International Biometric Society
Eastern North American Region

© 2006, Kevin C. Rose, AtlantaPhotos.com

March 11–14, 2007

Statistical Science
Solving Problems That Matter

Program and Abstracts

© 2006, Kevin C. Rose, AtlantaPhotos.com

Hyatt Regency Atlanta
Atlanta, Georgia

© 2006, Kevin C. Rose, AtlantaPhotos.com

IBS ENAR

Spring Meeting with IMS and Sections of ASA

Are you looking for the groundbreaking achievement, aggressive growth potential, and unlimited resources to excel? PPD is a leading global provider of discovery and development services and products for pharmaceutical, biotechnology and medical device companies. PPD empowers you to be truly great – to innovate, to explore and to make decisions as part of a global team working for a financially secure company that is positioned for strong growth.

break new ground

Our **Biostatistics** Department offers outstanding opportunities in clinical trials for candidates of all levels. If you are an expert in statistical methodology, have strong scientific writing ability and outstanding communication skills, we want you to apply today. We have positions available in the following drug development programs:

- **Associate Director, Biostatistics**
- **Manager, Biostatistics**
- **Sr. Biostatisticians**
- **Biostatisticians**
- **Associate Biostatisticians**
- **Programmers (All Levels)**
- **All titles available for Management and Technical Scientist career tracks**

Please visit our booth at the ENAR conference. We are giving away 2 iPod nano's daily at the booth, come by and register. Or visit us online at: www.ppdi.com and apply today.

We have career opportunities at these four locations:
Research Triangle Park, NC; Wilmington, NC; Austin, TX; and Hamilton, NJ.

At PPD, we strive to reward our employees' initiative with benefits that promote a work/life balance. We offer aggressive salaries and bonus potential, comprehensive benefits package, 401(k), stock purchase program and much more. Break into a new and exciting career challenge.

A diverse workforce enables us to achieve breakthrough advancements—an Equal Opportunity Employer.

PPD®

WWW.PPDI.COM

Table of Contents

Acknowledgements 4
Officers and Committees 5
Programs 6
Representatives 6
Fostering Diversity Workshop 7
Student Award Committee 7
Student Award Winners 7
Special Thanks 8
Presidential Invited Speaker 9
IMS Medallion Lecturer 9
Short Courses .... 10 – 13
Tutorials .... 14 – 15
Roundtables .... 16 – 17
Future Meetings of the International Biometric Society 18
Program Summary .... 19 – 23
Scientific Program: Poster Presentations .... 24 – 28
Scientific Program: Oral Presentations .... 29 – 63
Notes 64
Poster Presentation Abstracts .... 65 – 107
Oral Presentation Abstracts .... 109 – 381
Index of Participants .... 382 – 393
Notes 108, 394, 398
Hyatt Regency Atlanta Floor Plan .... Back Cover

Acknowledgements

ENAR would like to acknowledge the generous support of the 2007 Local Arrangements Committee, the faculty and staff of the Department of Biostatistics, Rollins School of Public Health, Emory University and the ENAR Sponsorship Committee – Frank Shen, B. Christine Clark, and Kannan Natarajan.

SPONSORS

We gratefully acknowledge the support of:

Abbott Laboratories
Amgen, Inc.
AstraZeneca
Biogen Idec
Bristol Myers Squibb Co.
Cephalon, Inc.
Cytel Inc.
Daiichi Sankyo
Emory University, Department of Biostatistics, Rollins School of Public Health
GlaxoSmithKline
Icon Clinical Research
Inspire Pharmaceuticals
Johnson & Johnson
Merck & Company, Inc.
Merck Inc.
Novartis Pharmaceuticals Inc.
Pfizer, Inc.
PPD, Inc.
Rho, Inc.
SAS
Schering-Plough Research Institute
Smith Hanley Associates LLC
Statistics Collaborative
Takeda Global Research & Development

EXHIBITORS

We gratefully acknowledge the support of:

Allergan, Inc.
Amgen, Inc.
Biostat, Inc.
Blackwell Publishing, Inc.
Cambridge University Press
The Cambridge Group, Ltd.
CRC Press – Taylor and Francis Group
Cytel Inc.
Insightful Corporation
JMP
Kforce Clinical Research Staffing
Oxford University Press
PDL BioPharma
PPD, Inc.
Salford Systems
SAS
SAS Publishing
Smith Hanley Associates LLC
Springer
Statistical Solutions
John Wiley & Sons

Officers and Committees

January – December 2007

EXECUTIVE COMMITTEE – OFFICERS

President	Lisa LaVange
Past President	Jane Pendergast
President-Elect	Eric (Rocky) Feuer
Secretary (2007-2008)	José Pinheiro
Treasurer (2006-2007)	Oliver Schabenberger

REGIONAL COMMITTEE (RECOM)

President (Chair) Lisa LaVange
Eight members (elected to 3-year terms):

2005-2007	2006-2008	2007-2009
Gregory Campbell	John Bailer	Karen Bandeen-Roche
Naisyin Wang	Stacy Lindborg	F. DuBois Bowman
	Tom Ten Have	Paul Rathouz

REGIONAL MEMBERS OF THE COUNCIL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Marie Davidian, Roderick Little, Ron Brookmeyer, Louise Ryan, Janet Wittes

APPOINTED MEMBERS OF REGIONAL ADVISORY BOARD (3-YEAR TERMS)

Chair: Scarlett Bellamy
Chair-Elect: Amy Herring

2005-2007	2006-2008	2007-2009
Barbara Bailey	Michael Hardin	Christopher S. Coffey
Sudipto Banerjee	Eileen King	Hormuzd A. Katki
Jason Connor	Carol Lin	Lan Kong
Todd Durham	Keith Muller	Yi Li
Kirk Easley	Soomin Park	Lillian Lin
Abie Ekangaki	Shyamal Peddada	Laura Meyerson
Deborah Ingram	Jeremy Taylor	Gene Pennello
Xuejen Peng	Melanie Wall	Tamara Pinkett
James Rosenberger	Position to be Filled	John Preisser
Maura Stokes	Position to be Filled	Douglas E. Schaubel

Programs

2007 Spring Meeting – Atlanta, GA
Program Chair: Amy Herring
Program Co-Chair: Gene Pennello
Local Arrangements Chair: Robert Lyles

2008 Spring Meeting – Crystal City, VA
Program Chair: Avital Cnaan
Program Co-Chair: Kyungmann Kim
Local Arrangements Co-Chairs: Guoqing Diao and Kimberly Drews

2007 Joint Statistical Meeting
Christopher Coffey

2008 Joint Statistical Meeting
Robert Johnson

Biometrics Executive Editor
Marie Davidian

Biometrics Co-Editors
Lawrence Freedman, Geert Molenberghs, and Naisyin Wang

Biometric Bulletin Editor
Ranny Dafni

ENAR Correspondent for the Biometric Bulletin
Rosalyn Stone

ENAR Executive Director
Kathy Hoskins

International Biometric Society Business Manager
Claire Shanley

Representatives

Committee Of Presidents Of Statistical Societies (Copss)
ENAR Representatives
Lisa LaVange (President)
Jane Pendergast (Past-President)
Eric Feuer (President-Elect)

ENAR Standing/Continuing Committee Chairs

Nominating (2006)	Peter Imrey
Nominating (2007)	Jane Pendergast
Sponsorship (2006 - 2007)	Frank Shen
Sponsorship (2007 - 2008)	B. Christine Clark
Information Technology Oversight (ITOC)	Bonnie LaFleur

ENAR Representative on the ASA Committee on Meetings
Maura Stokes (January 2006-December 2008)

American Association For The Advancement Of Science (Joint with WNAR) Terms through February 22, 2008

Section E, Geology and Geography	Stephen Rathbun
Section N, Medical Sciences	Joan Hilton
Section G, Biological Sciences	Geof Givens
Section U, Statistics	Mary Foulkes
Section O, Agriculture	Kenneth Porter

National Institute Of Statistical Sciences (ENAR President is also an ex-officio member) Board of Trustees
Member: Lisa LaVange

Visit the ENAR website (*www.enar.org*) for the most up to date source of information on ENAR activities.

Fostering Diversity Workshop

Scarlett Bellamy (co-organizer)
DuBois Bowman (co-organizer)
Cassandra Arroyo
Stacy Lindborg
Amita Manatunga
Renee Moore
Dionne Price
Dejuran Richardson
Louise Ryan
Kimberly Sellers
Keith Soper
Mahlet Tadesse
Tom Ten Have
Lance Waller

ENAR Student Award Committee

Peter B. Imrey (Chair)
Sudipto Banerjee
Jianwen Cai
Philip Dixon
David Dunson
Montserrat Fuentes
Elizabeth S. Garrett-Mayer
Liang Li
Jeffrey S. Morris
Jean Opsomer
Sunil Rao
Joshua Tebbs
Alicia Toledano

Student Award Winners

Van Ryzin Award Winner

Sijian Wang, University of Michigan

Award Winners

Ping Bai, University of North Carolina-Chapel Hill
Man-Hua Chen, University of Missouri-Columbia
Nam Hee Choi, University of Michigan
Wentao Feng, University of Pittsburgh
Jonathan Gelfond, University of North Carolina-Chapel Hill
Xin He, University of Missouri-Columbia
Emma Huang, University of North Carolina-Chapel Hill
Satkartar Kinney, Duke University
Shengde Liang, University of Minnesota
Yan Lin, University of Pittsburgh
Minggen Lu, University of Iowa
Tao Lu, University of South Florida
Sheng Luo, The Johns Hopkins University
Abel Rodriguez, Duke University
James Slaughter, University of North Carolina at Chapel Hill
Wenguang Sun, University of Pennsylvania
Luping Zhao, University of Minnesota

SPECIAL THANKS

2007 ENAR Program Committee
Amy H. Herring (Chair), University of North Carolina at Chapel Hill
Gene Pennello (Co-Chair), US Food and Drug Administration
Constantine Frangakis, Johns Hopkins University
Debashis Ghosh, University of Michigan
Laura Meyerson, Biogen IDEC, Inc
Jeffrey Morris, University of Texas MD Anderson Cancer Center

ASA Section Representatives
Amit Bhattacharyya (Biopharmaceutical Section), GlaxoSmithKline
Larry Cox (Section on Statistics in Defense and National Security),
National Center for Health Statistics
Steven Heeringa (Section on Survey Research Methods), University of Michigan
John Preisser (Biometrics Section), University of North Carolina at Chapel Hill
Duane Steffey (Section on Risk Analysis), Exponent, Inc.
Patrick Tarwater (Section on Epidemiology and Section on Teaching Statistics in the Health Sciences),
University of Texas
Linda Young (Section on Statistical Education), University of Florida
Jun Zhu (Section on Statistics and the Environment), University of Wisconsin

IMS Program Chair
David Banks, Duke University

ENAR Education Advisory Committee
John Bailer, Miami University
David B. Dunson, National Institute of Environmental Health Sciences
Barry Graubard, National Cancer Institute
Maura Stokes, SAS Institute
Jane Pendergast, University of Iowa

Local Arrangements Chair
Robert Lyles, Emory University

ENAR Student Awards Chair
Peter Imrey, Cleveland Clinic Foundation

ENAR Diversity Workshop Chair
Scarlett Bellamy, University of Pennsylvania

ENAR Presidential Invited Speaker

Frank W. Rockhold, Ph.D.

Frank Rockhold is currently Senior Vice President, Biomedical Data Sciences at GlaxoSmithKline Pharmaceuticals Research and Development. This includes statistics, epidemiology and health care informatics. In his 16 years at GSK he has also held management positions within the Statistics Department and Clinical Operations both in R&D and in the U.S. Pharmaceutical Business. Dr. Rockhold has previously held positions of Research Statistician, Lilly Research Laboratories (1979-1987) and Executive Director of Biostatistics, Data Management, and Health Economics, Merck Research Laboratories, (1994-1997). He has a BA in Statistics, from the University of Connecticut, a Sc.M. in Biostatistics, from Johns Hopkins University, and a Ph.D. in Biostatistics, from the Medical College of Virginia. He is Past-President, Society for Clinical Trials, Past Chair, PhRMA Biostatistics Steering Committee, and a member of the ICH E-9 and E-10 Expert Working Groups. He has previously served as Associate Editor for *Controlled Clinical Trials*. He has held several academic appointments in his career at Butler University, Indiana University and currently is Adjunct Professor of Health Evaluation Sciences, Penn State University and Adjunct Scholar in the Department of Epidemiology and Biostatistics at the University of Pennsylvania. Dr. Rockhold is currently on the Board of Directors of the Clinical Data Interchange Standards Consortium (CDISC), a member of the WHO Scientific Advisory Group on Clinical Trial Registration, a member of the Institute of Medicine Committee on Clinical Trial Registers, a member of the National Library of Medicine Advisory group for the ClinicalTrials.Gov, Senior Advisor for the PhRMA Biostatistics Technical Group, and a member of the PhRMA Clinical Leadership Committee. He is a Fellow of the American Statistical Association and on the Board of Trustees of the National Institute of Statistical Sciences. Dr. Rockhold has over 100 publications and external presentations.

IMS Medallion Lecturer

Robert John Tibshirani, Ph.D.

Robert Tibshirani is Professor of Health Research and Policy and Professor of Statistics at Stanford University. He received bachelor's degrees in statistics and computer science at the University of Waterloo, a master's degree in statistics at the University of Toronto, and a Ph.D. in statistics at Stanford under the guidance of Professor Bradley Efron. Before returning to Stanford, Dr. Tibshirani was Professor of Statistics at the University of Toronto. Dr. Tibshirani is a Fellow of the Royal Society of Canada, the Institute of Mathematical Statistics, and the American Statistical Association. His many honors include the COPSS Award, the CRM-SSC Prize in Statistics, a NSERC E.W.R. Steacie Fellowship, and a J. Guggenheim Foundation Fellowship. He has served as Associate Editor of *JASA Theory and Methods*, *PLOS Biology*, and the *Canadian Journal of Statistics* and is currently Associate Editor of the *Annals of Applied Statistics*. He has chaired numerous professional committees and is a reviewer for the National Science Foundation and NIH. The author of three popular statistics texts, he has over 190 peer-reviewed manuscripts in print.

Short Courses

DATE: Sunday, March 11, 2007

Full Day Fee	
Members	$220 ($245 after 2/10)
Nonmembers	$270 ($290 after 2/10)

Half Day Fee	
Members	$145 ($170 after 2/10)
Nonmembers	$185 ($210 after 2/10)

Short Course Registration	
Saturday, March 10	3:00–5:00 p.m.
Sunday, March 11	7:00–8:30 a.m.

SC1: MISSING DATA METHODS IN REGRESSION MODELS

(Full Day: 8:30 A.M. – 5:00 P.M.
Hanover AB (Exhibit Level)

Instructors:
Joseph G. Ibrahim, University of North Carolina at Chapel Hill; Ming-Hui Chen, University of Connecticut

Description:
Statistical inference with missing data is a very important problem since missing values are frequently encountered in practice. In fact, most statistical problems can be considered incomplete because not all variables are observed for each unit (or possible unit) in a study. For example, randomization in a clinical trial generates missing values since the outcome that would have been observed had a subject been randomized to a different treatment group is not observed.

Missing values can be both planned and unplanned. Unplanned missing data can arise when study subjects fail to report to a clinic for monthly evaluations, when respondents refuse to answer certain questions on a questionnaire, or when data are lost. On the other hand, data can be missing by design in a randomized clinical trial or in a Latin square experimental design. Although the problems associated with incomplete data are well-known, they are often ignored, and the analysis is restricted to those observations with complete data. This method of analysis is still the default method in most software packages despite the development of statistical methods that handle missing data more appropriately. In particular, likelihood-based methods, multiple imputation, methods based on weighted estimating equations, and fully Bayesian methods have gained increasing popularity since they have become more computationally feasible in recent years. In this short course, we examine each of these methods in some detail, and compare and contrast them under various settings. In particular, we will examine missing covariate and response data in generalized linear models, random effects models, and survival models. Ignorable missingness as well as non-ignorable missingness will be presented for theses models, as well as frequentist and Bayesian methods for analysis. The newly developed statistical package XMISS (Cytel Software) will be used and demonstrated for several real data examples. In addition, live demos of the XMISS software and data analysis using the various models will be given using an LCD projector.

The course presents a balance between theory and applications, and for each class of methods and models discussed, detailed examples and analyses from case studies are presented whenever possible. The applications are all essentially from the health sciences including cancer, AIDS, epidemiology, and the environment. Overall, this course will be applied in nature and will focus on the applications of frequentist and Bayesian methods for research problems arising in the medical sciences. Live demo real data examples will be given using the XMISS software.

Pre-requisites:
The prerequisites include one course in statistical inference and Bayesian theory at the level of Casella and Berger (1990) and Box and Tiao (1992). The audience's background should include some familiarity with basic regression models as well as some Bayesian models. Exposure to two or three conjugate models should provide enough background to motivate the examination of the types of models upon which this short course will focus. The audience should also have some exposure to longitudinal data, generalized linear models, and survival analysis at the introductory level, as this will help motivate the goals of the many examples. Thus, this course would be most suitable for those who are second- or third-year graduate students in statistics or biostatistics, or who have received MS or Ph.D. degrees in statistics, biostatistics, or other related fields.

Short Courses

SC2: SEMIPARAMETRIC THEORY AND MISSING DATA

(Full Day: 8:30 A.M. - 5:00 P.M.)

Learning Center (Ballroom Level)

Instructor:
Anastasios A. Tsiatis, North Carolina State University

Description:
Semiparametric models, which involve both a parametric and non-parametric component, have gained great popularity because of their flexibility and applicability to many statistical problems. The most popular semiparametric model is undoubtedly the proportional hazards model, introduced by Cox (1972) for regression analysis of survival data, where the non-parametric component is the unspecified baseline hazard function. However, more general regression models are also semiparametric models. For example, the model

$$E(Y \mid X) = \mu(X, \beta)$$

where Y is a response variable, X is a vector of covariates, and $\mu(X, \beta)$ is a linear or nonlinear function of the covariates X involving a finite number of parameters β is a semiparametric model because, although the conditional expectation of Y given X is given in terms of a finite number of parameters β, other features of the joint distribution of Y and X are left unspecified. Considerable research on the theoretical properties of estimators for parameters such as β, carried out in the past 20 years, has led to a wealth of knowledge regarding inference under important semiparametric models.

Often, some data required to carry out an intended analysis may be missing. As is well known, failure to account appropriately for such missing data, e.g., by using so called “complete-case” analysis, can lead to severe biases in many instances. There has been a great deal of research on methods for handling missing data, including likelihood and imputation techniques, most of it focused on parametric models. In a seminal paper, Robins, Rotnitzky, and Zhao (1994) introduced the notion of augmented inverse probability weighted complete-case (AIPWCC) estimators for parameters in both parametric and semiparametric models, which offer another approach to handling missing data. AIPWCC methods have since generated widespread interest.

