes Approach to Integrate Multiple External Summary-Level Information into Current Study

Tian Gu*, University of Michigan
Jeremy M.G. Taylor, University of Michigan
Bhramar Mukherjee, University of Michigan

We consider the situation in which there are K external studies, each of which developed a prediction model for the same outcome. The parameters of the external models are known, but the external data is not available. The goal is to develop a prediction model that uses all the possible covariates, using data from an internal study and the summary-level information from the external studies. We propose a meta-analysis framework, in which the parameters of interest may differ between the internal and external populations, but are assumed to be drawn from a common distribution. We use an empirical Bayes estimation approach, which first separately incorporates the different summary information from each external study into the internal study, and then takes a weighted average of the resulting estimates to give a final overall estimate. Our approach does not require external models to have the same covariates. We show that the final overall estimate is more efficient than the simple analysis of the internal data. The approach also gives an empirical best linear unbiased estimator for each of the internal and external populations.

e-mail: gutian@umich.edu

Tradeoff between Fixed-Effect and Random-Effects Meta-Analyses

Yipeng Wang*, Florida State University
Lifeng Lin, Florida State University

We proposed a new method, called the tradeoff method, to achieve the compromise between two conventional model settings, the fixed-effect model and the random-effects model, in meta-analysis. For the maximum likelihood estimates of a meta-analysis, we employed a penalty on between-study variance and applied profile likelihood to simplify the bivariate minimization problem to the univariate one. Combineing the penalized-likelihood method with the cross-validation method to construct the loss function, applied to choose the optimal tuning parameter and corresponding outcomes for a meta-analysis. Simulation studies showed 95% confidence interval coverage probability from outcomes of the tradeoff method is much higher than that of two traditional models in general situations. For high proportion of outliers, both bias and mean squared errors of results from the new method outperformed. Three case studies showed the compromise could be achieved by the new method; it might improve the estimation of overall effect size for the third case study through the cross-validation process.

e-mail: yw17b@my.fsu.edu

Bayesian Approach to Assessing Publication Bias with Controlled False Positive Rate in Meta-Analyses of Odds Ratios

Linyu Shi*, Florida State University
Lifeng Lin, Florida State University

Publication bias (PB) is a major threat to research synthesis and has been found in many scientific fields. Egger's regression test is the most widely-used approach, because it is easily implemented and generally has satisfactory statistical powers. It assesses PB by examining the association between the observed effect sizes and their standard errors (SEs). However, its false positive rate may be seriously inflated due to the intrinsic association between the effect size estimates and their SEs even when no PB appears. Although various alternative methods are available to deal with this problem, they are powerful in specific cases and are less intuitive than Egger's regression. This article proposes a new approach to assessing PB in meta-analyses of odds ratios via Bayesian hierarchical models. It controls false positive rates by using true SEs to perform Egger-type regression; those true SEs can be feasibly estimated with the Markov chain Monte Carlo algorithm. We present extensive simulations and 3 case studies to illustrate the superior performance of the proposed method.

e-mail: ls16d@my.fsu.edu

ABSTRACTS & POSTER PRESENTATIONS

A Bayesian Hierarchical CACE Model Accounting for Incomplete Noncompliance Data in Meta-Analysis

Jincheng (Jeni) Zhou*, Amgen
James S. Hodges, University of Minnesota
Haitao Chu, University of Minnesota

Noncompliance to assigned treatments is a common challenge in the analysis and interpretation of a randomized trial. One approach to handle noncompliance is to estimate the complier-average causal effect (CACE) using the principal stratification framework, where CACE measures the impact of an intervention in the subgroup of the population that complies with its assigned treatment. When noncompliance data are reported in each trial, intuitively one can implement a two-step approach (first estimating CACE for each study and then combining them) to estimate CACE in a meta-analysis. However, it is common that some trials do not report noncompliance data. The two-step approach can be less efficient and potentially biased as trials with incomplete noncompliance data are excluded. In this paper, we propose a flexible Bayesian hierarchical CACE framework to simultaneously account for heterogeneous and incomplete noncompliance data in a meta-analysis of RCTs. The performance of the proposed method is evaluated by an example of a meta-analysis estimating the CACE of epidural analgesia on cesarean section, in which only 10 out of 27 studies reported complete noncompliance data.