This full-day short course will introduce some of the key theoretical and methodological developments leading to AIPWCC methods through two distinct sessions. In the morning session, theory and methods for general semiparametric models assuming there are no missing data; i.e., the full-data problem, will be introduced. Semiparametric models will be defined formally and theory for inference reviewed. Most practical estimators for the parameters in either parametric or semiparametric models are *asymptotically linear* in that they can be approximated by a sum of iid random variables referred to as the *influence function*. It will be demonstrated that the large sample properties of estimators in these models are directly related to those of their influence functions; specifically, the asymptotic variance of an estimator is equal to the variance of its influence function. The beauty of semiparametric theory is that influence functions can be viewed as geometric objects, i.e. vectors in a linear space where distance away from the origin is related to the variance of the influence function, and estimators whose influence functions have small variance (short distance) are desirable. This geometric perspective will be shown to be the key to construction of estimators with good properties in both parametric and semiparametric models, including the *efficient estimator*, that with the smallest asymptotic variance, through the use of projections onto appropriate linear subspaces. Several pedagogical examples will be used to illustrate these developments.

In the afternoon session, these ideas will be extended to missing data problems, and more generally to coarsened-data problems. The usual taxonomy of missing data mechanisms (missing completely at random, MCAR; missing at random, MAR; non-missing at random, NMAR) will be introduced. With a focus on settings where data are MAR, the geometric perspective for full-data problems involving semiparametric models will be extended to the missing data context, and it will be shown that this development leads naturally to the AIPWCC estimators and to a deeper understanding of the underlying structure of missing data problems. Use of this theory to construct parameter estimators that are as efficient as possible while maintaining practical feasibility will be demonstrated.

Pre-requisites:
Participants should have knowledge of probability and inference at the advanced doctoral level and should feel comfortable with large sample theory concepts, such as convergence in probability and in distribution. Some prior exposure to functional analysis and Hilbert spaces would be helpful, but is not required; key results regarding Hilbert spaces necessary for mastery of the material will be reviewed.

Short Courses

SC3: ADVANCES IN LATENT VARIABLE MIXTURE MODELING USING MPLUS

(Full Day: 8:30 A.M. - 5:00 P.M.)
Hanover C (Exhibit Level)

Instructor:
Bengt Muthén, UCLA

Description:
This short course discusses recent advances in latent variable modeling. Models and applications are discussed using the general modeling framework of the Mplus program (*www.statmodel.com*). The generality of the Mplus framework comes from the unique use of both continuous and categorical latent variables. While continuous latent variables have seen frequent use in factor analysis, structural equation modeling, and random effects growth modeling, modeling that includes categorical latent variables, i.e. finite mixture modeling, is less wide spread in practice. The short course focuses on recently developed models that use categorical latent variables, either alone or together with continuous latent variables. An overview of conventional and new techniques is given including complier-average causal effect estimation (regression mixture analysis), latent class analysis, factor and IRT mixture modeling, latent transition analysis, growth mixture modeling, and survival analysis with latent variables. For each topic, issues of model specification, identification, ML estimation, testing, and model modification are discussed. Several health examples are examined. Modeling strategies are presented. Mplus input setups are provided and Mplus output is used for interpretation of analysis results. The presentation is in lecture format with no need for computer analyses. Following is a list of topics to be covered.

Cross-sectional Modeling With Categorical Latent Variables

- Linear and logic regression mixture analysis
- Randomized response modeling of sensitive questions
- Complier-average causal effect (CACE) estimation in randomized trials
- Latent class analysis
- Latent class analysis with covariates
- Confirmatory latent class analysis. Twin modeling
- Multilevel and complex survey latent class analysis
- Violations of conditional independence
- Factor mixture modeling, IRT mixture modeling

Longitudinal Modeling With Categorical Latent Variables

- Hidden Markov modeling, latent transition analysis
- Latent class growth analysis
- Growth mixture modeling with latent trajectory classes
- Randomized trials and treatment effects varying across latent trajectory classes
- Latent class growth analysis vs. growth mixture modeling
- Numerical integration, mixtures, and non-parametric representation of factor (random effect) distributions
- 3-level growth mixture modeling
- Discrete- and continuous-time survival modeling with latent variables

Pre-requisites:
Attendees should have completed coursework in categorical data analysis and linear models and should be familiar with multinomial logistic regression and with mixed models. Experience with factor analysis is helpful, but not required.

SC4: STATISTICAL CONSIDERATIONS IN DESIGN AND ANALYSIS OF NON-INFERIORITY CLINICAL TRIALS

(Half Day: 8:00 A.M. – 12:00 P.M.)
Hanover DE (Exhibit Level)

Instructors:
H.M. James Hung and Sue-Jane Wang, U.S. Food and Drug Administration

Description:
Non-inferiority trial designs are increasingly employed for evaluating the efficacy and safety of medical products in recent years. Such designs, particularly without a placebo arm, are often controversial in practice. This short course gives an introduction of the essential design specifications and outlines some fundamental issues in design and analysis of non-inferiority trials. The presentation will focus on the challenges in applications for evaluation of drug products. Topics include objectives of non-inferiority trials, relevant parameters of primary interest, key assumptions and the implications to design and analysis, statistical methods for non-inferiority inference, statistical risks or errors associated with a false assertion, historical evidence on the selected active control, intent-to-treat analysis versus per-protocol analysis, and issues of multiplicity with multiple non-inferiority analyses or superiority testing. Some other emerging issues and challenges will also be discussed. A few typical clinical trial examples will be used for illustrative purposes.

Pre-requisites:
It is assumed that the attendees are familiar with statistical hypothesis testing, type I error and confidence intervals.

Short Courses

SC5: THE DESIGN OF TARGETED CLINICAL TRIALS (Half Day: 1:00 p.m. – 5:00 p.m.) Hanover DE (Exhibit Level)

Instructor:
Richard Simon, National Cancer Institute

Description:
New technology and biological knowledge make it increasingly feasible to predict which patients are most likely to benefit or suffer severe adverse events from a new treatment. Using genomic classifiers to target treatment can greatly improve the therapeutic ratio of benefit to adverse effects. This results in smaller clinical trials, improved likelihood that a treated patient benefits, and economic benefit for society.

Much of the conventional wisdom about how to develop and utilize predictive biomarker classifiers is flawed and does not lead to definitive evidence of treatment benefit for a well defined population. The data used to develop a predictive classifier must be distinct from the data used to test hypotheses about treatment effect in subsets determined by the classifier. Developmental studies are exploratory, but studies on which treatment effectiveness claims are to be based should be definitive studies that test a treatment hypothesis in a patient population completely pre-specified by the classifier. The purpose of a pivotal clinical trial of a new drug that utilizes a genomic classifier is to determine whether the new drug provides benefit to a population of patients defined based on the pre-defined classifier. The purpose is not to refine the classifier or to demonstrate that repeating the classifier development process on independent data results in the same classifier.

Phase III clinical trial designs should provide reliable evidence that a treatment provides benefit for a well defined population of patients. Trial designs are available that will support broad labeling indications in cases where drug activity is sufficient, and provide strong evidence of effectiveness for a prospectively defined subset where appropriate.

Prospectively specified analysis plans for phase III data are essential to achieve reliable results. Biomarker analysis does not mean exploratory analysis except in developmental studies. Biomarker classifiers used in pivotal studies of new drugs should be completely specified based on external data. In some cases, definitive evidence can be achieved from prospective analysis of patients in previously conducted clinical trials with extensive archival of pre-treatment specimens.

This workshop will describe clinical trial designs for utilizing biomarker classifiers in conjunction with new drug development. The results of Simon and Maitournam (1,2) evaluating the efficiency of designs that utilize biomarker classifiers for selecting patients for trial will be described. Randomized designs described by Simon and Wang (3) that do not restrict eligibility but permit evaluating the treatment overall for all randomized patients as well as for one pre-defined biomarker determined subset of patients will be described. The adaptive design of Freidlin and Simon (4) that permits development of a biomarker classifier during the initial phase of a phase III trial and then testing the treatment overall and for the identified subset will also be discussed. More recent extensions of these designs, including extensions involving several approaches to sample size planning, will be described. Reprints as well as interactive software for planning targeted clinical trials is available at *http://linus.nci.nih.gov/brb*.

1. Simon R. and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research 10:6759-63, 2004.
2. Maitnourim A. and Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine 24:329-339, 2005.
3. Simon R., Wang SJ. Use of genomic signatures in therapeutics development. The Pharmacogenomics Journal 6:166-173, 2006.
4. Freidlin B. and Simon R. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clinical Cancer Research 11:7872-8, 2005
5. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology 23:7332-41, 2005.

SC6: APPLIED LONGITUDINAL ANALYSIS (Full Day: 8:30 A.M. - 5:00 P.M.) Hanover FG (Exhibit Level)

Instructor:
Garrett Fitzmaurice, Harvard University

Description:
This course will provide an introduction to statistical methods for analyzing longitudinal data. It will emphasize practical aspects of longitudinal analysis, beginning with a review of established methods for analyzing longitudinal data when the response of interest is continuous. We will present an overview of marginal models and generalized linear mixed models. We will highlight the main distinctions between these types of models and discuss the scientific questions addressed by each. Attendees should have a strong background in linear regression and minimal exposure to generalized linear models (e.g., logistic regression). The course will be based on the textbook, Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). Applied Longitudinal Analysis. John Wiley and Sons. ISBN: 0-471-21487-6.

Pre-requisites:
Course attendees are expected to be familiar with the topic at the level of: Applied Regression Analysis and Multi-variable Methods by David G. Kleinbaum, Lawrence L. Kupper, Keith E. Muller, Azhar Nizam, Duxbury Press OR An Introduction to Generalized Linear Models by Annette J. Dobson, Chapman Hall/CRC Press.

ENAR 2007 Tutorials

T1: INTERMEDIATE BAYESIAN DATA ANALYSIS USING WINBUGS AND BRUGS

Date: Monday, March 12
Time: 8:30 – 10:15 a.m.
Learning Center (Ballroom Level)

Instructor:
Brad Carlin, University of Minnesota

Description:
Most ENAR members have by this time been exposed to Bayesian methods, and have some idea about the hierarchical modeling and other settings in which they offer substantial benefits. But actually obtaining these benefits remains out of reach for many, due to a lack of experience with modern Bayesian software in the analysis of real data. In this tutorial, we will offer a hands-on opportunity to explore the use of WinBUGS, the leading Bayesian software package, in a variety of important models, including (time permitting) regression, ANOVA, logistic regression, nonlinear regression, survival, and multivariate models. Basic elements such as model building, MCMC convergence diagnosis and acceleration, and posterior plotting and summarization will be covered, as well as important data-analytic procedures such as residual analysis, model adequacy (through Bayesian p-values and CPO statistics), variable selection, and model choice (through posterior probabilities and DIC statistics). In addition to WinBUGS, we will also provide a brief introduction to BRugs, the new version of BUGS available directly within the popular R package, which enables simultaneous use of the features of both languages.

Pre-requisites:
Students will be expected to *bring their own laptop computers* to the session, and to have *the latest versions of WinBUGS and R already installed* on their computers. Both of these programs are freely available from *http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml* and *http://www.r-project.org/* respectively. You may wish to arrive in the tutorial room 30 minutes early in order to make sure you have the correct versions of the software and wireless web access. The presentation will assume familiarity with basic Bayesian methods and MCMC algorithms, at the level of, say, Chapters 2 and 5 of Carlin and Louis (2000) or Chapters 2, 3, 5 and 11 of Gelman et al. (2004). The tutorial's goal is to make these methods come alive in the software through real data examples that the students try for themselves during the presentation. All necessary WinBUGS and BRugs code will be made available on the web, and experienced teaching assistants will also be on hand to assist those who become "stuck" for any reason.

T2: STATISTICAL ANALYSIS OF HAPLOTYPE-DISEASE ASSOCIATIONS USING HAPSTAT

Date: Monday, March 12
Time: 3:45 – 5:30 p.m.
Learning Center (Ballroom Level)

Instructor:
Danyu Lin, University of North Carolina at Chapel Hill

Description:
The creation of a catalog of common genetic variants, i.e., single nucleotide polymorphisms (SNPs), by the International HapMap project and the precipitous drop in genotyping costs together have made association studies the preferred approach to understanding the genetic etiology of complex human diseases. Indeed, virtually all epidemiologic studies that are conducted nowadays have genetic components, and genetic information is used increasingly in clinical studies as well. There is no commercial software to perform proper statistical analysis of genetic associations, particularly haplotype-disease associations. Because the HapMap project and current SNP platforms focus on cataloging common SNPs, the use of haplotypes (i.e., arrangements of nucleotides on a single chromosome) tends to provide a more powerful test of genetic association than the use of individual SNPs, especially for rare causative SNPs that are not measured. Inference on haplotype-disease associations is a difficult missing data problem because haplotypes are not directly observed. Although maximum likelihood methods have been developed in recent years to make proper inference about haplotype-disease associations, these methods have rarely been used in practice due the lack of software.

HAPSTAT is a user-friendly software interface for the statistical analysis of haplotype-disease associations. HAPSTAT allows one to estimate or test haplotype effects and haplotype-environment interactions by maximizing the (observed-data) likelihood that properly accounts for haplotype ambiguity. All commonly used study designs, including cross-sectional, case-control, cohort, case-cohort, nested case-control and family studies are included. The phenotype can be a disease indicator, a quantitative trait or a potentially censored time to disease variable. The effects of haplotypes in the phenotype are formulated through flexible regression models, which accommodate a variety of genetic mechanisms. HAPSTAT also provides an efficient way to conduct single SNP analysis. There are always missing genotypes in real studies. The common practice of deleting all of the individuals with missing data can be very inefficient, especially when different SNPs are missing on different individuals. HAPSTAT provides efficient maximum likelihood estimation of individual SNP effects under the assumption of missing at random. Genotype data may be missing by design under case-cohort and nested case-control sampling. HAPSTAT provides efficient maximum likelihood estimation under such designs as well.

The purpose of this tutorial is to introduce HAPSTAT to practicing statisticians who are involved in the analysis of genetic association studies. The tutorial may also be of benefit to statisticians who are interested in methodological research in this area. We will provide an overview of the underlying methodology and describe in detail how to use the software. Several real examples will be provided for illustrations.

Pre-requisites:
All material will be presented at a non-technical level that is accessible to a general statistical audience. Basic understanding of mathematical statistics and regression modeling is required. Knowledge in genetics is helpful, but not required. Attendees are encouraged to download the software at *http://www.bios.unc.edu/~lin/hapstat/* and bring their laptops for a hands-on learning experience.

T3: DRAW YOUR ASSUMPTIONS BEFORE YOUR CONCLUSIONS: GRAPHS FOR CAUSAL INFERENCE

Date: Tuesday, March 13
Time: 8:30 – 10:15 a.m.
Learning Center (Ballroom Level)

Instructor:
Miguel A. Hernán, Harvard University

Description:
Causal directed acyclic graphs (DAGs) can be used to summarize, clarify, and communicate one's qualitative assumptions about the causal structure of a problem. The use of causal DAGs is a natural and simple approach to causal inference from observational data. It is also a rigorous approach that leads to mathematical results that are equivalent to those of counterfactual theory. As a result, causal DAGs are increasingly used in epidemiologic research and teaching. This workshop will provide a non-technical overview of causal DAGs theory, its relation to counterfactual theory, and its applications to causal inference. It will describe how causal DAGs can be used to propose a systematic classification of biases in observational and randomized studies, including the biases induced by the use of conventional statistical methods for the analysis of longitudinal studies with time-varying exposures.

Pre-requisites:
A Master's level knowledge of statistics and general familiarity with linear and non-linear modeling and longitudinal data analyses.

T4: INTRODUCTION TO EPIDEMIC MODELS AND STATISTICAL ANALYSIS OF INFECTIOUS DISEASE DATA

Date: Tuesday, March 13
Time: 1:45 – 3:30 p.m.
Learning Center (Ballroom Level)

Instructor:
M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington

Description:
Analysis of infectious disease data and evaluation of intervention programs, such as vaccination or antiviral agents, pose particular challenges. Infectious agents are transmitted between hosts, so that the dynamics of contacts within the host population, transmission of the infectious agent through the host population, and the assumptions made about those processes affect the analysis and interpretation of infectious disease studies.

In 1916, Sir Ronald Ross wrote about his Theory of Happenings, differentiating events that depended on the number of people already affected, like infectious diseases, from events that were independent, like heart disease or most cancers. Due to these dependent happenings, interventions in infectious diseases, such as vaccination, can have indirect effects in people who do not receive the intervention, as well as in those who do.

This tutorial will cover many of the ideas particular to the analysis of infectious disease data. Topics include the basic reproductive number, transmission probability, epidemic versus endemic infections, thresholds for transmission, latent and incubation periods, serial interval, deterministic and stochastic epidemic models, and real-time evaluation. An overview will be given of the different types of effects of interventions and the requisite study designs and methods of analysis to estimate vaccine efficacy and effectiveness. Advantages of using small transmission units such as households for studies of infectious diseases and interventions will be discussed, as well as a few of the commonly used models and estimation procedures.

Pre-requisites:
A Master's level knowledge of statistics and a healthy curiosity about some things that are not purely statistical are required.

ENAR 2007 Roundtables

**ROUNDTABLE LUNCHEONS –
REGENCY VI (BALLROOM LEVEL)
DATE: MONDAY, MARCH 12, TIME: 12:15 - 1:30 P.M.**

R1: WRITING COLLABORATIVE GRANTS

Discussion Leader:
Tom Ten Have, University of Pennsylvania

Collaborating as a statistician on grant proposals is a two-way street and often entails give and take on a number of fronts. The first give and take focuses on the role of a statistician on a grant proposal. Although some distinguish between staff and faculty statisticians, the role should depend on how involved the statistician is with the development of the grant application and research of the investigator. The second give and take focuses on the rigor of the study design, power analysis, and analysis plan. The statistician is the protector of good science here, but often encounters budgetary limits, alternative habits on the part of scientific fields with respect to these issues, and investigator beliefs and history with these issues. The third give and take is with respect to budget for statistical programming and advice throughout the tenure of the grant. All of the above three components of give-and-take relate to and influence each other. The best approach to all of the above issues is to develop a good and trustful working relationship with the investigator before the grant application process begins.