e-mail: jzhou@umn.edu

Meta-Analysis of Gene Set Coexpression

Haocan Song*, Vanderbilt University Medical Center
Yan Guo, University of New Mexico
Fei Ye, Vanderbilt University Medical Center

The number of genomic datasets being made publicly available and the number of genomic studies being published are increasing exponentially. The results of individual studies are often insufficient to provide reproducible conclusions, as their results are often inconsistent. We propose a meta-analysis approach to systematically assess the coexpression of a specified gene set by pooling the results from individual studies to provide point and interval estimates. We first construct a coexpression network with gene expression data for a single gene set under two conditions, from which we obtain the vertices with weights estimated by some correlation distance measure. A nonparametric approach that accounts for the correlation structure between genes is used to test whether the gene set is differentially coexpresssed between the two conditions, and a permutation test p-value is computed. Bootstrapping is used to construct confidence intervals of point estimates. A meta-analysis of single proportions is then performed with two options: random-intercept logistic regression model and the inverse variance method, to yield conclusive results over a number of individual studies.

e-mail: haocan.song@vumc.org

124. LONGITUDINAL DATA ANALYSIS

Regression Analysis of Sparse Asynchronous Longitudinal Data with Informative Observation Times

Dayu Sun*, University of Missouri
Hui Zhao, Zhongnan University of Economics and Law
Jianguo Sun, University of Missouri

In longitudinal studies, the responses and covariates may be observed intermittently at different time points, leading to sparse asynchronous longitudinal data. Traditional methods, e.g., the last value carried forward method, could result in large bias and high variation in estimation. Moreover, observation times of responses may carry information about or correlate with response variables. Failing to consider the informative observation times results in biased estimates. In this study, we consider a regression analysis of sparse asynchronous longitudinal data with informative observation times, which has not been thoroughly investigated. A flexible semiparametric transformation conditional model is used to incorporate dependence between observation times and responses. Simple kernel-weighted estimating equations are proposed to deal with discrepancies among observation times of responses and covariates. Extensive simulation studies were carried out to demonstrate that the proposed method performs well and yields less bias than existing methods in the presence of informative observation times. An example of an HIV study illustrates that the proposed method works well in practice.

e-mail: dsryb@mail.missouri.edu

Modeling Continuous Longitudinal Response Data using Ordinal Regression

Yuqi Tian*, Vanderbilt University
Bryan E. Shepherd, Vanderbilt University
Chun Li, Case Western Reserve University
Jonathan S. Schildcrout, Vanderbilt University

Continuous data can be considered to be ordinal data and analyzed with ordinal cumulative probability models (CPMs), also known as “cumulative link models”. In these models, we assume the data follow a known linear model after some unspecified transformation and estimate the transformation nonparametrically. CPMs have many nice features but most research with these models has focused on cross-sectional data. Here we would investigate the use of CPMs in longitudinal settings. We describe extensions of CPMs using generalized estimation equations approaches. CPMs are readily extended to fitting repeated longitudinal data with a working correlation structure of independence. Other working correlation structures are more complicated; we will discuss alternating logistic regression approaches and computational challenges. Longitudinal CPMs will be applied to data from a smoking intervention program to study lung function.

e-mail: yuqi.tian@vanderbilt.edu

ABSTRACTS & POSTER PRESENTATIONS

Novel Joint Models for Identifying Determinants of Cognitive Decline in the Presence of Informative Drop-out and Observation Times

Kendra Davis-Plourde*, Boston University
Yorghos Tripodis, Boston University

Many statistical challenges exist in longitudinal studies of cognition. For instance, cognitive decline is associated with debilitating health consequences, such as death, making the drop-out mechanism informative. Further, participants experiencing cognitive decline might be more/less likely to agree to neuropsychological testing making the observation times informative. In many longitudinal settings a linear mixed effects model is used to estimate cognitive decline, but this method ignores informative drop-out and informative observation times leading to spurious and biased associations. Recently, joint models have become a promising solution to these issues by combining the linear mixed effects model with a Cox proportional hazards model, to take into account informative drop-out, and a frailty model, to take into account informative observation times. In our study we propose a novel joint model specifically tailored to modeling cognitive decline.