R2: WRITING STATISTICAL METHODOLOGY GRANTS

Discussion Leader:
Marie Davidian, North Carolina State University

This roundtable luncheon provides biostatisticians who are interested in applying for statistical methodology grants with information about the NIH process for applications and grant review. The review criteria, what constitutes good grant writing skills, and common problems with applications will be discussed.

R3: ENVIRONMENTAL IMPACT ON PUBLIC HEALTH

Discussion Leader:
Linda Young, University of Florida

The potential and real environmental impacts on public health are receiving increasing attention by the media and governments at all levels. The relationship between air quality and asthma and the effect of water quality on the birth weight of babies are but two examples. Through a CDC initiative, several states are beginning to conduct environmental health tracking. Answering the important questions being posed in a timely and yet statistically valid way is challenging. Participants in this round table will discuss what is being done, common statistical challenges, and possible approaches.

R4: OPPORTUNITIES FOR NEW AND EMERGING RESEARCH AREAS IN BIOSTATISTICS

Discussion Leader:
Michael Kosorok, University of North Carolina at Chapel Hill

Novel opportunities for new research areas (e.g., computational biology, functional genomics, proteomics, and imaging) and ways in which these new areas alter the traditional structure of a biostatistician's research activities will be discussed. Is it possible that some biostatisticians are more like basic scientists with wet labs? How might these changes affect the potential nature of industry-academia collaboration? How might new biostatistical research areas impact the theoretical foundations of statistics (eg., from very high dimensional data)?

R5: THE ROLE OF STATISTICIANS IN HEALTH POLICY ISSUES

Discussion Leader:
Eric Feuer, National Cancer Institute

Many people who work with statisticians find that we are often frustratingly non-committal when it comes to taking stands on controversial issues. There seems to be a wide range of opinions among our profession on the role that statisticians should play in health policy issues. Some feel that statisticians are "arbiters" of science, and while it is appropriate to debate methodological issues, they should leave it to others to sort through the evidence to take "sides" on a particular substantive issue. Others feel that perhaps nobody understands the strengths and weaknesses of data supporting each side of an issue better than the statistician who has analyzed/reviewed the data. Participants in this roundtable are welcome to bring examples in which they were faced with a decision of choosing a "side", and how they dealt with the situation.

R6: CONFLICT OF INTEREST

Discussion Leader:
Keith Soper, Merck Research Labs

Recent controversies about conclusions of certain research papers have led some medical journals (including JAMA) to require additional analysis of industry-sponsored studies. In this roundtable we will briefly review current status of the issue, consider whether there is need for balanced changes to increase confidence in the quality of science, and explore what those changes might be.

R7: CLIMATE CHANGE INVESTIGATIONS

Discussion Leader:
Timothy G. Gregoire, Yale University

The purpose of this roundtable discussion is to bring together statisticians with research projects that directly or indirectly deal with climate-change issues. Climate change was the topic of a few sessions at the 2006 JSM, and it seems likely that climate change investigations will remain a hot topic in the years ahead. This roundtable is being held in the hope that it will foster networking opportunities and a greater awareness of climate change activities within the statistical and biometrical community. My own work has dealt with investigations of change to the active layer of vegetation above permafrost on the tundra of the north slope of Alaska.

ENAR 2007 Roundtables

R8: EXPLORING ROADS TO SUCCESSFUL PUBLISHING

Discussion Leader:
Joel Greenhouse, Carnegie Mellon University

Dissemination of research results through publication remains the primary venue for communication in the statistical sciences. In this roundtable, Joel Greenhouse, a co-editor of Statistics in Medicine and past editor of the IMS Lecture Note and Monograph Series, will describe the review and publication process from the editor side and will facilitate a discussion on how to improve your chances of a successful submission. This roundtable will focus on issues of most concern to new researchers.

R9: THE INTERSECTION OF TWO STATISTICAL WORLDS: MEDICAL DEVICES AND PHARMACEUTICAL DRUGS

Discussion Leader:
Greg Campbell, U.S. Food and Drug Administration

There are many similarities and some differences between the statistical world of medical devices and that of pharmaceutical drugs. These two worlds are growing closer together in terms of the statistical issues. This can be seen most dramatically in the consideration of combination products. There are a number of important such examples, including drug-eluting stents; these are bare metal coronary stents on which a particular drug has been coated. Another example of the close cooperation of the two worlds is in the area of pharmacogenomics, in which a particular new drug is targeted based on the outcome of a particular new diagnostic test. One challenge is the design of such studies to demonstrate the efficacy of the drug as well as to evaluate the performance of the diagnostic test. The many challenges of the intersection of these two worlds will be discussed.

R10: CHOOSING AND ADJUSTING SAMPLE SIZE IN DESIGNING A STUDY

Discussion Leader:
Keith Muller, University of Florida

When the threshold for a scientifically important difference can be specified, choosing a sample size reduces to finding appropriate values for nuisance parameters, such as variances. Uncertainty due to bias and random estimates leads to uncertainty about sample size and a desire to adjust it based on early data. Discussion will center on overcoming operational and statistical challenges to study validity due to such uncertainty.

R11: MODEL ASSESSMENT IN LONGITUDINAL DATA ANALYSES

Discussion Leader:
Jane Pendergast, University of Iowa

The need for longitudinal data modeling arises frequently in both designed experiments and observational studies. Classical methods, such as repeated measures ANOVA models and MANOVA models, have been supplemented by work over the past 20 years on generalized linear and nonlinear mixed models, marginal models fit using generalized estimating equations, hierarchical models, etc. While software is readily available to fit a wide variety of models, analysts are faced with issues surrounding how to determine if a model fits the data well and how to choose among competing models. What strategies, approaches, and measures will help? What are their strengths and weaknesses? Where is there need for more research?

R12: IS THERE A FUTURE FOR SURROGATE MARKER EVALUATION IN RANDOMIZED CLINICAL STUDIES?

Discussion Leader:
Geert Molenberghs, Hasselt University

For a number of reasons, surrogate endpoints are considered instead of the so-called true endpoint in clinical studies, especially when such endpoints can be measured earlier, and/or in a less burdensome fashion for patient and experimenter. Surrogate endpoints may occur more frequently than their standard counterparts. For such reasons, it is not surprising that the use of surrogate endpoints in clinical practice is increasing, in spite of early skepticism. Building on the seminal work of Prentice (1989) and Freedman et al. (1992), Buyse et al. (2000) framed the evaluation exercise within a meta-analytic setting, in an effort to overcome difficulties that necessarily surround evaluation efforts based on a single trial. The meta-analytic framework has been extended to non-normal and longitudinal settings, and proposals have been made to unify the somewhat disparate collection of validation measures currently on the market. Implications for design and for predicting the effect of treatment in a new trial, based on the surrogate, have been addressed. Buyse, Molenberghs, and Burzykowski (2005) devoted a monograph to the topic. Is there a future for surrogate marker evaluation and, if so, which steps need to be taken to improve the spread and application of the methodology?

Future Meetings of the International Biometric Society

2008 ENAR Spring Meeting

Crystal City, VA

2009 ENAR Spring Meeting

San Antonio, TX

XXIV International Biometric Conference

Dublin, Ireland

13-18, July 2008

I B S ENAR

Program Summary

SATURDAY, MARCH 10

3:00 p.m.–5:30 p.m. Grand Foyer (Exhibit Level)	**Conference Registration**

SUNDAY, MARCH 11

7:30 a.m.–6:30 p.m. Grand Foyer (Exhibit Level)	**Conference Registration**
8:00 a.m.–12:00 p.m. Hanover DE (Exhibit Level)	**Short Courses** **SC4:** Statistical Considerations in Design and Analysis of Non-Inferiority Clinical Trials
8:30 a.m.–5:00 p.m. Hanover AB (Exhibit Level) Learning Center (Ballroom Level) Hanover C (Exhibit Level) Hanover FG (Exhibit Level)	**Short Courses** **SC1:** Missing Data Methods in Regression Models **SC2:** Semiparametric Theory and Missing Data Analysis **SC3:** Advances in Latent Variable Mixture Modeling Using MPlus **SC6:** Applied Longitudinal Analysis
12:30 p.m.–5:00 p.m. Courtland (ACC Level)	**Fostering Diversity Workshop**
1:00 p.m.–5:00 p.m. Hanover DE (Exhibit Level)	**Short Courses** **SC5:** The Design of Targeted Clinical Trials
4:00 p.m.–6:00 p.m. Grand Foyer (Exhibit Level)	**Exhibits Open**
4:00 p.m.–7:00 p.m. Executive Conference Room 219 (ACC Level)	**ENAR Executive Committee Meeting (Closed)**
4:30 p.m.–6:30 p.m. Chicago Rooms (Exhibit Level)	**Placement Service Opens**
7:30 p.m.–8:00 p.m. Regency VI (Ballroom Level)	**New Member Welcome Reception**
8:00 p.m.–11:00 p.m. Regency VI (Ballroom Level)	**Social Mixer and Poster Presentations**

MONDAY, MARCH 12

7:30 a.m.–8:30 a.m. Regency VI (Ballroom Level)	**Student Breakfast**
7:30 a.m.–5:00 p.m. Grand Foyer (Exhibit Level)	**Conference Registration**
7:30 a.m.–5:00 p.m. Harris Room (ACC Level)	**Speaker Ready Room**
9:00 a.m.–5:00 p.m. Chicago Rooms (Exhibit Level)	**Placement Service**
8:30 a.m.–5:00 p.m. Grand Foyer (Exhibit Level)	**Exhibits Open**

Program Summary

8:30 a.m.–10:15 a.m.	**Tutorial**
Learning Center (Ballroom Level)	**T1: Intermediate Bayesian Data Analysis Using Winbugs and Brugs**
	Scientific Program
Baker (ACC Level)	2. Dynamic Network Models
Regency V (Ballroom Level)	3. Genetical Genomics: Combining Expression and Allelic Variation Data for Complex Diseases
Hanover E (Exhibit Level)	4. Item Response Theory
Hanover AB (Exhibit Level)	5. Multiplicity and Reproducibility in Scientific Studies: Results from a SAMSI Workshop
Hanover F (Exhibit Level)	6. Recent Developments in Bayesian Survival Analysis
Spring (ACC Level)	7. Reconciling Differences in Estimation Methods: A Case Study of Manatee Population Dynamics
Courtland (ACC Level)	8. Contributed Papers: Finding the Best Dose in Clinical Trials
Inman (ACC Level)	9. Contributed Papers: Missing Data and Measurement Error in Epidemiology
Hanover D (Exhibit Level)	10. Contributed Papers: Longitudinal Data, Including Missing Data and Markov Models
Piedmont (ACC Level)	11. Contributed Papers: Genomics: Pathways and Networks
Dunwoody (ACC Level)	12. Contributed Papers: Genome-Wide Association Studies
Hanover C (Exhibit Level)	13. Contributed Papers: Nonparametric Survival Analysis
10:15 a.m.–10:30 a.m.	**Break**
Grand Foyer (Exhibit Level)	
10:30 a.m.–12:15 p.m.	**Scientific Program**
Regency V (Ballroom Level)	14. Innovations in Clinical Trials Design
Hanover D (Exhibit Level)	15. Metabolomics
Courtland (ACC Level)	16. Statistical Methods in HIV Genomics
Dunwoody (ACC Level)	17. New Approaches for Analyzing Functional Data
Hanover AB (Exhibit Level)	18. New Methods for Genetic Association Studies
Hanover F (Exhibit Level)	19. Misclassified or Mismeasured Data in Epidemiology
Hanover E (Exhibit Level)	20. Contributed Papers: Causal Inference
Piedmont (ACC Level)	21. Contributed Papers: Longitudinal Data Applications
Hanover C (Exhibit Level)	22. Contributed Papers: Missing Data
Spring (ACC Level)	23. Contributed Papers: Power Analysis and Sample Size
Baker (ACC Level)	24. Contributed Papers: Multivariate Survival, Including Adjustment for Quality of Life
Inman (ACC Level)	25. Contributed Papers: General Methods I
12:15 p.m.–1:30 p.m.	**Roundtable Luncheons (Registration Required)**
Regency VI (Ballroom Level)	
12:30 p.m.–4:30 p.m.	**Regional Advisory Board (RAB) Luncheon Meeting (By Invitation Only)**
Greenbriar (ACC Level)	
1:45 p.m.–3:30 p.m.	**Scientific Program**
Dunwoody (ACC Level)	26. Functional and Structural Neuro-imaging Data: Modeling and Inference
Hanover AB (Exhibit Level)	27. Methods for Vaccine Trials with Rare Events, Small Sample Sizes, and Missing Data
Regency V (Ballroom Level)	28. Recent Advances in Regression Modeling with Survival Data
Hanover F (Exhibit Level)	29. Statistical Safety Analysis of Time-Dependent Endpoints in Clinical Trials
Spring (ACC Level)	30. Statistics Education in K-12 and Its Potential Impacts
Courtland (ACC Level)	31. Variable Selection for High Dimensional Data
Baker (ACC Level)	32. C. Frederick Mosteller: Biostatistical Scientist – Educator - Mentor
Hanover C (Exhibit Level)	33. Contributed Papers: Clinical Trials
Piedmont (ACC Level)	34. Contributed Papers: Microarray Analysis I
Inman (ACC Level)	35. Contributed Papers: General Methods: Categorical and Survival Data
Hanover E (Exhibit Level)	36. Contributed Papers: Recent Advances in Assessing Agreement
Hanover D (Exhibit Level)	37. Contributed Papers: General Methods II
3:30 p.m.–3:45 p.m.	**Refreshment Break**
Grand Foyer (Exhibit Level)	

Program Summary

3:45 p.m.–5:30 p.m. Learning Center (Ballroom Level)	**Tutorial** **T2: Statistical Analysis of Haplotype-Disease Associations Using Hapstat**
	Scientific Program
Inman (ACC Level)	38. Challenges and Solutions in the Analysis of Observational Data
Courtland (ACC Level)	39. Design and Analysis of Behavioral Intervention Studies
Regency V (Ballroom Level)	40. Integromics
Dunwoody (ACC Level)	41. Interval Censoring in Studies of Infectious Disease
Hanover F (Exhibit Level)	42. Nonparametric Bayes Clustering for Complex Biological Data
Piedmont (ACC Level)	43. Statistical Data Mining for Adverse Drug Event Surveillance
Hanover AB (Exhibit Level)	44. Contributed Papers: Adaptive Design
Baker (ACC Level)	45. Panel on the New FDA Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials
Hanover C (Exhibit Level)	46. Contributed Papers: Classification with High Dimensional Data: Genomics, Proteomics, and Metabolomics
Hanover E (Exhibit Level)	47. Contributed Papers: Imaging
Hanover D (Exhibit Level)	48. Contributed Papers: ROC Analysis
Spring (ACC Level)	49. Contributed Papers: Variable Selection
6:00 p.m.–7:30 p.m. Executive Conference Rooms 219 & 222 (ACC Level)	**President's Reception (By Invitation Only)**

TUESDAY, MARCH 13

7:30 a.m.–5:00 p.m. Grand Foyer (Exhibit Level)	**Conference Registration**
7:30 a.m.–5:00 p.m. Harris Room (ACC Level)	**Speaker Ready Room**
9:00 a.m.–4:00 p.m. Chicago Rooms (Exhibit Level)	**Placement Service**
8:30 a.m.–5:00 p.m. Grand Foyer (Exhibit Level)	**Exhibits Open**
8:30 a.m.–10:15 a.m. Learning Center (Ballroom Level)	**Tutorial**
	T3: Draw Your Assumptions Before Your Conclusions: Graphs for Causal Inference
	Scientific Program
Regency V (Ballroom Level)	50. Adaptive Designs: Where Are We Now, and Where Should We Be Going?
Dunwoody (ACCLevel)	51. Advances in Methodology for the Analysis of Recurrent Event Data
Baker (ACC Level)	52. Computing Your Risk of Disease: Absolute Risk Models
Hanover F (Exhibit Level)	53. Hierarchical Modeling of Large Biomedical Datasets
Hanover AB (Exhibit Level)	54. Semiparametric Regression Methods for Longitudinal Data Analysis
Courtland (ACC Level)	55. Model Selection and Assessment in GEE
Hanover E (Exhibit Level)	56. Contributed Papers: Recurrent Events
Spring (ACC Level)	57. Contributed Papers: Diagnostics I
Inman (ACC Level)	58. Contributed Papers: Health Services Research and Medical Cost
Piedmont (ACC Level)	59. Contributed Papers: Microarray Expression Analysis and Array CGH
Hanover C (Exhibit Level)	60. Contributed Papers: Multiple Testing Procedures, Including Gatekeeping, and FDR
Hanover D (Exhibit Level)	61. Contributed Papers: Spatial/Temporal Methodology and Applications
10:15 a.m.–10:30 a.m. Grand Foyer (Exhibit Level)	**Refreshment Break**
10:30 a.m.–12:15 p.m. Regency VI-VII (Ballroom Level)	**Presidential Invited Address**

Program Summary

12:15 p.m. – 1:30 p.m.	**Lunch on your own**
12:30 p.m.–4:30 p.m. Greenbriar (ACC Level)	**Regional Committee (RECOM) Luncheon Meeting (By Invitation Only)**
1:45 p.m.–3:30 p.m. Learning Center (Ballroom Level)	**Tutorial** **T4: Introduction to Epidemic Models and Statistical Analysis of Infectious Disease Data**
	Scientific Program
Regency V (Ballroom Level)	63. IMS Medallion Lecture
Dunwoody (ACC Level)	64. Covariance Selection and Variance Component Testing in Modeling Longitudinal Data
Piedmont (ACC Level)	65. Diagnostic Medical Imaging
Courtland (ACC Level)	66. New Strategies in Designing Combined-Phase Clinical Trials
Hanover AB (Exhibit Level)	67. Recent Innovations in Dynamic Treatment Regimes
Inman (ACC Level)	68. Statistical Modeling in Ecology
Hanover C (Exhibit Level)	69. Contributed Papers: Microarray Analysis II
Hanover E (Exhibit Level)	70. Contributed Papers: Functional Data Analysis
Baker (ACC Level)	71. Contributed Papers: Latent Variable Applications, Including Structural Equations and Factor Analysis
Hanover F (Exhibit Level)	72. Contributed Papers: Genetic Epidemiology/Statistical Genetics
Hanover D (Exhibit Level)	73. Contributed Papers: Survival Data: Variable Selection and Competing Risks
Spring (ACC Level)	74. Contributed Papers: Bayesian Methods
3:30 p.m.–3:45 p.m. Grand Foyer (Exhibit Level)	**Break**
3:45 p.m.–5:30 p.m.	**Scientific Program**
Courtland (ACC Level)	75. Analysis of Very Large Geostatistical Datasets
Regency V (Ballroom Level)	76. Diagnostics for Mixed Models
Dunwoody (ACC Level)	77. Discovering Structure in Multivariate Data Using Latent Class and Latent Feature Models
Hanover AB (Exhibit Level)	78. Rethinking the FDA (Panel Discussion)
Baker (ACC Level)	79. Statistical Methods for Interpreting and Analyzing Protein Mass-Spectrometry Data
Hanover F (Exhibit Level)	80. Survival Analysis and its Applications in Genetics/Genomics
Piedmont (ACC Level)	81. Contributed Papers: Density Estimation and Empirical Likelihood
Hanover E (Exhibit Level)	82. Contributed Papers: Measurement Error and Surrogate Endpoints
Spring (ACC Level)	83. Contributed Papers: Survival Data: Frailty Models & Cure Rates
Inman (ACC Level)	84. Contributed Papers: Cancer Applications, Including Spatial Cluster Detection
Hanover C (Exhibit Level)	85. Contributed Papers: Variable Selection Methods and Applications
Hanover D (Exhibit Level)	86. Contributed Papers: General Methods and Applications
5:30 p.m.–6:15 p.m. Hanover D (Exhibit Level)	**ENAR Business Meeting (Open to all ENAR Members**)
6:30 p.m.–9:30 p.m.	**Tuesday Night Event – Dinner at the City Grill (Registration Required)** (Shuttle buses will depart from the hotel at 6:15 p.m.)

PROGRAM SUMMARY

WEDNESDAY, MARCH 14

7:30 a.m.–9:00 a.m.	**2008 Spring Meeting Planning Committee Breakfast Meeting (Closed**)
Fairlie Room (ACC Level)	
7:30 a.m.–12:00 noon	**Speaker Ready Room**
Harris Room (ACC Level)	
8:00 a.m.–12:30 p.m.	**Conference Registration**
Grand Foyer (Exhibit Level)	
8:30 a.m.–10:15 a.m.	**Scientific Program**
Hanover C (Exhibit Level)	87. Biosurveillance and Anomaly Detection
Baker (ACC Level)	88. Innovations in Survival Analysis Methodology for Public Health Problems
Hanover AB (Exhibit Level)	89. Instrumental Variable Methods for Causal Inference
Regency V (Ballroom Level)	90. Dose-Finding in Clinical Trials
Hanover E (Exhibit Level)	91. Non-Stationary Time Series Analysis with Applications to Biomedical Data
Regency VI (Ballroom Level)	92. Software Packages for Handling Missing Data (Introductory Lecture Session)
Inman (ACC Level)	93. Contributed Papers: Analysis of Laboratory Experiments
Spring (ACC Level)	94. Contributed Papers: Longitudinal Count Data
Dunwoody (ACC Level)	95. Contributed Papers: Microarray Analysis III
Courtland (ACC Level)	96. Contributed Papers: Quantitative Trait Loci
Piedmont (ACC Level)	97. Contributed Papers: Non- and Semi-Parametrics
Hanover D (Exhibit Level)	98. Contributed Papers: Kinetic and Other Modeling, Including PK/PD
10:15 a.m.–10:30 a.m.	**Break**
Grand Foyer (Exhibit Level)	
10:30 a.m.–12:15 p.m.	**Scientific Program**
Regency VI (Ballroom Level)	99. Error Probabilities, Sample Size, and Multiplicity: Likelihood Methods
Courtland (ACC Level)	100. Group Randomized Trials (Introductory Lecture Session)
Regency V (Ballroom Level)	101. Marginalized Models
Inman (ACC Level)	102. New Methods Using Statistical Graphics as Tools in Addressing Research Questions
Hanover AB (Exhibit Level)	103. Role of Biostatisticians in Policy Issues (Panel Discussion)
Dunwoody (ACC Level)	104. Contributed Papers: Environmental Statistics
Baker (ACC Level)	105. Contributed Papers: Epidemiology Using Bayesian and Empirical Bayes Methods
Piedmont (ACC Level)	106. Contributed Papers: Methods for High Dimensional Data
Hanover C (Exhibit Level)	107. Contributed Papers: Multivariate and Correlated Data
Spring (ACC Level)	108. Contributed Papers: Diagnostics II
Hanover E (Exhibit Level)	109. Contributed Papers: Statistical Models for Genetic Data
Hanover D (Exhibit Level)	110. Contributed Papers: Log-Rank or Other Comparisons of Survival Curves in Independent or Matched Samples

Scientific Program: Poster Presentations

Times may change slightly prior to the meetings. Please check the on-site program for final times. Asterisks (*) indicate paper presenters.

SUNDAY, MARCH 11
8:00–11:00 p.m.
Opening Mixer and Poster Session
Regency VI (Ballroom Level)

1. POSTER PRESENTATIONS

SPONSOR: ENAR

AGREEMENT

1) Selecting an Acceptable Population from K Multivariate Normal Populations
Weixing Cai*, Syracuse University

2) A Generalized Linear Model Based Approach for Estimating Conditional Correlations
Xueya Cai*, Gregory E. Wilding and Alan Hutson, The State University of New York at Buffalo

3) Resampling Dependent Concordance Correlation Coefficients
John M. Williamson and Sara B. Crawford, Centers for Disease Control and Prevention, Hung-Mo Lin*, Penn State College of Medicine

CATEGORICAL DATA

4) A Predictive Model for Binary Paired Outcome with Two-stage Sampling
Jieqiong Bao*, Emory University, Ming Yuan, Georgia Institute of Technology, Jose N.G. Binongo, Andrew Taylor and Amita Manatunga, Emory University

5) Estimation Using the Noncentral Hypergeometric Distribution for Combining 2x2 Tables
Kurex Sidik, Pfizer, Inc., Jeffrey N. Jonkman*, Mississippi State University

6) Hierarchical Bayesian Analysis of Bivariate Binary Data
Ananya Roy*, Malay Ghosh and Ming-Hui Chen, University of Florida

CLINICAL TRIALS

7) Calibration of the Continual Reassessment Method and Implementation in R
Ken Cheung*, Columbia University

8) Challenges in Applying the Intent-to-Treat Principles to a School-based Group Randomized Trial
Kimberly Drews* and Laure El ghormli, The George Washington University Biostatistics Center

9) A Step-down Procedure with Feedback for Eliminating Inferior Treatments in Clinical Trials
Chen-ju Lin*, Georgia Institute of Technology, Anthony J. Hayter, University of Denver

10) Calculating Number Needed to Treat for Continuous Outcomes Using Receiver-operator Characteristic Analyses
David R. Nelson* and Haitao Gao, Eli Lilly & Company

11) Some Considerations for ITT and Treatment Emergent Adverse Event Analyses
Hui Quan*, Qiankun Sun and Ji Zhang, Sanofi-Aventis, Weichung J. Shih, University of Medicine and Dentistry of New Jersey

12) Just Say No to Change from Baseline Analyses
Timothy W. Victor*, Endo Pharmaceuticals, Richard E. White

ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

13) Effect of Air Pollution (PM2.5 & PM10) on Low Birthweight in North Carolina
Simone Gray*, Kerry Williams, Sharon Edwards, Eric Tassone, Geeta Swamy, Alan Gelfand and Marie Lynn Miranda, Duke University

14) Adjusting for Genotyping Error in Non-invasive DNA-based Mark-recapture Population Studies
Shannon M. Knapp*, Bruce A. Craig, Purdue University, Katy Simonsen, Purdue University and Bristol-Myers Squibb

15) Estimating the Number of Species from a Censored Sample
Chang Xuan Mao and Junmei Liu*, University of California-Riverside

16) Archaeological Application of Generalized Additive Models
Yuemei Wang* and Lance A. Waller, Rollins School of Public Health, Emory University, Zev Ross, ZevRoss Spatial Analysis

Scientific Program: Poster Presentations

EPIDEMIOLOGIC METHODS

17) The Dangers of Categorizing BMI in Studies Investigating In-hospital Mortality Following Cardiac Surgery
Giovanni Filardo* Institute for Health Care Research and Improvement, Baylor Research Institute / Southern Methodist University, Cody Hamilton, Institute for Health Care Research and Improvement, Baylor Research Institute, Baron Hamman, Baylor University Medical Center, Hon KT Ng, Southern Methodist University, Paul Grayburn, Baylor University Medical Center

18) Local Multiplicity Adjustments for Spatial Cluster Detection
Ronald E. Gangnon*, University of Wisconsin-Madison

19) Causal Intermediate Effects in HIV Prevention Research
Giovanni Filardo and Cody Hamilton*, Institute for Health Care Research and Improvement

20) Statistical Methods for Associations Between Exposure Profiles and Response
Amy H. Herring*, University of North Carolina at Chapel Hill, David A. Savitz, Mount Sinai School of Medicine

21) Evaluating Spatial Methods for Cluster Detection of Cancer Cases
Lan Huang*, Barnali Das and Linda Pickle, National Cancer Institute

22) Modeling Survival: An Alternative to Proportional Hazards in Alzheimer's Disease
Elizabeth A. Johnson*, Johns Hopkins University, Kathryn Ziegler-Graham, St. Olaf College, Ron Brookmeyer, Johns Hopkins University

23) Structural Equation Modeling of Genotype by Environment Interaction in Coronary Heart Disease
Xiaojuan Mi* and Kent M. Eskridge, University of Nebraska-Lincoln

24) Sensitivity Analysis for the Nested Case Control Data
Kenichi Yoshimura*, National Cancer Center-Japan

GENERAL METHODS

25) Internal Pilot Design with Interim Analysis
John A. Kairalla*, University of North Carolina-Chapel Hill, Keith E. Muller, University of Florida, Christopher S. Coffey, University of Alabama-Birmingham

26) Underreporting in a Zero-Inflated Generalized Poisson Regression Model
Mavis Pararai*, Georgia Southern University

27) A Unified Approach for Computing Power and Sample Size Estimates for the Dorfman-Berbaum-Metz and Obuchowski-Rockette Methods for Analyzing Multireader ROC Studies
Stephen L. Hillis*, VA Iowa City Health Care System

28) New Large-Sample Confidence Intervals for a Binomial Contrast
Joshua M. Tebbs*, University of South Carolina, Scott A. Roths, Penn State University

29) A Nonparametric Mean Estimator for Judgment Post-stratified Data
Xinlei Wang*, Southern Methodist University, Johan Lim, Texas A & M University, Lynne Stokes, Southern Methodist University

30) Distributions for Sums of Exchangeable Bernoulli Random Variables
Chang Yu*, Vanderbilt University Medical Center, Daniel Zelterman, Yale University

GENOMICS

31) Transcription Factor Analysis in Gene Expression Data
William T. Barry*, Fred A. Wright and Mayetri Gupta, University of North Carolina

32) Detecting QTLs by Bayesian Hierarchical Regression Model
Susan J. Simmons and Yi Chen*, The University of North Carolina at Wilmington

33) Maximum Likelihood Factor Analysis when $n < p$
Karen E. Chiswell*, GlaxoSmithKline, John F. Monahan, North Carolina State University

Scientific Program:
Poster Presentations

34) Latent Variable Approach for Meta-analysis of Gene Expression Data from Multiple Microarray Experiments
Hyungwon Choi*, Ronglai Shen, Debashis Ghosh and Arul M. Chinnaiyan, University of Michigan

35) About Dichotomized Continuous Traits in Family-based Association Tests: Do You Really Need Quantitative Traits?
David W. Fardo*, Harvard School of Public Health, Juan C. Celedon, Benjamin A. Raby, and Scott T. Weiss, Channing Laboratory-Brigham and Women's Hospital-Harvard Medical School, Christoph Lange, Harvard School of Public Health-Channing Laboratory-Brigham and Women's Hospital-Harvard Medical School

36) An Association-based Method to Detect Epistasis for Quantitative Traits in Family Data
Guimin Gao*, University of Alabama at Birmingham, Hua Li, Stowers Institute for Medical Research

37) A Mixed Effects Model Implementation of the S-Score Algorithm
Richard E. Kennedy* and Kellie J. Archer, Virginia Commonwealth University

38) Network Neighborhood Analysis With the Multi-Node Topological Overlap Measure
Ai Li* and Steve Horvath, University of California-Los Angeles

39) Domain Enhanced Analysis of Microarray Data Using the Gene Ontology
Jiajun Liu* and Jacqueline M. Hughes-Oliver, North Carolina State University, Alan J. Menius, Glaxo Smith Kline Inc.

40) A Fast Bayesian Method for eQTL Linkage Analysis in Experimental Crosses
Jinze Liu*, Fred Wright, Fei Zou and Yu-Ling Chang, University of North Carolina-Chapel Hill

41) Gene Profiling Data of the Red Light Signaling Pathways in Roots
Xiaobo Li, Richard Lee, Xueli Liu*, Melanie J. Correll & Gary F. Peter, University of Florida,

42) Hierarchical Bayesian Model for QTL Detection
Caroline A. Pearson*, AAIPharma Inc./University of North Carolina at Wilmington

43) Weighted Rank Aggregation of Cluster Validation Measures: A Monte Carlo Cross-entropy Approach
Vasyl Pihur*, Susmita Datta and Somnath Datta, University of Louisville

44) Statistical Issues and Analyses of In Vitro Genomic Data in Order to Identify Clinically Relevant Profiles In Vivo
Laila M. Poisson* and Debashis Ghosh, University of Michigan

45) A Wavelet-based Method for Determining Persistent DNA Copy Number Variation Across Multiple Subjects
William R. Prucka* and Christopher S. Coffey, University of Alabama at Birmingham

46) Statistical Methods for the Discovery of Gene Regulatory Networks
Pingping Qu*, Mayetri Gupta and Joseph G. Ibrahim, University of North Carolina at Chapel Hill

47) A Likelihood Ratio Test of Incomplete Dominance Versus Overdominance and/or Underdominance with Application to Gene Expression Levels
Kai Wang*, University of Iowa

48) Multivariate Correlation Estimator for Replicated Omics Data
Dongxiao Zhu*, Stowers Institute for Medical Research, Youjuan Li, University of Michigan

49) A Hidden Markov Model for Inferring Haplotype Structure from Mouse SNP Data
Jin P. Szatkiewicz*,
Glen L. Beane and Gary A. Churchill, The Jackson Laboratory

HEALTH POLICY APPLICATIONS

50) Implications of Joint Effects within the Family for Treatment Choices
John A. Myers*, University of Louisville

51) Suppression Rules for Unreliable Estimates of Health Indicators
Jennifer D. Parker* and Diane M. Makuc, National Center for Health Statistics/CDC

IMAGING

52) Bayesian Analysis of a Bivariate Autoradiographic Image of Tumors: Local vs. Global Correlation
Timothy D. Johnson*, University of Michigan

53) Examining Modulation of Functional Networks in Multi-Session Multi-Subject fMRI
Rajan S. Patel*, Amgen, Inc.

Scientific Program: Poster Presentations

54) Test-Statistics by Shrinking Variance Components with an Application to fMRI
Shuchih Su* and Brian Caffo, Johns Hopkins Bloomberg School of Public Health, Elizabeth Garrett-Mayer, Johns Hopkins Kimmel Cancer Center, Susan Spear Bassett, Johns Hopkins University

LONGITUDINAL DATA, TIME SERIES, AND FUNCTIONAL DATA ANALYSIS

55) Comparison of Multivariate Strategies for the Analysis of Skewed Psychometric Data with Many Zeros
G.K. Balasubramani* and Stephen R. Wisniewski, Epidemiology Data Center, GSPH, University of Pittsburgh

56) Using a Bivariate Binomial Mixture Model to Estimate Bivariate Component Distributions in a Continuous Mixture Model
Tatiana A. Benaglia* and Thomas Hettmansperger, Pennsylvania State University

57) Factors Affecting Vehicular Lateral Control Measures in Driving Simulators
Jeffrey D. Dawson*, Joshua D. Cosman, Yang Lei, Elizabeth Dastrup, JonDavid Sparks and Matthew Rizzo, University of Iowa

58) High Breakdown Inference for the Constrained Mixed-VARX Model
Mark A. Gamalo*, University of Missouri-Kansas City

59) Multiple Indicator Hidden Markov Model with an Application to Medical Utilization Data
Ran Li*, Melanie Wall and Tianming Gao, University of Minnesota

60) A Nonlinear Latent Class Model for Joint Analysis of Multivariate Longitudinal Data and a Time-to-Event
Cecile Proust-Lima* and Helene Jacqmin-Gadda, University of Michigan

61) An RKHS Formulation of Discrimination and Classification for Stochastic Processes
Hyejin Shin*, Auburn University

MISSING DATA, MEASUREMENT ERROR, AND CAUSAL INFERENCE

62) Propensity Score Subclassification for the Effects of Time-Dependent Treatments: Applications to Diabetes
Constantine Frangakis, Aristide Achy-Brou* and Michael Griswold, Johns Hopkins University

63) Estimation of the Mean Response in Sample Selection Models with Endogenous Covariates
Joseph Gardiner* and Zhehui Luo, Michigan State University

64) A Comparison of Two Procedures for Accommodating Attrition in Substance Abuse Clinical Trials
Sarra L. Hedden*, Robert F. Woolson and Robert J. Malcolm, Medical University of South Carolina

65) Regression Analysis for Group Testing Data with Covariate Measurement Error
Xianzheng Huang* and Joshua M. Tebbs, University of South Carolina

66) Generalized Ridge Regression Models for Estimating a Treatment Effect Using Surrogate Marker Data
Yun Li*, Jeremy M.G. Taylor and Roderick J.A. Little, University of Michigan

67) A Comparison of Methods for Correlation Analysis of Biomarker Data When Some Observations Are Below the Analytic Limit of Detection
Hongwei Wang, Louisiana State University Health Sciences Center, Stephen W. Looney* and Siuli Mukhopadhyay, Medical College of Georgia

68) Estimating the Causal Effect of Race on Stroke Outcome
Megan E. Price*, Vicki S. Hertzberg, Kerrie Krompf and Michael R. Frankel, Emory University

69) Bayesian Approach for Analyzing Cluster Randomized Trial and Adjusting for Misclassification in GLMM
Dianxu Ren*, University of Pittsburgh

70) Bayesian Model Averaging in Latent Variable Models
Benjamin R. Saville* and Amy H. Herring, University of North Carolina at Chapel Hill

Scientific Program:
Poster Presentations

71) Inference and Sensitivity Analysis for the Malaria Attributable Fraction
Dylan Small*, University of Pennsylvania

72) Testing the Mediating Effect in Mediational Models with Multiple Outcomes
Kang Sun*, Sati Mazumdar and Wesley Thompson, University of Pittsburgh, Patricia R. Houck, University of Pittsburgh Medical Center

73) Predicting the Treatment Effect from a Surrogate Marker Using a Potential Outcomes Approach
Jeremy MG Taylor*, Yun Li and Michael Elliott, University of Michigan

74) The Effect of Intravenous Levocarnitine Among Hemodialysis Patients: Marginal Structural Modeling in a Very Large Sample
Eric D. Weinhandl*, David T. Gilbertson and Allan J. Collins, Minneapolis Medical Research Foundation

75) A Comparison of Missing Data Methods for SF36 Quality of Life Data
Liping Zhao*, Paul Kolm and William S. Weintraub, Christiana Care Health System

76) Statistical Analysis of Standardized Uptake Values with Negative PET Scans
Qin Zhou*, Richard Wong, Steven M. Larson and Mithat Gönen, Memorial Sloan-Kettering Cancer Center

SURVIVAL ANALYSIS

77) Evaluation of Recurrent Event Analyses in Pediatric Firearm Victims' Emergency Department Visits
Hyun J. Lim*, University of Saskatchewan-Canada, Marlene Melzer-Lange, Medical College of Wisconsin

78) Comparing Treatments for Twin-Twin Transfusion Syndrome: An Application of Survival Analysis
David M. Shera*, The Children's Hospital of Philadelphia and The University of Pennsylvania, Timothy Crombleholme, Cincinnati Children's Hospital

VARIABLE SUBSET SELECTION AND MODEL SELECTION

79) A Simulation Study of a Model Selection Procedure for Nonlinear Logistic Regression
Scott W. Keith* and David B. Allison, University of Alabama at Birmingham

80) Variable Selection in Clustering via Dirichlet Process Mixture Models
Sinae Kim*, The University of Michigan, Mahlet G. Tadesse, University of Pennsylvania School of Medicine, Marina Vannucci, Texas A&M University

81) Identifying Biomarkers from Mass Spectrometry Data with Ordinal Outcomes
Deukwoo Kwon*, National Cancer Institute, Mahlet G. Tadesse, University of Pennsylvania, Naijun Sha, University of Texas at El Paso, Ruth M. Pfeiffer, National Cancer Institute, Marina Vannucci, Texas A&M University

82) Regressions by Enhanced Leaps-and-Bounds via Additional Optimality Tests (LBOT)
Xuelei (Sherry) Ni*, Kennesaw State University, Xiaoming Huo, Georgia Institute of Technology

83) Variable Selection and Estimation in the Partially Linear AFT Model with High-Dimensional Covariates
Huaming Tan* and Jian Huang, The University of Iowa

Scientific Program

Distinguished Student Award Winner presentations appear in ***boldface italics.***

MONDAY, MARCH 12
8:30 A.M.–10:30 A.M.