e-mail: kldavis@bu.edu

Multiple Imputation of an Expensive Covariate in Outcome Dependent Sampling Designs for Longitudinal Data

Chiara Di Gravio*, Vanderbilt University
Ran Tao, Vanderbilt University
Jonathan S. Schildcrout, Vanderbilt University

Outcome dependent sampling (ODS) designs are efficient when exposure ascertainment costs limit sample size. In longitudinal studies with a continuous outcome, ODS designs can lead to efficiency gains by sampling based on low-dimensional summaries of individual longitudinal trajectories. Analyses can be conducted using only the sampled subjects with data on exposure, outcome, and confounder. Alternatively, one can combine data on the sample subjects with the partial data (i.e., outcome and confounder) from the unsampled subjects to conduct full-likelihood or imputation approaches. In this talk, we will discuss a relatively easy to conduct imputation approach for ODS designs that may be more efficient than the complete data approach, may be easier to implement than full-likelihood approaches, and may be more broadly applicable than other imputation approaches. We will examine finite sampling operating characteristics of the imputation approach under several ODS designs and longitudinal data features (e.g., balance and unbalanced data, equal and unequal cluster sizes), and we will apply the imputation approach to the analysis of lung function in the Lung Health Study.

e-mail: chiara.di.gravio@vanderbilt.edu

Real-Time Regression Analysis of Streaming Clustered Data with Possible Abnormal Data Batches

Lan Luo*, University of Michigan
Peter X.K. Song, University of Michigan

This paper develops an incremental data analytic based on quadratic inference function (QIF) to analyze streaming datasets with correlated outcomes such as longitudinal and clustered data. We propose a renewable QIF (RenewQIF) method in a paradigm of renewable estimation and incremental inference, in which parameter estimates are recursively renewed with current data and summary statistics of historical data, but with no use of any historical subject-level raw data. We show theoretically and numerically that our renewable estimation method is asymptotically equivalent to the oracle generalized estimating equations (GEE) approach that directly processes the entire cumulative subject-level data. We also consider checking the homogeneity assumption of regression coefficients via a sequential goodness-of-fit test as a screening procedure on occurrences of abnormal data batches. Through extensive simulation studies we demonstrate that RenewQIF enjoys both statistical and computational efficiencies. In addition, we illustrate the proposed method by an analysis of streaming car crash datasets from the National Automotive Sampling System-Crashworthiness Data System (NASS CDS).

e-mail: luolsph@umich.edu

Modeling Disease Progression with Time-Dependent Risk Factors and Time-Varying Effects using Longitudinal Data

Jacquelyn E. Neal*, Vanderbilt University
Dandan Liu, Vanderbilt University

Reliable statistical models for risk of disease progression are important, especially for diseases with long duration and slow progression. Given longitudinal data with long follow-up period, commonly used approaches, such as cross-sectional models only using baseline risk factors or Markov assumption based transition model using consecutive visits, could not fully address challenges due to time-dependent risk factors with time-varying effects. To maximize information from a longitudinal perspective, we propose to model disease progression using partly conditional models for ordinal outcome. Advantages of this method include direct modeling of disease progression, use of longitudinal risk factors, allowing time-varying effects, and flexibility in target period of progression. Competing risk of death for the observation process, a common phenomenon for chronic disease, will also be considered. Data from the National Alzheimer's Coordinating Center will be used to illustrate this method.

e-mail: jacquelyn.e.neal@vanderbilt.edu

ABSTRACTS & POSTER PRESENTATIONS

Informative Visit Processes in Longitudinal Data from the Health Sciences

Fridtjof Thomas*, University of Tennessee Health Science Center
Csaba P. Kovesdy, Memphis VA Medical Center
Yunusa Olufadi, University of Memphis