2. DYNAMIC NETWORK MODELS Baker (ACC Level)

SPONSOR: IMS
ORGANIZER: ERIC XING, CARNEGIE MELLON UNIVERSITY
CHAIR: ERIC XING, CARNEGIE MELLON UNIVERSITY

8:30 Analyzing Brain Networks with Granger Causality
Mingzhou Ding*, University of Florida
8:55 Understanding Protein Function on a Genome-scale Using Networks
Mark Gerstein*, Yale University
9:20 Graphical Models for Temporal Data
Paola Sebastiani*, Boston University
9:45 Network Biclustering: Identify Condition-specific Network Modules Across Massive Biological Networks
Haiyan Hu, Yu Huang and Xianghong J. Zhou*, University of Southern California
10:10 Floor Discussion

3. GENETICAL GENOMICS: COMBINING EXPRESSION AND ALLELIC VARIATION DATA FOR COMPLEX DISEASES
Regency V (Ballroom)

SPONSOR: ENAR
ORGANIZER: DEBASHIS GHOSH, UNIVERSITY OF MICHIGAN
CHAIR: SINAE KIM, UNIVERSITY OF MICHIGAN

8:30 Dimension Reduction Methods in the Study of the Genetics of Gene Expression
Stephanie A. Monks* and Qiang Guo, Oklahoma State University, Kathy Hanford, University of Nebraska
9:00 Combined Correlation, Linkage, and Enrichment Analysis in eQTL Studies
Christina Kendziorski*, University of Wisconsin
9:30 Using Weighted Gene Co-expression Networks for Integrating Gene Expression, Genetic Marker and Complex Trait Data
Steve Horvath*, University of California-Los Angeles
10:00 Floor Discussion

4. ITEM RESPONSE THEORY
Hanover E (Exhibit Level)

SPONSOR: ENAR
ORGANIZER: JEFFREY S. MORRIS, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER
CHAIR: JEFFREY S. MORRIS, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER

8:30 Nonparametric Bayesian Item Response Theory
Kristin A. Duncan, San Diego State University, Steven N. MacEachern*, The Ohio State University
9:00 Differential Item Functioning in a Graded Response IRT Model: A Bayesian Approach to Item Discrimination
Mark E. Glickman*, Susan Eisen and Pradipta Seal, Boston University
9:30 A Markov Chain Monte Carlo Approach to Confirmatory Item Factor Analysis
Michael C. Edwards*, The Ohio State University
10:00 Floor Discussion

5. MULTIPLICITY AND REPRODUCIBILITY IN SCIENTIFIC STUDIES: RESULTS FROM A SAMSI WORKSHOP
Hanover AB (Exhibit Level)

SPONSORS: ENAR, IMS
ORGANIZERS: JIM BERGER, DUKE UNIVERSITY; SAMSI; PETER MÜLLER, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER
CHAIR: PETER MÜLLER, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER

8:30 Identifying Meaningful Patient Subgroups
Robert L. Obenchain*, Eli Lilly and Company
8:55 Some Aspects of Multiple Testing
Susie Bayarri*, University of Valencia and SAMSI, James Berger, SAMSI and Duke University
9:20 Subgroup Analysis - A Stylized Bayes Approach
Siva Sivaganesan*, University of Cincinnati, Prakash Laud, Medical College of Wisconsin, Peter Müeller, The University of Texas-M.D. Anderson Cancer Center
9:45 Bayesian Decision Theory for Multiplicities
Kenneth M. Rice*, University of Washington
10:10 Floor Discussion

Scientific Program

6. RECENT DEVELOPMENTS IN BAYESIAN SURVIVAL ANALYSIS

Hanover F (Exhibit Level)

SPONSOR: ENAR
ORGANIZER: JOSEPH G. IBRAHIM, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAIR: SUDIPTO BANERJEE, UNIVERSITY OF MINNESOTA

8:30 Prior Elicitation and Variable Selection in Regression Models With High Dimensional Data
Joseph G. Ibrahim* and Mayetri Gupta, University of North Carolina at Chapel Hill

9:00 Joint Modeling of Longitudinal and Survival Data Using Mixtures of Polya Trees
Timothy Hanson, University of Minnesota, Adam Branscum, University of Kentucky, Wesley Johnson*, University of California-Irvine

9:30 Longitudinal Studies with Outcome-Dependent Follow-Up: Models and Bayesian Regression
Bani Mallick*, Texas A&M University, Duchwan Ryu, Chase, Debajyoti Sinha and Stuart Lipsitz, MUSC

10:00 Floor Discussion

7. RECONCILING DIFFERENCES IN ESTIMATION METHODS: A CASE STUDY OF MANATEE POPULATION DYNAMICS

Spring (ACC Level)

SPONSORS: ASA SECTION ON STATISTICS AND THE ENVIRONMENT, ENAR
ORGANIZER: BRUCE A. CRAIG, PURDUE UNIVERSITY
CHAIR: KEVIN GROSS, NORTH CAROLINA STATE UNIVERSITY

8:30 Population Growth Rate of Florida Manatees: Contradictory Estimates from Independent Methods
John E. Reynolds, III*, Mote Marine Laboratory, Michael C. Runge, USGS Patuxent Wildlife Research Center

8:55 Bayesian Trend Analysis of the Florida Manatee Along the Atlantic Coast via Aerial Surveys
Bruce A. Craig*, Purdue University

9:20 Mark-recapture Estimates of Adult Manatee Survival Rate: Modeling Possible Sources of Conflict with Aerial Survey Estimates of Growth
Catherine A. Langtimm*, USGS-Florida Integrated Science Center

9:45 Population Growth Rate of Florida Manatees: Reconciling Estimates from Independent Methods
Michael C. Runge*, USGS Patuxent Wildlife Research Center, Bruce A, Craig, Purdue University, Catherine A. Langtimm, USGS Florida Integrated Science Center, John E. Reynolds, Mote Marine Laboratory

10:10 Floor Discussion

8. CONTRIBUTED PAPERS: FINDING THE BEST DOSE IN CLINICAL TRIALS

Courtland (ACC Level)

SPONSORS: ASA BIOPHARMACEUTICAL SECTION, ENAR
CHAIR: KEN CHEUNG, COLUMBIA UNIVERSITY

8:30 Finding an Acceptable Region in Cancer Dose-Finding Studies Modeling both Toxicity and Efficacy
Bo Huang* and Rick Chappell, University of Wisconsin-Madison

8:45 Assessing the Precision of MED Estimators in Dose-Response Studies
Shanhong Guan*, University of Wisconsin-Madison, Jose C. Pinheiro, Novartis Pharmaceuticals, Frank Bretz, Novartis Pharma AG

9:00 On Identifying Minimum Efficacious Doses in Combination Drug Trials
Julia N. Soulakova*, University of Nebraska-Lincoln, Allan R. Sampson, University of Pittsburgh

9:15 Bayesian Dose-Finding in Clinical Trials for Drug Combinations
Guosheng Yin* and Ying Yuan, The University of Texas-M. D. Anderson Cancer Center

9:30 A Statistical Look at Relative Potency in a Crossover Design
Guoyong Jiang*, Cephalon, Inc.

9:45 Optimization in Multivariate Generalized Linear Models
Siuli Mukhopadhyay*, Medical College of Georgia, Andre I. Khuri, University of Florida

10:00 A Unified Approach to Proof of Activity and Dose Estimation for Binary Data
Bernhard Klingenberg*, Williams College

9. CONTRIBUTED PAPERS: MISSING DATA AND MEASUREMENT ERROR IN EPIDEMIOLOGY

Inman (ACC Level)

SPONSORS: ASA SECTION ON STATISTICS IN EPIDEMIOLOGY, ENAR
CHAIR: HAITAO CHU, THE JOHNS HOPKINS UNIVERSITY

8:30 Estimation of Incubation Periods Distribution under

Different Scenarios of Infectious Disease Outbreak
Xiaojun You* and Ron Brookmeyer, The Johns Hopkins University

8:45 Double Sampling Designs for Addressing Loss to Follow-up in Estimating Mortality
Ming-Wen An* and Constantine E. Frangakis, Th Johns Hopkins University, Donald B. Rubin, Harvard University, Constantin T. Yiannoutsos, Indiana University School of Medicine

9:00 A Combined Analysis of Matched and Unmatched Case-control Studies: a Latent Group Approach
Mulugeta G. Gebregziabher* and Paulo Guimaraes, Medical University of South Carolina

9:15 Non ignorable Missing Data in Matched Case-control Study
Samiran Sinha*, Texas A&M University, Tapabrata Maiti, Iowa State University

9:30 Analysis of a Disease and Probability of Exposure Association Using a Replicated Error-Prone Exposure Assessment
Chengxing Lu* and Robert H. Lyles, Emory University

9:45 Imputation for Missing Continuous Outcomes in Community Intervention Trials
Monica Taljaard*, Ottawa Health Research Institute and University of Ottawa, Allan Donner, Schulich School of Medicine-University of Western Ontario and Robarts Clinical Trials, Neil Klar, Schulich School of Medicine-University of Western Ontario

10:00 Accounting for Error due to Misclassification of Exposures in Case-Control Studies of Gene-Environment Interaction
Li Zhang*, The Cleveland Clinic Foundation, Bhramar Mukherjee, University of Michigan, Malay Ghosh, University of Florida

10. CONTRIBUTED PAPERS: LONGITUDINAL DATA, INCLUDING MISSING DATA AND MARKOV MODELS

Hanover D (Exhibit Level)

SPONSORS: ASA BIOMETRICS SECTION, ENAR
CHAIR: GEERT MOLENBERGHS, HASSELT UNIVERSITY

8:30 Two Local Influence Approaches for Two Binary Outcomes with Non-monotone Missingness
Caroline Beunckens*, Cristina L. Sotto and Geert Molenberghs, Hasselt University

8:45 A Simulation Study Comparing Weighted Estimating Equations with Multiple Imputation Based Estimating Equations
Caroline Beunckens, Cristina L. Sotto*, Geert Molenberghs, Hasselt University

9:00 A Pattern Mixture Model with Least Square Splines in the Analysis of Longitudinal Data with Non-ignorable Dropout: RCT Data from Lidoderm Patch 5% for Pain from Osteoarthritis of the Knee
Qinfang Xiang* and Suna Barlas, Endo Pharmaceuticals

9:15 Non-deletion Approach to Detect Discordant Subjects in Repeated Measurements
Jungwon Mun*, University of Wisconsin-Madison

9:30 Testing for Trends in a Two-state Markov Model with Applications in Smoking Cessation Studies
Charles G. Minard*, The University of Texas M.D. Anderson Cancer Center and The University of Texas – Houston, Wenyaw Chan, The University of Texas – Houston, David Wetter and Carol J. Etzel, The University of Texas M.D. Anderson Cancer Center

9:45 Hidden Markov Models for Alcohol Data
Kenneth E. Shirley* and Dylan Small, The Wharton School, University of Pennsylvania, Kevin Lynch, University of Pennsylvania

10:00 A Bayesian Semiparametric Hidden Markov Model for Incidence Estimation
Alejandro Jara*, María José García-Zattera and Emmanuel Lesaffre, Catholic University of Leuven

11. CONTRIBUTED PAPERS: GENOMICS: PATHWAYS AND NETWORKS

Piedmont (ACC Level)

SPONSOR: ENAR
CHAIR: STEVEN MA, YALE UNIVERSITY

8:30 Incorporating Gene Functional Annotations into Regression Analysis of DNA-protein Binding Data and Gene Expression Data to Construct Transcriptional Networks
Peng Wei* and Wei Pan, University of Minnesota

8:45 Cluster-network Model
Lurdes Y.T. Inoue*, University of Washington, Mauricio Neira, Colleen Nelson and Martin Gleave, Vancouver General Hospital, Ruth Etzioni, Fred Hutchinson Cancer Center

9:00 Bayesian Semiparametric Method for Pathway Analysis
Inyoung Kim*, Herbert Pang and Hongyu Zhao, Yale University

9:15 A Genome-Based Statistical Clustering Method using Dynamic Patterns: Application to Budding Yeast Data Following Glucose-Galactose Shift
Wonsuk Yoo*, Wayne State School of Medicine, Sungchul Ji, Rutgers University

9:30 Joint Modeling of Gene Expression and Motif Profile Data to Infer Transcriptional Modules and Module Specific Activation of Motifs
Roxana A. Alexandridis* and Sunduz Keles, University of Wisconsin-Madison, Rebecka Jornsten, Rutgers University

9:45 ***Variable Selection with Strong Hierarchy Constraints and Its Application to Identification of Gene-Gene and Gene-Environment Interactions***
Nam Hee Choi* and Ji Zhu, University of Michigan

10:00 Incorporating Covariates in Mapping Heterogeneous Traits - A Hierarchical Model Using Empirical Bayes Estimation
Swati Biswas*, University of North Texas Health Science Center, Shili Lin, The Ohio State University

12. CONTRIBUTED PAPERS: GENOME-WIDE ASSOCIATION STUDIES

Dunwoody (ACC Level)

SPONSOR: ENAR
CHAIR: WENBIN LU, NORTH CAROLINA STATE UNIVERSITY

8:30 Genetic Similarity Matching for Genome-wide Association Studies
Weihua Guan*, Liming Liang, Michael Boehnke and Gonçalo R. Abecasis, University of Michigan

8:45 Winner's Curse in Genetic Association Studies
Rui Xiao* and Michael Boehnke, University of Michigan

9:00 A General Population Genetic Model for Haplotyping a Complex Trait in Genetic Association Studies
Min Lin*, Duke University, Rongling Wu, University of Florida

9:15 Application of Bayesian Variable Selection Incorporating Linkage Disequilibrium for Genetic Association Studies
Brooke L. Fridley * and Mariza de Andrade, Mayo Clinic

9:30 A Wald's SPRT-Based Group-Sequential Testing Procedure for Genome-Wide Association Scans
Andres Azuero* and David T. Redden, University of Alabama at Birmingham

9:45 A Bayesian Model to Describe the Association Between Genotype Data and a Phenotype Defined in a Limited Range
Ling Wang*, Vikki Nolan, Clinton T. Baldwin, Martin H. Steinberg and Paola Sebastiani, Boston University School of Public Health

10:00 Simultaneous Confidence Intervals for Odds Ratios in Candidate Gene Studies
Melinda H. McCann* and Stephanie A. Monks, Oklahoma State University

13. CONTRIBUTED PAPERS: NONPARAMETRIC SURVIVAL ANALYSIS

SPONSORS: IMS, ENAR
CHAIR: DANYU LIN, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Dunwoody (ACC Level)

8:30 Nonparametric Estimation of Mean Residual Life Function Using Scale Mixtures
Sujit K. Ghosh* and Shufang Liu, North Carolina State University

8:45 Semiparametric Regression with Time-Dependent Coefficient for Failure Time Data Analysis
Zhangsheng Yu*, The Ohio State University, Xihong Lin, Harvard School of Public Health

9:00 Efficient Estimation in Accelerated Failure Time Model
Donglin Zeng* and Danyu Lin, University of North Carolina

9:15 Exponential Tilt Models in the Presence of Censoring
Chi Wang* and Zhiqiang Tan, The Johns Hopkins University

9:30 Modeling Regression Mean Residual Life Function Using Scale Mixtures
Sujit K. Ghosh and Shufang Liu*, North Carolina State University

9:45 Estimation of Hazard Function under Shape Restrictions
Desale H. Habtzghi*, Georgia College and State University, Mary C. Meyer, University of Georgia, Somnath Datta, University of Louisville

10:00 Nonparametric Inference on Median Residual Life Function
Jong-Hyeon Jeong*, University of Pittsburgh, Sin-Ho Jung, Duke University, Joseph P. Costantino, University of Pittsburgh

MONDAY, MARCH 12
10:15–10:30 A.M.

BREAK Grand Foyer (Exhibit Level)

MONDAY, MARCH 12
10:30 A.M.–12:15 P.M.