Electronic health records (EHR) allow researchers to access longitudinal data from large hospital networks and that can be obtained on a massive scale as “big data”. Such data originates from patient-doctor interactions and exists for one of two reasons: It is either obtained as part of a regular health screening/monitoring process or intentionally created out of a specific concern for a patient’s health status, often triggered by alarming high or low values as found in previous visits. Most standard models for longitudinal data assume that the existence of a specific measurement value is independent of its expected value and this assumption allows the likelihood function in parametric models to factor and to be analytically tractable. Contrary to that assumption, an informative visit process is defined by the fact that the existence of a measurement is not independent of its expected value. We discuss how to identify such informative visit processes and their effect on parameter estimation and health monitoring in general, and the monitoring of kidney function based on estimated glomerular filtration rate (eGFR) in particular.

e-mail: fthomas4@uthsc.edu

125. HIGH DIMENSIONAL DATA ANALYSIS: THE BIG PICTURE

Capturing Skewness and Sparsity in High Dimensions

Xiaoqiang Wu*, Florida State University
Yiyuan She, Florida State University
Debajyoti Sinha, Florida State University

This paper studies how to model skewed, heteroscedastic, and continuous or semi-continuous (zero-inflated) responses and address the associated variable selection possibly in high dimensions. We propose a Skewed Sparse Regression (SSR) which includes the unknowns of skewness, scale and mean in high dimensions simultaneously. To tackle the computational challenge from non-convex and non-smoothness and diversity of penalties, we combine Blockwise Coordinate Descent and Majorize-Minimization algorithms to develop a highly scalable and efficient algorithm with guaranteed convergence. Our statistical analysis ensures that the computation obtained blockwise-SSR estimators, though not necessarily globally optimal, still enjoy the minimax rate up to a logarithm term at the occurrence of skewness. Extensive simulation studies show that in comparison with state of the art methods, SSR can achieve better estimation accuracy and selection consistency with, however, substantially reduced computation cost. We also demonstrate how a hurdle model derived from SSR can help analyze the Medical Expenditure Panel Survey data.

e-mail: xw15@my.fsu.edu

Efficient Greedy Search for High-Dimensional Linear Discriminant Analysis

Hannan Yang*, University of North Carolina, Chapel Hill
Quefeng Li, University of North Carolina, Chapel Hill

High dimensional classification commonly appears in biomedical research, where traditional linear discriminant analysis is known to result in poor classification per- formance due to error accumulation in estimating the unknown parameters. Most of the recently proposed regularized discriminant analysis methods rely on solving large-scale optimization problems or matrix inversion, which are computationally prohibitive for some ultra-high dimensional biomedical data, such as SNP data. To improve the computational efficiency, we propose an efficient greedy search algorithm for linear discriminant analysis based on learning the increment of the Mahalanobis distance, which only relies on closed-form formula and is scalable to handle ultra-high dimensional data. We give the theoretical guarantee of its statistical properties that the variable selection and error rate consistency are established under some mild as- sumptions on the underlying distributions. We compare our method with the existing regularized methods in simulations and demonstrate that our method has profound improvement of computational efficiency and superior classification performance.

e-mail: hnyang@live.unc.edu

Parallelized Large-Scale Estimation and Inference for High-Dimensional Clustered Data with Binary Outcomes

Wenbo Wu*, University of Michigan School of Public Health
Kevin He, University of Michigan School of Public Health
Jian Kang, University of Michigan School of Public Health

Large-scale electronic health records derived from national disease registries have proven useful in monitoring health care processes and outcomes. However, using high-dimensional data introduces statistical and computational obstacles. When performing a national evaluation of kidney dialysis facilities in the United States, we are confronted with multiple challenges: facility count can be very large, estimation and inference on high-dimensional facility effects are problematic, and traditional methods suitable for small and moderate-sized data cannot be applied to a large-scale context. To address these issues, we propose ParBRNR, a shared-memory parallelized block-relaxation Newton--Raphson algorithm designed for performant estimation on high-dimensional facility effects. Simulated and real data analyses demonstrate its computational efficiency. Further, we develop PET, a Poisson binomial distribution based exact test that provides robust inference, especially on small-sized facility effects. A simulated data experiment justifies its scalability with respect to facility size and count. Funnel plots illustrate PET’s application in visualizing provider profiling.

e-mail: wenbowu@umich.edu

ABSTRACTS & POSTER PRESENTATIONS

A Generalized Framework for High-Dimensional Inference based on Leave-One-Covariate-Out LASSO Path