14. INNOVATIONS IN CLINICAL TRIALS DESIGN

Regency V (Ballroom Level)

SPONSOR: ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: W. Y. WENDY LOU, UNIVERSITY OF TORONTO
CHAIR: W. Y. WENDY LOU, UNIVERSITY OF TORONTO

10:30 Advances in Clinical Trial Design and Emerging Problems
H.M. James Hung*, U.S. Food and Drug Administration
11:00 Data-driven Interim Analyses with Intent to Cheat
KuangKuo G Lan* and Peter Hu, Johnson & Johnson
11:30 Adaptive Sample Size Based on a Nuisance Parameter: Should We Condition on the Sample Size Actually Used?
Michael A. Proschan* and Martha C. Nason, National Institute of Allergy and Infectious Diseases
12:00 Floor Discussion

15. METABOLOMICS

Hanover D (Exhibit Level)

SPONSOR: IMS
ORGANIZER: JACQUELINE HUGHES-OLIVER, NORTH CAROLINA STATE UNIVERSITY
CHAIR: DAVID BANKS, DUKE UNIVERSITY

10:30 Linking Genetic Profiles to Biological Outcome
S. Stanley Young*, Metabolon, Inc., NISS, Paul Fogel, Consultant, France, Doug Hawkins, University of Minnesota
11:00 Exploring a Complex Metabolomics Data Set
Susan J. Simmons*, University of North Carolina-Wilmington, Emilea Norris, AAI Pharma, Matthew Mitchell, Metabolon
11:30 Pathway-Based Analysis of Metabolic Profiles
Jacqueline M. Hughes-Oliver*, North Carolina State University
12:00 Floor Discussion

16. STATISTICAL METHODS IN HIV GENOMICS

Courtland (ACC Level)

SPONSORS: ENAR, ASA BIOMETRICS SECTION
ORGANIZER: ANDREA S. FOULKES, UNIVERSITY OF MASSACHUSETTS
CHAIR: MICHAEL HUDGENS, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

10:30 Graphical Models for the Accumulation of HIV Drug Resistance Mutations
Niko Beerenwinkel*, Harvard University
10:55 A Resampling-based Approach to Multiple Testing with Uncertainty in Phase
Andrea S. Foulkes*, University of Massachusetts, Victor G. DeGruttola, Harvard School of Public Health
11:20 Non- and Semi-parametric Analysis of Multiple Categorical Predictors and Several Outcomes
A. Gregory DiRienzo*, Harvard University
11:45 Finding Low Dimensional Structure in High Dimensional Data
Hugh Chipman, Acadia University, Andrea Foulkes, University of Massachusetts, Edward George*, University of Pennsylvania, Robert McCulloch, University of Chicago
12:10 Floor Discussion

17. NEW APPROACHES FOR ANALYZING FUNCTIONAL DATA

Dunwoody (ACC Level)

SPONSOR: ENAR
ORGANIZER: VEERA BALADANDAYUTHAPANI, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER
CHAIR: VEERA BALADANDAYUTHAPANI, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER

10:30 Bayesian Method for Curve Classification Using Wavelets
Xiaohui S. Wang*, University of Texas-Pan American, Shubhankar Ray, Merck Research Laboratories, Bani K. Mallick, Texas A&M University
11:00 Functional Data Analysis for Gene Expression Time Courses
Hans-Georg Müller*, University of California-Davis
11:30 A Bayesian Model for Sparse Functional Data
Wesley K. Thompson*, University of Pittsburgh, Ori Rosen, University of Texas-El Paso
12:00 Floor Discussion

18. NEW METHODS FOR GENETIC ASSOCIATION STUDIES

Hanover AB (Exhibit Level)

SPONSOR: ENAR
ORGANIZER: BEVAN HUANG, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAIR: BEVAN HUANG, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

10:30 Statistical Approaches to Whole Genome Association Testing
Andrew Clark*, Cornell University

11:00 Design and Analysis of Genome Wide Association Studies: Application to Type 2 Diabetes
Michael Boehnke*, University of Michigan, Andrew Skol, University of Chicago School of Medicine, Goncalo Abecasis and Laura Scott, University of Michigan

11:30 Likelihood-Based Inference on Haplotype-Disease Associations
Danyu Lin* and Donglin Zeng, University of North Carolina

12:00 Floor Discussion

19. MISCLASSIFIED OR MISMEASURED DATA IN EPIDEMIOLOGY

Hanover F (Exhibit Level)

SPONSOR: ASA SECTION ON STATISTICS IN EPIDEMIOLOGY
ORGANIZER: HAITAO CHU, THE JOHNS HOPKINS UNIVERSITY
CHAIR: PATRICK M. TARWATER, UNIVERSITY OF TEXAS

10:30 Inferring Exposure-disease Relationships when Exposure is Mismeasured: No Adjustment Versus Sensitivity Analysis Versus Bayesian Adjustment
Paul Gustafson*, University of British Columbia

10:55 Multivariate Meta-analysis of Diagnostic Tests without a Gold Standard
Haitao Chu*, Sining Chen, The Johns Hopkins Bloomberg School of Public Health

11:20 Identifiability in the Presence of Misclassification Error
Daniel O. Scharfstein*, Yong Chen and Shuai Wang, The Johns Hopkins Bloomberg School of Public Health

11:45 Classifying Healthcare Associated Infections Using Date of Onset
Justin Lessler* and Ronald Brookmeyer, The Johns Hopkins Bloomberg School of Public Health, Trish M. Perl, Johns Hopkins Hospital

12:10 Floor Discussion

20. CONTRIBUTED PAPERS: CAUSAL INFERENCE

Hanover E (Exhibit Level)

SPONSORS: ASA SECTION ON STATISTICS IN EPIDEMIOLOGY, ENAR
CHAIR: BYRON J. GAJEWSKI, UNIVERSITY OF KANSAS MEDICAL CENTER

10:30 Recursive Path Analysis for Categorical Variables
Haihong Li*, University of Florida

10:45 Multivariate Path Models and the Calculus of Coefficients
Youngju Pak* and Randy L. Carter, The SUNY at Buffalo

11:00 Doubly Robust Estimators for Direct Effects
Sylvie Goetgeluk* and Stijn Vansteelandt, Ghent University, Ghent, Belgium, Els Goetghebeur, Ghent University, Ghent, Belgium and Harvard School of Public Health

11:15 Sensitivity Analysis to Account for a Non-ignorable Missing Covariate in the Estimation of SACE
Brian L. Egleston*, Fox Chase Cancer Center, Daniel O. Scharfstein, The Johns Hopkins University

11:30 Estimating a Causal Treatment Effect on a Mark Variable with Complications of Failure Time Censoring
Jing Ning* and Mei-Cheng Wang, The Johns Hopkins University

11:45 The Sign of the Bias of Unmeasured Confounding
Tyler J. VanderWeele*, University of Chicago, Miguel A. Hernan and James M. Robins, Harvard School of Public Health

12:00 Combining Information From Randomized and Observational Data: A Simulation Study
Eloise E. Kaizar*, The Ohio State University, Howard Seltman and Joel Greenhouse, Carnegie Mellon University

21. CONTRIBUTED PAPERS: LONGITUDINAL DATA APPLICATIONS

Piedmont (ACC Level)

SPONSOR: ENAR
CHAIR: XIN HE, UNIVERSITY OF MISSOURI-COLUMBIA

10:30 Classification Rules for Triply Multivariate Data with an AR(1) Correlation Structure on the Repeated Measures Over Time
Anuradha Roy*, The University of Texas at San Antonio, Ricardo Leiva, F.C.E., Universidad Nacional de Cuyo-Argentina

10:45 Flexible Estimation of Serial Correlation in Linear Mixed Models
Jan Serroyen*, Geert Molenberghs and Marc Aerts, Hasselt University, Belgium, Geert Verbeke, Katholieke Universiteit Leuven, Belgium

11:00 Implementation of a New Correlation Structure in Framework of GEE with R Software
Jichun Xie* and Justine Shults, University of Pennsylvania

11:15 Implementing Semiparametric Varying-Coefficient Partially Linear Models for Longitudinal Data
Jialiang Li* and Yingcun Xia, National University of Singapore, Mari Palta, University of Wisconsin,-Madison

11:30 Modeling Multivariate Latent Trajectories as Predictors of a Univariate Outcome
Sujata M. Patil*, Memorial Sloan-Kettering Cancer Center, Trivellore E. Raghunathan and Jean T. Shope, University of Michigan

11:45 Joint Longitudinal Analysis of Alcohol and Drug Uses for Young Adults
Liang Zhu*, Jianguo Sun and Phillip Wood, University of Missouri-Columbia

12:00 Flexible Modeling of Exposure Data with Informative Number of Repeated Measurements
Huichao Chen*, Amita K. Manatunga, Robert H. Lyles, Limin Peng and Michele Marcus, Emory University

22. CONTRIBUTED PAPERS: MISSING DATA

Hanover C (Exhibit Level)

SPONSOR: ENAR
CHAIR: MICHAEL D. LARSEN, IOWA STATE UNIVERSITY

10:30 Detecting Multiple Sources of Informative Dropout in Clustered Longitudinal Data
Sara B. Crawford* and John J. Hanfelt, Emory University

10:45 Handling Missing Responses in Generalized Linear Mixed Model Without Specifying Missing Mechanism
Hui Zhang* and Myunghee Cho Paik, Columbia University

11:00 Alternatives to Top-coding for Statistical Disclosure Control
Di An* and Roderick J.A. Little, University of Michigan

11:15 Bayesian Analysis of Incomplete Data in Crossover Trials
Sanjib Basu and Sourav Santra*, Northern Illinois University

11:30 Correlation Analysis for Longitudinal Data: Applications to HIV and Psychosocial Research
Yan Ma* and Xin Tu, University of Rochester

11:45 Estimating Mean Cost Under Dependent Censoring
Wenqin Pan*, Duke University, Donglin Zeng, University of North Carolina

12:00 Floor Discussion

23. CONTRIBUTED PAPERS: POWER ANALYSIS AND SAMPLE SIZE

Spring (ACC Level)

SPONSOR: ENAR
CHAIR: CHRIS COFFEY, UNIVERSITY OF ALABAMA AT BIRMINGHAM

10:30 Approximate Confidence Intervals for Power in UNIREP Analyses
Matthew J. Gribbin* and Jacqueline L. Johnson, University of North Carolina, Keith E. Muller, University of Florida

10:45 Sample Size for Tumor Xenograft Studies
Carin J. Kim* and Daniel F. Heitjan, University of Pennsylvania School of Medicine

11:00 Sample Size Calculation for the Wilcoxon-Mann-Whitney Test Adjusting for Ties
Yan Zhao*, Eli Lilly and Company, Dewi Rahardja, University of Indianapolis, Yongming Qu, Eli Lilly and Company

11:15 A Likelihood Approach in Sample Size Calculation
Yong Chen* and Charles Rohde, The Johns Hopkins University

11:30 Designing Longitudinal Studies to Optimize the Number of Subjects and Number of Repeated Measurements
Xavier Basagana* and Donna Spiegelman, Harvard School of Public Health

11:45 A General Approach for Sample Size and Statistical Power Calculations Assessing the Effects of Interventions Using a Mixture Model in the Presence of Detection Limits
Lei Nie*, Georgetown University, Haitao Chu and Stephen Cole, The Johns Hopkins University

12:00 Preemptive Power and the Consulting Statistician
David F. Sawrie*, University of Alabama at Birmingham

24. CONTRIBUTED PAPERS: MULTIVARIATE SURVIVAL, INCLUDING ADJUSTMENT FOR QUALITY OF LIFE

Baker (ACC Level)

SPONSOR: ENAR
CHAIR: ZHANGSHENG YU, THE OHIO STATE UNIVERSITY

10:30 Estimation Methods for Multiple Health-Related Quality-of-Life Adjusted Times-to-Event
Adin-Cristian Andrei*, University of Wisconsin-Madison, Fernando J. Martinez, University of Michigan

10:45 Nonparametric Inference for Paired Quality-of-Life Adjusted Time-to-Event Data
Kristine L. Cooper and Susan Murray*, University of Michigan, Hongwei Zhao, University of Rochester

11:00 On Consistency of Kendall's Tau Under Censoring
David Oakes*, University of Rochester Medical Center

11:15 Efficient Estimation for the Proportional Hazards Model
Lianming Wang*, Biostatistics Branch-National Institute of Environmental Health Sciences, Jianguo Sun, University of Missouri, Xingwei Tong, Beijing Normal University

11:30 On the Association Measure in Copula Models with Application to a HIV Study
Suhong Zhang* and Ying J. Zhang, University of Iowa

11:45 Analyzing Time-to-Event Data for a Composite Endpoint Having a Silent Component
Peng Zhang* and Stephen W. Lagakos, Harvard University School of Public Health

12:00 ***The Proportional Odds Model for Multivariate Interval-censored Failure Time***
Man-Hua Chen*, University of Missouri-Columbia, Xingwei Tong, Beijing Normal University, Jianguo Sun, University of Missouri-Columbia

25. CONTRIBUTED PAPERS: GENERAL METHODS I

Inman (ACC Level)

SPONSOR: ENAR
CHAIR: JASON ROY, UNIVERSITY OF ROCHESTER

10:30 Fiducial Intervals for Variance Components in an Unbalanced Two-component Normal Mixed Linear Model
Lidong E*, Hannig Jan and Hari K. Iyer, Colorado State University

10:45 Multidimensional Array-based Group Testing in the Presence of Test Error
Hae-Young Kim* and Michael G. Hudgens, University of North Carolina at Chapel Hill

11:00 Relativity of Tests' Optimality, with Applications to Change Point Detection and Mixture Type Testing
Albert Vexler*, Chengqing Wu and Kai F. Yu, National Institute of Child Health and Human Development

11:15 Finite Mixture Inference using the Quadratic Inference Function
Daeyoung Kim* and Bruce G. Lindsay, The Penn State University

11:30 Statistical Modeling of Adverse Event Counts in Clinical Trials
Michael A. O'Connell*, Tim Hesterberg and David Henderson, Insightful Corporation

11:45 Some Practical Concerns on Bayesian Sample Size Criteria ACC and ALC
Jing Cao*, Southern Methodist University, Jack J. Lee, M D Anderson Cancer Center

12:00 Problems with Exact Two-sided Tests and the Associated Confidence Intervals for Discrete Distributions
Paul W. Vos and Suzanne S. Hudson*, East Carolina University

MONDAY, MARCH 12
12:15–1:30 P.M.

ROUNDTABLE LUNCHEONS

Regency VI (Ballroom Level)

MONDAY, MARCH 12
1:45–3:30 P.M.

26. FUNCTIONAL AND STRUCTURAL NEURO-IMAGING DATA: MODELING AND INFERENCE

Dunwoody (ACC Level)

SPONSOR: IMS
ORGANIZERS: F. DUBOIS BOWMAN, EMORY UNIVERSITY; THOMAS NICHOLS, UNIVERSITY OF MICHIGAN
CHAIR: LYNN EBERLY, UNIVERSITY OF MINNESOTA

1:45 Bayesian Hierarchical Modeling of Functional Neuroimaging Data
F. DuBois Bowman*, Emory University, Brian Caffo, Johns Hopkins University

2:15 Lead Exposure, Behavior and Neuronal Volume
Brian S. Caffo*, Sining Chen and Brian Schwartz, Johns Hopkins University

2:45 Improving Cluster-wise Inference with Different Types Combined Statistics
Hui Zhang*, University of Michigan, Thomas E. Nichols, GlaxoSmithKline, Timothy D. Johnson, University of Michigan
3:15 Floor Discussion

27. METHODS FOR VACCINE TRIALS WITH RARE EVENTS, SMALL SAMPLE SIZES, AND MISSING DATA

Hanover AB (Exhibit Level)

SPONSOR: ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: CRAIG B. BORKOWF, CENTERS FOR DISEASE CONTROL AND PREVENTION
CHAIR: BRIAN D. PLIKAYTIS, CENTERS FOR DISEASE CONTROL AND PREVENTION

1:45 On Comparing Infecteds in Randomized Vaccine Trials
Dean A. Follmann*, Michael Fay, and Michael Proschan, National Institute of Allergy and Infectious Diseases
2:10 A Prototype Proof-of-Concept Trial Design for Cell Mediated Immunity-Based Vaccines
Devan V. Mehrotra*, Merck Research Laboratories
2:35 Evaluation of Multiple Imputation in an Vaccine Immunogenicity Trial
Michela Baccini, University of Florence, Constantine E. Frangakis, Johns Hopkins University, Fan Li, Harvard Medical School, Fabrizia Mealli, University of Florence, Brian D. Plikaytis and Charles E. Rose, Jr., Centers for Disease Control and Prevention, Donald B. Rubin, Harvard University, Elizabeth R. Zell*, Centers for Disease Control and Prevention
3:00 Modeling Vaccine Adverse Event Count Data Using Zero-Inflated and Hurdle Models
Charles E. Rose, Jr.*, Stacey W. Martin, Kathleen A. Wannemuehler and Brian D. Plikaytis, Centers for Disease Control and Prevention
3:25 Floor Discussion

28. RECENT ADVANCES IN REGRESSION MODELING WITH SURVIVAL DATA

Regency V (Ballroom Level)

SPONSOR: ENAR
ORGANIZERS: LIMIN PENG AND YIJIAN HUANG, EMORY UNIVERSITY
CHAIR: YIJIAN HUANG, EMORY UNIVERSITY

1:45 On General Transformation Model for Censored Data
Zhezhen Jin*, Columbia University
2:10 Survival Analysis with Temporal Covariate Effects
Limin Peng* and Yijian Huang, Rollins School of Public Health-Emory University
2:35 Regression Analysis of Recurrent Episodes Data: the Length-Frequency Tradeoff
Jason P. Fine*, University of Wisconsin-Madison, Jun Yan, University of Iowa
3:00 Analysis of Recurrent-marker Process Data using Model-based Nonparametric and Semiparametric Models
Mei-Cheng Wang *, Bloomberg School of Public Health-Johns Hopkins University
3:25 Floor Discussion

29. STATISTICAL SAFETY ANALYSIS OF TIME-DEPENDENT ENDPOINTS IN CLINICAL TRIALS

Hanover F (Exhibit Level)

SPONSOR: ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: RAFIA BHORE, FOOD AND DRUG ADMINISTRATION
CHAIR: MATT GRIBBIN, UNIVERSITY OF NORTH CAROLINA CHAPEL HILL

1:45 Change-Point Analysis and Survival Data
Ramin B. Arani*, Bristol Myers Squibb Company, Seng-Jaw Soong, Comprehensive Cancer Center-University of Alabama at Birmingham
2:10 Estimating the Location and Confidence Interval of Unknown Change Points in Hazard Rate Models with and without a Binary Grouping Covariate
Sandra L. Gardner*, Sunnybrook Health Sciences Centre, Rafia Bhore, Food and Drug Administration
2:35 Testing for Change-Points in Analyzing Time-to-Event Endpoints
Thomas Hammerstrom, Rafia Bhore* and Mohammad Huque, Food and Drug Administration
3:00 Discussant: Ralph D'Agostino, Boston University