Xiangyang Cao*, University of South Carolina
Karl Gregory, University of South Carolina
Dewei Wang, University of South Carolina

We propose a new measure of variable importance in high-dimensional regression based on the change in the LASSO solution path when one covariate is left out. The proposed procedure provides a novel way to calculate variable importance and conduct variable screening. In addition, our procedure allows for the construction of p-values for testing whether each coefficient is equal to zero as well as for simultaneous testing of multiple hypotheses; bootstrap techniques are used to construct the null distribution. For low-dimensional linear models, our method has essentially the same power as the t-test. Extensive simulations are provided to show the effectiveness of our method. In the high-dimensional setting, our proposed solution path based test achieves greater power than some other recently developed high-dimensional inference methods.

e-mail: caoxiangyang93@gmail.com

Iterative Algorithm to Select Vine Copula According to Expert Knowledge and Pairwise Correlations

Philippe Saint Pierre*, University of Toulouse
Nazih Benoumechiara, Sorbonnes University
Nicolas J. Savy, University of Toulouse

Estimation of a multivariate distribution becomes tricky when the number of covariates increases since the correlations between variables should be captured to obtain a credible estimation. The copula approach presents some drawbacks when the number of variables is greater than two. Vine model has been proposed to deal with more variables. However difficulties still occur in the context of high dimension to estimate the vine copula in particular when databases contain limited numbers of individuals. Moreover, in several applications, expert knowledges should be taken into account in the vine copula construction to obtain a more realistic model from a clinical point of view. We propose a greedy algorithm starting from the independence copula and adding one pair of variables to the R-vine structure at each iteration. The ranking of candidate pairs of variables is fixed according to pairwise correlations and expert knowledge. The R-vine structure is filled with independent pair-copulas except for the selected pairs of variables. Doing so, the estimation of the R-vine structure is relevant from a clinical point of view and can be performed in a reasonable computational time.

e-mail: Philippe.Saint-Pierre@math.univ-toulouse.fr

126. CLINICAL 'TRIALS AND TRIBULATIONS'

Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Adjustment for Additional Baseline Variables

Bingkai Wang*, Johns Hopkins University
Michael Rosenblum, Johns Hopkins University
Ryoko Susukida, Johns Hopkins University
Ramin Mojtabai, Johns Hopkins University
Masoumeh Aminesmaeili, Johns Hopkins University

We focus on estimating the average treatment effect in clinical trials that involve stratified randomization. It is important to understand the large sample properties of estimators that adjust for stratum variables and additional baseline variables, since this can lead to substantial gains in precision and power. Surprisingly, this is an open problem. It was only recently that a simpler problem was solved by Bugni et al. (2018) for the case with no additional baseline variables, continuous outcomes and the ANCOVA estimator. We generalize their results in three directions. First, we handle binary and time-to-event outcomes. Second, we allow adjustment for additional baseline variables. Third, we handle missing outcomes under the missing at random assumption. We prove that a wide class of estimators is asymptotically normally distributed under stratified randomization and has equal or smaller asymptotic variance than under simple randomization. For each estimator in this class, we give a consistent variance estimator. The above results also hold for the biased-coin design. We demonstrate our results using completed trial data sets of treatments for substance use disorder.

e-mail: bingkai.w@gmail.com

ABSTRACTS & POSTER PRESENTATIONS

Dynamic Borrowing in the Presence of Treatment Effect Heterogeneity

Ales Kotalik*, University of Minnesota
David Vock, University of Minnesota
Eric Donny, Wake Forest School of Medicine
Dorothy Hatsukami, University of Minnesota
Joseph Koopmeiners, University of Minnesota

A number of statistical approaches have been proposed for incorporating supplemental information in randomized trials. Existing methods often compare the marginal treatment effects to evaluate the degree of consistency between sources. Dissimilar marginal treatment effects would either lead to increased bias or down-weighting of the supplemental data. This represents a limitation in the presence of treatment effect heterogeneity. We introduce the concept of conditional exchangeability, where differences in the marginal treatment effect can be explained by the different distributions of the effect modifiers. The potential outcomes framework is used to conceptualize conditional and marginal exchangeability. We utilize a linear model and the multisource exchangeability models framework to facilitate borrowing when conditional exchangeability holds. We investigate the operating characteristics of our method using simulations. We also illustrate our method using data from clinical trials of very low nicotine cigarettes. Our method has the ability to incorporate supplemental information in a wider variety of situations than when only marginal exchangeability is considered.

e-mail: kotal004@umn.edu

Bayesian Methods to Compare Dose Levels to Placebo in a Small n Sequential Multiple Assignment Randomized Trial (snSMART)

Kimberly A. Hochstedler*, University of Michigan
Fang Fang, University of Michigan
Roy N. Tamura, University of South Florida
Thomas M. Braun, University of Michigan
Kelley M. Kidwell, University of Michigan

Clinical trials studying treatments for rare diseases are challenging to design and conduct due to the limited number of patients eligible for the trial. One design used to address this challenge is the small n, sequential, multiple assignment, randomized trial (snSMART). We propose a new snSMART design to investigate the response rates of a drug tested at a low and high dose compared to placebo. Patients are randomized to an initial treatment (stage 1). In stage 2, patients are re-randomized, depending on their stage 1 treatment and response, to either the same or a different dose of treatment. Data from both stages are used to determine the efficacy of the active treatment. We present a Bayesian approach where information is borrowed between stages 1 and 2, and compare it to methods using only stage 1 data and a log-Poisson joint stage model. Our Bayesian method has smaller root-mean-square-error and 95% credible interval widths than standard methods in the tested scenarios. We conclude that it is advantageous to utilize data from both stages for a primary efficacy analysis and that our snSMART design can be used to register a drug for the treatment of rare diseases.

e-mail: hochsted@umich.edu

Sample Size Calculation in Comparative Clinical Trials with Longitudinal Count Data: Incorporation of Misspecification of the Variance Function and Correlation Matrix

Masataka Igeta*, Hyogo College of Medicine
Shigeyuki Matsui, Nagoya University Graduate School of Medicine

Longitudinal count data are frequently compared between treatment groups in clinical trials where the primary endpoint is occurrence of clinical events. Some sample size formulae for such trials have been developed under longitudinal gamma frailty models or semi-parametric marginal models with a specified covariance matrix, involving a specified variance function and correlation matrix. In this paper, we propose a sample size formula anticipating misspecification of the covariance matrix. We derived the formula from an asymptotic distribution of the Wald-type test with a sandwich-type robust variance estimator under the null hypothesis based on a generalized estimating equation. The asymptotic variance of the treatment effect estimator involves both true and working covariance matrices under a misspecification of the working covariance matrix. Our formula includes some existing formulae as special cases under some specifications of the true and working covariance matrices. We consider a sensitivity analysis for misspecifications of the true covariance matrix. The adequacy of our formulae is evaluated via simulation studies. An application to a clinical trial is provided.

e-mail: masataka.igeta@gmail.com

Sequential Interval Estimation of Patient Accrual Rate in Clinical Trials

Dongyun Kim*, National Heart Lung and Blood Institute, National Institutes of Health
Sung-Min Han, OSEHRA

An adequate patient accrual is a critical component for the success of a clinical trial. In this talk we introduce a new monitoring method for patient accrual by constructing a confidence interval using a fully sequential procedure as new accrual data become available. The confidence interval is asymptotically optimal, and it does not require a priori estimation of population variance. We illustrate the method using real data from two well-known phase III clinical trials.

e-mail: kimd10@nhlbi.nih.gov

Statistical Analysis of Glucose Variability

Jiangtao Luo*, Eastern Virginia Medical School
Ismail El Moudden, Eastern Virginia Medical School
Mohan Pant, Eastern Virginia Medical School

Glucose is a significant variable is the study of diabetes. There are several methods to describe the glucose variability. We will study their relationship from statistical viewpoint. In addition simulation and real data example are given to demonstrate the relationship.