30. STATISTICS EDUCATION IN K-12 AND ITS POTENTIAL IMPACTS

Spring (ACC Level)

SPONSOR: ASA SECTION ON STATISTICAL EDUCATION
ORGANIZER: BONNIE LAFLEUR, VANDERBILT UNIVERSITY
CHAIR: BONNIE LAFLEUR, VANDERBILT UNIVERSITY

1:45 A Framework for Teaching and Learning Statistics in Elementary Grades
Christine A. Franklin*, University of Georgia
2:15 A Framework for Teaching and Learning Statistics in Middle Grades
Gary D. Kader*, Appalachian State University
2:45 An Update on AP Statistics
Linda J. Young*, University of Florida
3:15 Floor Discussion

31. VARIABLE SELECTION FOR HIGH DIMENSIONAL DATA

Courtland (ACC Level)

SPONSORS: ENAR, IMS
ORGANIZER: AMY H. HERRING, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAIR: AMY H. HERRING, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

1:45 Objective Bayes Variable Selection: Some Methods and Some Theory
George Casella*, University of Florida
2:15 Spike and Slab Variable Selection
Hemant Ishwaran*, Cleveland Clinic
2:45 Hierarchical Sparsity Priors in High-Dimensional Variable Selection
Joe Lucas, Carlos Carvalho and Mike West*, Duke University
3:15 Floor Discussion

32. SPECIAL CONTRIBUTED SESSION: C. FREDERICK MOSTELLER: BIOSTATISTICAL SCIENTIST -- EDUCATOR -- MENTOR

Baker (ACC Level)

SPONSORS: ENAR, IMS, ASA
ORGANIZER: DEPARTMENT OF BIOSTATISTICS, HARVARD SCHOOL OF PUBLIC HEALTH
MODERATOR: MARVIN ZELEN, HARVARD UNIVERSITY

1:45 Panel Discussion

33. CONTRIBUTED PAPERS: CLINICAL TRIALS

Hanover C (Exhibit Level)

SPONSORS: ASA BIOPHARMACEUTICAL SECTION, ENAR
CHAIR: AFISI SEGUN ISMAILA, MCMASTER UNIVERSITY

1:45 Estimating the Current Treatment Effect with Historical Control Data
Zhiwei Zhang*, Food and Drug Administration
2:00 The Effects of an Inaccurate Marker on Clinical Trials
Nancy Wang, University of California-Berkeley, Nusrat Rabbee*, Genentech, Inc.
2:15 The Effect of Interim Safety Monitoring on End-of-Trial Estimates of Risk
Michael J. Dallas*, Merck Research Laboratories
2:30 On Treatment Selection in Accelerated Drug Development
Ying Wan*, Temple University
2:45 Analysis of Binary Outcomes in Neonatal Clinical Trials with Twin Births
Michele L. Shaffer* and Allen R. Kunselman, Penn State College of Medicine
3:00 Application of New Levene Type Tests for Hypotheses about Dispersion Differences in Clinical Data Analysis
Xiaoni Liu*, Bristol Myers Squibb Company, Dennis Boos and Cavell Brownie, North Carolina State University
3:15 The Effect of Herd Immunity on Vaccine Trial Design
Blake F. Charvat*, Ronald Brookmeyer and Jay Herson, Johns Hopkins Bloomberg School of Public Health

34. CONTRIBUTED PAPERS: MICROARRAY ANALYSIS I

Piedmont (ACC Level)

SPONSOR: ENAR
CHAIR: JASMINE ZHOU, UNIVERSITY OF SOUTHERN CALIFORNIA

1:45 A Classification Method using the Fiedler Vector, with Application to Micorarray Data
Choongrak Kim*, Minkyung Oh, Soonphil Hong, Eunyeong Yoon and Jinmee Kim, Pusan National University-Pusan, South Korea
2:00 Statistical Methods for Identifying Differentially Expressed Gene Combinations
Yen-Yi Ho* and Leslie Cope, Johns Hopkins University, Marcel Dettling, Zurcher Hochschule-Winterthur, Switzerland, Giovanni Parmigiani, Johns Hopkins University

2:15 ***Proximity Model for Expression Quantitative Trait Loci (eQTL) Detection***
Jonathan A. Gelfond*, Joseph G. Ibrahim and Fei Zou, University of North Carolina

2:30 Sequential Sample Size for Microarrays
Rossell David*, Rice University - MD Anderson Cancer Center, Peter Müller, MD Anderson Cancer Center

2:45 Estimation of Survival Time using cDNA Microarray Data
Jimin Choi*, Minho Kang and Choongrak Kim, Pusan National University

3:00 Exploration, Normalization, and Genotype Calls of High Density Oligonucleotide SNP Array Data
Benilton S. Carvalho*, Johns Hopkins University, Henrik Bengtsson, University of California-Berkeley, Terence P. Speed, WEHI, Melbourne, Australia and University of California-Berkeley, Rafael A. Irizarry, Johns Hopkins University

3:15 Multi-class Cancer Outlier Differential Gene Expression Detection
Fang Liu* and Baolin Wu, University of Minnesota

35. CONTRIBUTED PAPERS: GENERAL METHODS: CATEGORICAL AND SURVIVAL DATA

Inman (ACC Level)

SPONSOR: ENAR
CHAIR: BRODERICK OLUYEDE, GEORGIA SOUTHERN UNIVERSITY

1:45 Impact of Failure to Properly Adjust for Onset of a Time-dependent Exposure
Ayumi K. Shintani*, Patrick G. Arbogast and E. Wesley Ely, Vanderbilt University

2:00 Maximum Likelihood Estimation for Tied Survival Data under Cox's Regression Model via the EM Algorithm
Thomas H. Scheike, University of Copenhagen, Yanqing Sun*, University of North Caroline-Charlotte

2:15 Fractional Imputation Methods for Missing Categorical Data
Michael D. Larsen*, Iowa State University, Shinsoo Kang, KwanDong University, Kenneth J. Koehler, Iowa State University

2:30 'Smooth' Semiparametric Regression Analysis for Arbitrarily Censored Time-to-Event Data
Min Zhang* and Marie Davidian, North Carolina State University

2:45 Algorithmic Prediction of Suicide Attempts
Steven P. Ellis*, Hanga Galfalvy, Maria Oquendo and John Mann, Columbia University

3:00 Exact Likelihood Ratio Trend Test Using Correlated Binary Data
Jonghyeon Kim* and Neal Oden, The EMMES Corporation, Sungyoung Auh, National Institute of Neurological Disorders and Stroke

3:15 Ranges of Measures of Associations for Familial Binary Variables
Yihao Deng*, Indiana University - Purdue University Fort Wayne

36. CONTRIBUTED PAPERS: RECENT ADVANCES IN ASSESSING AGREEMENT

Hanover E (Exhibit Level)

SPONSOR: ENAR
CHAIR: JOHN WILLIAMSON, CENTERS FOR DISEASE CONTROL AND PREVENTION

1:45 A Unified Approach for Assessing Agreement
Lawrence Lin, Baxter Healthcare, Sam Hedayat, University of Illinois at Chicago, Wenting Wu*, Mayo Clinic

2:00 A Class of Repeated Measures Concordance Correlation Coefficients
Tonya S. King* and Vernon M. Chinchilli, Pennsylvania State University College of Medicine, Josep L. Carrasco, University of Barcelona

2:15 A New Approach to Assessing Agreement between Quantitative Measurements using Replicated Observations
Michael Haber*, Emory University. Huiman X. Barnhart, Duke University

2:30 Comparison of Concordance Correlation Coefficient and Coefficient of Individual Agreement in Assessing Agreement
Huiman X. Barnhart*, Duke University, Michael Haber, Emory University, Yuliya Lokhnygina and Andrzej S. Kosinski, Duke University

2:45 A Bayesian Approach for Modeling Censored Method Comparison Data and Assessing Agreement using Tolerance Intervals
Pankaj K. Choudhary*, University of Texas at Dallas, Swati Biswas, University of North Texas Health Science Center

3:00 Comparison of ICC and CCC for Assessing Agreement for Data without and with Replications
Chia-Cheng Chen*, North Carolina State University, Huiman X. Barnhart, Duke University

3:15 Floor Discussion

37. CONTRIBUTED PAPERS: GENERAL METHODS II

Hanover D (Exhibit Level)

SPONSOR: ENAR
CHAIR: HAE-YOUNG KIM, UNIVERSITY OF NORTH CAROLINA-CHAPEL HILL

1:45 Modelling And Inference For An Ordinal Measure Of Stochastic Superiority
Euijung Ryu*, University of Florida

2:00 Random Cluster Size, Within-cluster Resampling and Generalized Estimating Equations
Eugenio Andraca-Carrera*, Bahjat Qaqish and John Preisser, University of North Carolina at Chapel Hill

2:15 A Method for the Significance Analysis of a Treatment Effect Among a Group of Genes and its Application
Taewon Lee*, Robert R. Delongchamp and Varsha G. Desai, National Center for Toxicological Research, Cruz Velasco, Louisiana State University-Health Science Center

2:30 Correlation Estimation from Incomplete Bivariate Samples
Qinying He* and H. N. Nagaraja, The Ohio State University

2:45 Inverse Prediction: A Clinical Application
Jay Mandrekar*, Mayo Clinic

3:00 Latent Trait Models for Development of a Summary Index in Infant Health Care
Xuefeng Liu*, Wayne State University

3:15 Meta-analysis of Trends in Dietary Studies
Michael P. LaValley*, Boston University School of Public Health

MONDAY, MARCH 12
3:30–3:45 P.M.

BREAK

Grand Foyer (Exhibit Level)

MONDAY, MARCH 12
3:45–5:30 P.M.

38. CHALLENGES AND SOLUTIONS IN THE ANALYSIS OF OBSERVATIONAL DATA

Inman (ACC Level)

SPONSOR: ASA SECTION ON STATISTICS IN EPIDEMIOLOGY
ORGANIZER: MARSHALL M. JOFFE, UNIVERSITY OF PENNSYLVANIA
CHAIR: PATRICK HEAGERTY, UNIVERSITY OF WASHINGTON

3:45 Assessing the Safety of Recombinant Human Erythropoietin Using Instrumental Variable Methods
M. Alan Brookhart*, Brigham and Women's Hospital and Harvard Medical School

4:15 An Application of Marginal Structural Models in the Analysis of Anemia Management in Patients on Hemodialysis
David T. Gilbertson* and Eric Weinhandl, Minneapolis Medical Research Foundation

4:45 Novel Approaches to Dealing with Missing Confounder Information in Hemodialysis Management
Marshall M. Joffe*, University of Pennsylvania

5:15 Floor Discussion

39. DESIGN AND ANALYSIS OF BEHAVIORAL INTERVENTION STUDIES

Courtland (ACC Level)

SPONSOR: ASA BIOMETRICS SECTION
ORGANIZER: JOSEPH HOGAN, BROWN UNIVERSITY
CHAIR: JOSEPH HOGAN, BROWN UNIVERSITY

3:45 Inference About Causal Contrasts Between Two Active Treatments in Randomized Trials with Noncompliance: The Effect of Supervised Exercise to Promote Smoking Cessation
Jason Roy*, University of Rochester, Joseph W. Hogan, Brown University

4:10 Post-Randomization Interaction Analyses with Rank Preserving Models
Thomas R. Ten Have*, University of Pennsylvania School of Medicine, Jennifer Faerber, University of Pennsylvania, Marshall M. Joffe, University of Pennsylvania School of Medicine

4:35 The Multi-phase Optimization Strategy: A Novel Way to Develop Multi-component Behavioral Interventions
Bibhas Chakraborty*, University of Michigan, Linda M. Collins, Pennsylvania State University, Susan A. Murphy, Vijayan N. Nair and Victor J. Strecher, University of Michigan

5:00 Heterogeneous Between- and Within-Subjects Variance Models for Longitudinal Behavioral Data
Donald Hedeker* and Robin Mermelstein, University of Illinois at Chicago

5:25 Floor Discussion

40. INTEGROMICS

Regency V (Ballroom Level)

SPONSOR: IMS
ORGANIZER: INGO RUCZINSKI, JOHNS HOPKINS UNIVERSITY
CHAIR: BRIAN CAFFO, JOHNS HOPKINS UNIVERSITY

3:45 Genomic Integration of Copy Number and Expression Data
Debashis Ghosh*, University of Michigan

4:15 A Hidden Markov Model for Combining Estimates of Copy Number and Genotype Calls in High Throughput SNP Arrays
Robert B. Scharpf*, Giovanni Parmigiani, Jonathan Pevsner and Ingo Ruczinski, Johns Hopkins University

4:45 Statistical Methods for Reconstructing Transcriptional Regulatory Networks
Ning Sun, Yale University School of Medicine, Raymond J. Carroll, Texas A&M University, Hongyu Zhao*, Yale University School of Medicine

5:15 Floor Discussion

41. INTERVAL CENSORING IN STUDIES OF INFECTIOUS DISEASE

Dunwoody (ACC Level)

SPONSORS: ENAR, ASA BIOMETRICS SECTION
ORGANIZERS: ZHIGANG ZHANG, OKLAHOMA STATE UNIVERSITY; LEI NIE, GEORGETOWN UNIVERSITY
CHAIR: ZHIGANG ZHANG, OKLAHOMA STATE UNIVERSITY

3:45 Nonparametric Estimation of the Joint Distribution of a Survival Time Subject to Interval Censoring and a Continuous Mark Variable
Michael G. Hudgens*, University of North Carolina-Chapel Hill, Marloes H. Maathuis, University of Washington, Peter B. Gilbert, Fred Hutchinson Cancer Research Center and University of Washington

4:10 Relating a Health Outcome to Subject-Specific Characteristics Based on Left- or Interval-Censored Longitudinal Exposure Data
Robert H. Lyles*, Kathleen A. Wannemeuhler and Amita K. Manatunga, Emory University, Renee H. Moore, University of Pennsylvania, Michele Marcus, Emory University

4:35 Targeted Maximum Likelihood Estimation: Application to Interval Censored Data
Mark van der Laan*, University of California-Berkeley

5:00 Discussant: Somnath Datta, University of Louisville

42. NONPARAMETRIC BAYES CLUSTERING FOR COMPLEX BIOLOGICAL DATA

Hanover F (Exhibit Level)

SPONSORS: ENAR, IMS
ORGANIZER: MAHLET TADESSE, UNIVERSITY OF PENNSYLVANIA
CHAIR: MAHLET TADESSE, UNIVERSITY OF PENNSYLVANIA

3:45 Enriched Stick Breaking Processes for Functional Data
David B. Dunson*, NIEHS/NIH, Bruno Scarpa, University of Padua, Italy

4:15 Nonparametric Bayesian Methods for Genomic Data
Marina Vannucci*, Texas A&M University

4:45 A Hidden Markov Dirichlet Process Model for Joint Inference of Population Structure, Linkage Disequilibrium, and Recombination Hotspots
Eric Xing* and Kyung-Ah Sohn, Carnegie Mellon University

5:15 Floor Discussion

43. STATISTICAL DATA MINING FOR ADVERSE DRUG EVENT SURVEILLANCE

Piedmont (ACC Level)

SPONSORS: ASA SECTION ON STATISTICS IN DEFENSE AND NATIONAL SECURITY, ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: MARTIN KULLDORFF, HARVARD MEDICAL SCHOOL
CHAIR: MARTIN KULLDORFF, HARVARD MEDICAL SCHOOL

3:45 Data Mining the WHO Database of Suspected Adverse Drug Reactions
Andrew Bate*, WHO Collaborating Centre for International Drug Monitoring

4:10 Vaccine Adverse Event Surveillance
Robert L. Davis*, CDC

4:35 Data Mining for Treatment-Comparator Differences in Adverse Event Rates Within Collections of Clinical Trial Results
William DuMouchel* and Robert Zambarano, Lincoln Technologies Division of Phase Forward, Inc.

5:00 Bayesian Logistic Regression for Drug Safety Surveillance
David Madigan*, Rutgers University, Aimin Feng, Novartis, Ivan Zorych, New Jersey Institute of Technology

5:25 Floor Discussion

44. CONTRIBUTED PAPERS: ADAPTIVE DESIGN

Hanover AB (Exhibit Level)

SPONSORS: ASA BIOPHARMACEUTICAL SECTION, ENAR
CHAIR: MATTHEW J. GURKA, UNIVERSITY OF VIRGINIA

3:45 Estimation Following an Adaptive Group Sequential Test
Cyrus R. Mehta*, Cytel Inc.