e-mail: luoj@evms.edu

ABSTRACTS & POSTER PRESENTATIONS

The Impact of Precision on Go/No-Go Decision in Proof-of-Concept Trials

Macaulay Okwuokenye*, Brio Dexteri Pharmaceutical Consultants

Pharmaceutical sponsors are increasingly deploying quantitative approaches to inform Go/No-Go (GNG) decision in Proof-of-Concept (PoC) trials. These trials are relatively small and expectedly at risk for producing high random variation in the treatment effect estimate. This begs for attention to efficient study design and method of analysis. It is advocated that PoC trials may be designed based on number of patients needed to obtain a desired probability of indeterminate decision. We note that under certain conditions, designing proof-of-concept (PoC) trials based on the number of patients needed to obtain a desired probability of indeterminate (Consider zone) decision is equivalent to designing PoC trials based on approximate target precision of treatment effect. We evaluate the impact of precision on operating characteristics under three-outcome GNG decision framework. Results suggest that precision and precision improvement through efficient study design and analysis methods impact three-outcome GNG decision. Urinary albumin-to-creatinine ratio (UACR) in kidney disease is used for illustration.

e-mail: Chiefendorce@hotmail.com

127. COUNT DATA: THE THOUGHT THAT COUNTS

Probabilistic Canonical Correlation Analysis for Sparse Count Data

Lin Qiu*, The Pennsylvania State University
Vernon M. Chinchilli, The Pennsylvania State University

Integrative analysis between two high-dimensional omics data sets becomes more and more popular. Canonical correlation analysis (CCA) is a classical and important multivariate technique for exploring the relationship between two sets of continuous variables. Although some sparse CCA methods are developed to deal with high-dimensional problems, they are designed specifically for continuous data and do not consider the integer values from Next generation sequencing platform with very low counts for some important features. We propose, for the first time, a model-based probabilistic approach for correlation and canonical correlation analysis for two sparse count data sets (SPCCA). We demonstrate through extensive simulation studies that SPCCA outperforms existing approaches on both correlation and canonical correlation estimation. We further apply the SPCCA method to study the association between miRNA and mRNA zero-inflated count expression data sets from a squamous cell lung cancer study, finding that SPCCA can uncover a large number of strongly correlated pairs than standard correlation approaches and other sparse CCA.

e-mail: luq7@psu.edu

Bayesian Credible Subgroups for Count Regression and Its Application to Safety Evaluation in Clinical Studies

Duy Ngo*, Western Michigan University
Patrick Schnell, The Ohio State University
Shahrul Mt-Isa, MSD Research Laboratories
Jie Chen, Merck & Co., Inc.
Greg Ball, Merck & Co., Inc.
Dai Feng, AbbVie Inc.
Richard Baumgartner, Merck & Co., Inc.

Evaluation of safety in clinical trials is generally concerned with the occurrence of adverse events. A common approach in safety analysis typically entails testing for differences in adverse event rates between treatment groups. In this context, Poisson regression can be used for statistical analysis of count data. We focus on subgroup identification with respect to counts of adverse events. To this end, we propose a zero-inflated Bayesian credible subgroup approach for adverse event count data in which the data exhibit excess zeros. Our approach enables control of multiplicity and identification of a subgroup of patients with a non-inferiority analysis. Furthermore, we discuss potential extensions to simultaneous modeling of efficacy and safety with respect to benefit-risk considerations. We demonstrate the performance of our method in a simulation study and in a real-world application.

e-mail: duy.ngo@wmich.edu

Analysis of Panel Count Data with Time-Dependent Coefficient and Covariate Effects

Yuanyuan Guo*, University of Missouri, Columbia
Jianguo Sun, University of Missouri, Columbia

Panel count data occurred when events are observed at finite fixed time points and the visit times vary from subject to subject, and the exact event times are unknown. In the past decades, there have been extensive researches to study the proportional mean model for panel count data, if we consider only time independent coefficients and covariates. When we account for time dependent coefficient, spline method is an important method, generally used to study for survival data. Nevertheless, limited work has been done in panel count data. Furthermore, in practice, the time-dependent covariate effects situation are common. Limited research has been found here. In this paper, we consider situations where coefficient and covariate effects are time-dependent simultaneously. Based on the conditional estimating equations method developed for time-dependent covariates, we approximate the coefficients by Bernstein splines, hence allow both coefficients and covariates to be time-dependent. We conduct a comprehensive simulation study to prove the operational characteristics of the proposed method. A real data example will be analyzed for illustration.

e-mail: yg882@mail.missouri.edu

ABSTRACTS & POSTER PRESENTATIONS

Semi-Parametric Generalized Linear Model for Binary Count Data with Varying Cluster Sizes

Xinran Qi*, Medical College of Wisconsin
Aniko Szabo, Medical College of Wisconsin

The semi-parametric generalized linear model (SPGLM) developed by Rathouz and Gao (2009) assumes that the response is from a general exponential family with unspecified reference distribution and can be applied to model the distribution of binary count data with a constant cluster size. We extend the SPGLM to model response distributions of such data with varying cluster sizes by assuming marginal compatibility. The proposed model combines a non-parametric reference describing the within-cluster dependence structure with a parametric density ratio characterizing the covariate effect. It avoids making parametric assumptions about higher-order dependence and is more parsimonious than non-parametric models. We fit the SPGLM with an Expectation Maximization Newton-Raphson algorithm to a developmental toxicology dataset and compare the estimates with existing methods.

e-mail: xinqi@mcw.edu

Drug Safety Evaluation Using Panel Count Model

Yizhao Zhou*, Georgetown University
Ao Yuan, Georgetown University
Ming Tan, Georgetown University

In FDA drug safety evaluation, thousands of reported adverse events (AE) are associated with thousands of drugs under the adverse event reporting system. The data is in the form of a large I*J table, with nij being the reported number of AE's for the i-th AE and j-th drug. The data are collected for a large number of users over multiple years. As no adverse events have been observed for a lot of drugs for many years, the challenges are how to handle the large number of excessive zero counts, and incorporate potential covariates. To handle these problems, we propose a panel count model which assumes a non-homogenous Poisson process yij=yij (t) for counts having conditional mean E(yij (t)|zj,xj) = G(t)exp(βT xij+αi zij), where xij are covariates, zj is the vector of length I: zij=I(nij>0); and G(·)≥ 0 is an unspecified monotone increasing function. The EM algorithm and isotonic technique are used to estimate parameters.

email: yz459@georgetown.edu

Measurement Error Modeling for Count Data

Cornelis J. Potgieter*, Texas Christian University

Count variables often have a fixed set of possible outcomes {0,...,N}. Such variables pose a unique problem when measurement error is present in observed data. Specifically, as both the true count X and the noisy count W lie in the fixed set of values, the typical additive model W = X + U with independent measurement error U is unsuitable. While it is possible to extend misclassification models for categorical variables to the count variable framework, the resulting model is tedious and has a large number of parameters to estimate. We explore how a transformation model W = max(0, min(X+U)) can be used to account for measurement error. In this formulation, the additivity and independence of measurement error can still be utilized. The goal of our study is two-fold: Firstly, we propose discrete parametric distributions for the measurement error U that are symmetric and zero mean. Secondly, we consider the companion problems of estimating the pmf of X based on W data and estimating the conditional mean function E[X|W]. The methodology is illustrated in an application involving the assessment of reading abilities of elementary school children.

email: c.potgieter@tcu.edu

Conditional Mutual Information Estimation for Discrete and Continuous Data with Nearest Neighbors

Octavio Mesner*, Carnegie Mellon University
Cosma Rohilla Shalizi, Carnegie Mellon University

Mutual information is an attractive statistic for many applications because it completely captures the dependence between two random variables or vectors. Conditional mutual information (CMI) is particularly useful in settings, e.g. causal discovery, where it is necessary to quantify dependence between a pair of variables that may be mediated by other variables. CMI's usage is rare in fields such as epidemiology, public policy, and social sciences due to its inability to handle mixtures of continuous and discrete random variables. While progress has been made on estimating mutual information for discrete and continuous variables, a CMI estimation method does not currently exist. This paper builds on prior research to develop a novel method for non-parametric CMI estimation for discrete and/or continuous variables. For each point, the method locally estimates CMI using its nearest neighbors then averages all local estimates. If a point's nearest neighbor occupies that same location, the method recognizes that the point is likely discrete and alters the counting process. We prove that this estimator is consistent theoretically and demonstrate its performance empirically as well.

email: omesner@cmu.edu

ENAR

11130 Sunrise Valley Drive, Suite 350
Reston, Virginia 20191
Phone: 703-234-4146
Fax: 703-234-4147
www.enar.org