4:00 Using Weighted Estimates to Increase the Power of a Seamless Phase II/III Trial
Kenneth Klesczewski*, Bristol-Myers Squib

4:15 A Computationally Intensive Approach to Optimizing Bayesian Group Sequential Clinical Trials
J. Kyle Wathen* and Peter F. Thall, University of Texas: M.D. Anderson Cancer Center

4:30 Self-designing Multiple Dose Study via Adaptive Randomization
Lin Wang* and Lu Cui, Sanofi-Aventis

4:45 Doubly Adaptive Biased Coin Design with Heterogeneous Responses
Liangliang Duan* and Feifang Hu, University of Virginia

5:00 A Likelihood Approach For Treatment Schedule Finding Using A Mixture Cure Model With A Sectional Weibull Hazard
Changying A. Liu* and Tom Braun, University of Michigan

5:15 Extracting Information from an On-going Blinded Trial
Jitendra Ganju*, Biostatistical Consulting, Biao Xing, Genentech

45. PANEL ON THE NEW FDA GUIDANCE FOR THE USE OF BAYESIAN STATISTICS IN MEDICAL DEVICE CLINICAL TRIALS

Baker (ACC Level)

SPONSORS: ENAR, ASA SECTION ON BIOPHARMACEUTICAL STATISTICS
CHAIR: LARRY KESSLER, FOOD AND DRUG ADMINISTRATION

Panelists:

Gregory Campbell, Food and Drug Administration
Gene Pennello, Food and Drug Administration
Telba Irony, Food and Drug Administration
Gerry Gray, Food and Drug Administration

46. CONTRIBUTED PAPERS: CLASSIFICATION WITH HIGH DIMENSIONAL DATA: GENOMICS, PROTEOMICS, AND METABOLOMICS

Hanover C (Exhibit Level)

SPONSOR: ENAR
CHAIR: CHOONGRAK KIM, PUSAN NATIONAL UNIVERSITY

3:45 Towards Better Visualization, Identification, and Classification in Metabolomics
Seoung Bum Kim*, University of Texas-Arlington, Dean P. Jones, Emory University

4:00 Locally Adaptive Discriminant Analysis of Proteomics Data
Yuping Wu*, Cleveland State University, R. Webster West, Texas A&M University

4:15 Quantification and Classification of Mass Spectrometry Data by Kernel Methods
Shuo Chen* and Yu Shyr, Vanderbilt Ingram Cancer Center

4:30 Identification of Major Metabolite Features in High-Resolution NMR
Yajun Mei*, Georgia Institute of Technology, Seoung Bum Kim, University of Texas-Arlington, Kwok Leung Tsui, Georgia Institute of Technology

4:45 Detection of Biomarker for Fetal Alcohol Syndrome (FAS) using High Throughput (MALDI-TOF) Proteomics Data
Susmita Datta*, University of Louisville

5:00 Confidence Intervals for Cross-validated Prediction Accuracy Estimates
Kevin K. Dobbin*, National Cancer Institute/NIH

5:15 A Classification Algorithm for the Development of Genomic Signatures from High-Dimensional Data
Songjoon Baek* and Hojin Moon, National Center for Toxicological Research/FDA, Hongshik Ahn, Stony Brook University, Ralph L. Kodell and James J. Chen, National Center for Toxicological Research/FDA

47. CONTRIBUTED PAPERS: IMAGING

Hanover E (Exhibit Level)

SPONSOR: ENAR
CHAIR: NICHOLAS PETRICK, FOOD AND DRUG ADMINISTRATION

3:45 Robust Fitting for Neuroreceptor Mapping
Chung Chang* and Robert Todd Ogden, Columbia University

4:00 Modeling the Spatial and Temporal Dependence in fMRI Data: An Application to an Inhibitory Control Study of Cocaine Addicts
Gordana Derado* and F. D. Bowman, Emory University

4:15 ***Robust Independent Component Analysis for fMRI***
Ping Bai*, Haipeng Shen and Young Truong, University of North Carolina at Chapel Hill

4:30 Modeling Progression of Cerebrovascular Disease with Longitudinal MRI Data via Spatio-Temporal Transition Models
Qian Weng*, Danielle J. Harvey and Laurel A. Beckett, University of California-Davis

4:45 Longitudinal Image Analysis via Stochastic EM Algorithm
Xiaoxi Zhang*, Timothy D. Johnson and Roderick R.A. Little, University of Michigan

5:00 Bayesian Image Reconstruction using Inverse Regularization
Margaret B. Short*, University of Alaska – Fairbanks, David E. Sigeti, Derek Armstrong, Diane Vaughn and David M. Higdon, Los Alamos National Laboratory

5:15 Bayesian Hidden Markov Normal Mixture Model with Application to MRI Tissue Classification
Dai Feng*, The University of Iowa

48. CONTRIBUTED PAPERS: ROC ANALYSIS

Hanover D (Exhibit Level)

SPONSOR: ENAR
CHAIR: JIALIANG LI, NATIONAL UNIVERSITY OF SINGAPORE

3:45 An Alternative Approach to Permutation Tests for Comparing Correlated ROC Curves
Thomas M. Braun*, University of Michigan, Todd A. Alonzo, University of Southern California

4:00 Multi-Reader ROC Methods: Explicit Formulations and Relationships between DBM, Multi-WMW, BWC and Gallas's Methods
Andriy Bandos* and Howard E. Rockette, University of Pittsburgh, Brandon D. Gallas, NIBIB/CDRH, David Gur, University of Pittsburgh

4:15 Strong Approximations for Resample Quantile Processes and Application to ROC Methodology
Jiezhun Gu* and Subhashis Ghosal, North Carolina State University-Raleigh

4:30 Inferences for a Nonparametric Competing Probability
Yongming Qu and Yan D. Zhao*, Eli Lilly and Company, Dewi Rahardja, University of Indianapolis

4:45 Semiparametric Least Squares Analysis of Clustered ROC Curve Data
Liansheng Tang* and Xiaohua A. Zhou, University of Washington

5:00 Comparison of Nonparametric Confidence Intervals for the Area Under the ROC Curve of a Continuous-Scale Diagnostic Test
Gengsheng Qin* and Lejla Hotilova, Georgia State University

5:15 Evaluating Markers for Selecting a Patient's Treatment
Xiao Song*, University of Georgia, Margaret S. Pepe, University of Washington

49. CONTRIBUTED PAPERS: VARIABLE SELECTION

Spring (ACC Level)

SPONSOR: ENAR
CHAIR: ABEL RODRIGUEZ, DUKE UNIVERSITY

3:45 Evaluation and Comparison of Regularization Based Variable Selection Methods
Michael C. Wu*, Tianxi Cai and Xihong Lin, Harvard School of Public Health

4:00 Model Estimation and Selection via Double Penalized Least Squares in Partial Linear Models
Xiao Ni*, Hao Helen Zhang and Daowen Zhang, North Carolina State University

4:15 Sparse Partial Least Squares Regression with an Application to the Genome Scale Transcription Factor Activity Analysis
Hyonho Chun* and Sunduz Keles, University of Wisconsin - Madison

4:30 Model Complexity for the AIC Statistic with Neural Networks
Doug Landsittel* and Dustin Ferris, Duquesne University

4:45 Simultaneous Subset Selection via Rate-distortion Theory, with Applications to Gene Clustering and Significance Analysis of Differential Expression
Rebecka J. Jornsten*, Rutgers University

5:00 ***Fixed and Random Effects Selection in Linear and Logistic Models***
Satkartar K. Kinney*, Duke University, David B. Dunson, National Institute of Environmental Health Sciences

5:15 Ramifications of Pre-processing on Feature Discovery in Microarray Experiments
Kouros Owzar* and Sin-Ho Jung, Duke University Medical Center

TUESDAY, MARCH 13
8:30–10:15 A.M.

50. ADAPTIVE DESIGNS: WHERE ARE WE NOW, AND WHERE SHOULD WE BE GOING?

Regency V (Ballroom Leverl)

SPONSOR: ASA BIOPHARMACEUTICAL SECTION
ORGANIZERS: JEFF MACA AND JOSÉ PINHEIRO, NOVARTIS PHARMACEUTICALS
CHAIR: JOSÉ PINHEIRO, NOVARTIS PHARMACEUTICALS

8:30 Handling Uncertainty About the Likely Treatment Effect: The Roles of Group Sequential Designs and Adaptive Sample Size Re-estimation
Christopher Jennison*, University of Bath-U.K.

9:00 Assessing and Quantifying the Operational Bias of Adaptive Designs
Armin Koch*, BfArM (German Institute for Drugs and Medical Devices)

9:30 Comparison of Statistical Methods and Characteristics of Adaptive Seamless Phase II/III Trials
Jeff D. Maca*, Novartis Pharmacueticals

10:00 Discussant: Robert O'Neill, Food and Drug Administration

51. ADVANCES IN METHODOLOGY FOR THE ANALYSIS OF RECURRENT EVENT DATA

Dunwoody (ACC Level)

SPONSOR: ENAR
ORGANIZER: DOUGLAS SCHAUBEL, UNIVERSITY OF MICHIGAN
CHAIR: DOUGLAS SCHAUBEL, UNIVERSITY OF MICHIGAN

8:30 Estimation of Marginal Characteristics for Recurrent Event Processes
Richard J. Cook* and Jerry F. Lawless, University of Waterloo

8:55 Finite and Asymptotic Properties of Estimators for a General Recurrent Event Model
Edsel A. Peña*, University of South Carolina

9:20 A Semiparametric Nested Frailty Model for Clustered Bivariate Processes, with Applications to Recurrent and Terminal Events
Emmanuel Sharef and Robert L. Strawderman*, Cornell University

9:45 Semiparametric Analysis of Correlated Recurrent and Terminal Events
Yining Ye*, Amgen Inc., Jack D. Kalbfleisch and Douglas E. Schaubel, University of Michigan

10:00 Floor Discussion

52. COMPUTING YOUR RISK OF DISEASE: ABSOLUTE RISK MODELS

Baker (ACC Level)

SPONSOR: ENAR
ORGANIZER: BARRY GRAUBARD, NATIONAL CANCER INSTITUTE
CHAIR: BARRY GRAUBARD, NATIONAL CANCER INSTITUTE

8:30 Lung Cancer Risk Prediction
Peter B. Bach*, Memorial Sloan-Kettering Cancer Center

8:55 Development of a Colorectal Cancer Risk Assessment Tool
Ruth Pfeiffer*, Andrew Freedman and Mitchell Gail, National Cancer Institute, Marty Slattery, University of Utah

9:20 Risk Prediction in the Framingham Heart Study
Lisa M. Sullivan*, Boston University, Michael Pencina, Ramachandrun Vasan, Ralph D'Agostino

9:45 Discussant: Mitch Gail, National Cancer Institute

53. HIERARCHICAL MODELING OF LARGE BIOMEDICAL DATASETS

Hanover F (Exhibit Level)

SPONSOR: ENAR
ORGANIZER: RICHARD F. MACLEHOSE, NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES
CHAIR: RICHARD F. MACLEHOSE, NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES

8:30 Latent Aspects Analysis for Gene Expression Data
Edoardo M. Airoldi*, Princeton University, Stephen E. Fienberg and Eric P. Xing, Carnegie Mellon University

9:00 A Two Stage Method for Fitting Semiparametric Random Effects Models to Large Data Sets
Michael L. Pennell*, The Ohio State University, David B. Dunson, NIEHS

9:30 Computationally Efficient Analysis of Spatially Varying Disease Rates for Large Data Sets
Louise M. Ryan*, Harvard University

10:00 Floor Discussion

54. SEMIPARAMETRIC REGRESSION METHODS FOR LONGITUDINAL DATA ANALYSIS
Hanover AB (Exhibit Level)

SPONSOR: IMS
ORGANIZER: RUNZE LI, PENN STATE UNIVERSITY
CHAIR: RUNZE LI, PENN STATE UNIVERSITY

8:30 Nonparametric/Semiparametric Regression for Incomplete Longitudinal Data Using EM
Xihong Lin*, Harvard School of Public Health, Raymond Carroll, Texas A&M University
8:55 Generalized Semiparametric Linear Mixed Effects Models for Longitudinal Data
Tatiyana V. Apanasovich*, Cornell University, School of ORIE
9:20 Markov Chain Marginal Bootstrap for Longitudinal Data
Di Li and Xuming He*, University of Illinois at Urbana-Champaign
9:45 Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function
Jianqing Fan*, Princeton University, Tao Huang, University of Virginia, Runze Li, Pennsylvania State University
10:00 Floor Discussion

55. MODEL SELECTION AND ASSESSMENT IN GEE
Courtland (ACC Level)

SPONSORS: ASA BIOMETRICS SECTION, ENAR
ORGANIZERS: JOHN PREISSER, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL, RAO CHAGANTY, OLD DOMINION UNIVERSITY
CHAIR: BING LU, MEMORIAL HOSPITAL OF RHODE ISLAND

8:30 Covariance Model Selection in GEE
Vincent J. Carey*, Harvard Medical School, You-Gan Wang, CSIRO, Lin Y. Hin, Private Medical Practitioner
8:55 Some Aspects of Model Assessment when Using GEE
Bahjat F. Qaqish* and John Preisser, University of North Carolina at Chapel Hill
9:20 Consistent Model Selection and Data-driven Smooth Tests for Longitudinal Data in the Estimating Equation Approach
Lan Wang, University of Minnesota, Annie Qu*, Oregon State University
9:45 On the Impact and Likelihood of a Violation of Bounds for the Correlation in GEE Analyses of Binary Data from Longitudinal Trials and What We Can Do to Address This Problem
Justine Shults* and Wenguang Sun, University of Pennsylvania School of Medicine, Xin Tu, University of Rochester, Tom Ten Have, University of Pennsylvania School of Medicine
10:00 Floor Discussion

56. CONTRIBUTED PAPERS: RECURRENT EVENTS
Hanover E (Exhibit Level)

SPONSOR: ENAR
CHAIR: JING QIAN, EMORY UNIVERSITY

8:30 Robust Methods for Analyzing Recurrent Events in Presence of Terminal Events
Rajeshwari Sundaram*, National Institute of Child Health and Human Development, NIH
8:45 Accelerated Failure Time Marginal Means Models for Recurrent Events with a Terminal Event
Xiaoyan Wang* and Jianwen Cai, University of North Carolina at Chapel Hill
9:00 Accelerated Testing with Recurrent Events
Alexander C. McLain* and Edsel A. Peña, University of South Carolina
9:15 Semiparametric Estimation of the Gap-Time Distribution with Recurrent Event Data Under an Informative Monitoring Period
Akim Adekpedjou* and Edsel A. Peña, University of South Carolina
9:30 Goodness of Fit for Composite Hypothesis in the General Recurrent Event Setting
Jonathan T. Quiton* and Edsel A. Peña, University of South Carolina
9:45 Recurrent Competing Risks with Random Monitoring Period
Laura L. Taylor* and Edsel A. Peña, University of South Carolina
10:00 A Maximum Likelihood Approach to Accommodating Censoring in Autoregressive Linear Models Assuming Multivariate Normality with Applications to Modeling
HIV RNA Concentration in Plasma
Caitlin Ravichandran* and Victor DeGruttola, Harvard University

57. CONTRIBUTED PAPERS: DIAGNOSTICS I

Spring (ACC Level)

SPONSOR: ENAR
CHAIR: GENGSHENG QIN, GEORGIA STATE UNIVERSITY

8:30 Two New Measures of Cardiac Dyssynchrony Based on Eigenvalue Decompositions of Correlation Matrices
Peter M. Meyer*, Matthew Weinberg, Richard Miller, Jeffrey Neiger and Steven B. Feinstein, Rush University Medical Center

8:45 Bayesian Estimation of Summary ROC Curve for Meta-analysis of Diagnostic Test Performance
Scott W. Miller*, Debajyoti Sinha, Elizabeth H. Slate,, Donald Garrow and Joseph Romagnuolo, Medical University of South Carolina

9:00 Exploring Parallel and Combined Testing Strategies for Mammography and Other Imaging Tests
Deborah H. Glueck* and Molly M. Lamb, University of Colorado at Denver and Health Sciences Center, John M. Lewin, Diversified Radiology of Colorado, Pisano D. Etta, University of North Carolina at Chapel Hill School of Medicine, Keith E. Muller, University of Florida

9:15 Multivariate Mixtures of Polya Trees for Modeling ROC Data
Tim Hanson, University of Minnesota, Adam Branscum*, University of Kentucky, Ian Gardner, University of California-Davis

9:30 ML Inference on ROC Surfaces in the Presence of Differential Verification
Yueh-Yun Chi* and Xiao-Hua Zhou, University of Washington

9:45 An Exact Test for Detecting Inconsistency in Reader's Interpretation Screening Mammograms
Ji-Hyun Lee* and Steven Eschrich, H. Lee Moffitt Cancer Center & Research Institute, The University of South Florida

10:00 When Sensitivity Depends on Age and Time Spent in Preclinical State in Cancer Screening
Dongfeng Wu*, Mississippi State University, Gary L. Rosner and Lyle D. Broemeling, University of Texas, M.D. Anderson Cancer Center

58. CONTRIBUTED PAPERS: HEALTH SERVICES RESEARCH AND MEDICAL COST

Inman (ACC Level)

SPONSOR: ENAR
CHAIR: LEI LIU, UNIVERSITY OF VIRGINIA

8:30 A Randomization Test of Rater Agreement with Groups of Raters
A. John Bailer* and Robert B. Noble, Miami University

8:45 Analyzing Pressure Ulcer Development of 36 Nursing Homes Using Bayesian Hierarchical Modeling
Jing Zhang*, Zhuoqiong He and David R. Mehr, University of Missouri

9:00 Distribution of Data Envelopment Analysis Efficiency Scores: An Application to Nursing Homes' Care Planning Process
Byron J. Gajewski*, Robert Lee, Marjorie Bott, Ubolrat Piamjariyakul and Roma Lee Taunton, University of Kansas

9:15 Bayesian Sample Size Calculation for Counted Data with Excess Zeroes
Chunyao Feng*, Amgen Inc., James Stamey and John Seaman, Baylor University

9:30 Using Hierarchical Models to Estimate a Weighted Average of Stratum-Specific Parameters
Babette A. Brumback*, Larry H. Winner, George Casella, Allyson Hall and Paul Duncan, University of Florida

9:45 Empirical Likelihood Inference for the Calibration Regression Model with Lifetime Medical Cost
Yichuan Zhao* and Min Lu, Georgia State University

10:00 Adjustment of Multiple Cardiovascular Risk Factors with a Summary Risk Score
Patrick G. Arbogast*, Hua Ding and Wayne A. Ray, Vanderbilt University

59. CONTRIBUTED PAPERS: MICROARRAY EXPRESSION ANALYSIS AND ARRAY CGH

Piedmont (ACC Level)

SPONSOR: ENAR
CHAIR: YUN-LING XU, FOOD AND DRUG ADMINISTRATION

8:30 Hierarchical Mixture Model for Paired Microarray Experiments
Haiyan Wu*, Emory University, Ming Yuan, Georgia Institute of Technology, Rama Akondy, Emory University, M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington

8:45 Variable Selection for Varying Coefficients Models, With Applications to Analysis of Microarray Time Course Gene Expression Data
Lifeng Wang* and Hongzhe Li, University of Pennsylvania School of Medicine

9:00 Are BOEC Cells More Like Large Vessel or Microvascular Endothelial Cells or Neither of Them?
Aixiang Jiang*, Vanderbilt University, Wei Pan, Liming, Milbauer and Robert Hebbel, University of Minnesota

9:15 A Method for CGH Microarray Data Analysis
Wenqing He*, Ian McLeod and Shahidul Islam, The University of Western Ontario

9:30 Practical Issues Associated with the Analysis of Array Comparative Genomic Hybridization Data
Daniel P. Gaile* and Jeffrey C. Miecznikowski, University at Buffalo New York Center of Excellence in Bioinformatics and Life Sciences, Lori Shepherd, New York Center of Excellence in Bioinformatics and Life Sciences, Norma J. Nowak, Roswell Park Cancer Institute New York Center of Excellence in Bioinformatics and Life Sciences

9:45 Analysis of DNA Copy Number Variations Using Penalized Least Absolute
Xiaoli Gao* and Jian Huang, University of Iowa

10:00 ***Smarter Clustering Methods for High-throughput SNP Genotype Calling***
Yan Lin* and George C. Tseng, University of Pittsburgh, Lora J.H. Bean and Stephanie L. Sherman, Emory University, Eleanor Feingold, University of Pittsburgh

60. CONTRIBUTED PAPERS: MULTIPLE TESTING PROCEDURES: GATEKEEPING, AND FDR

Hanover C (Exhibit Level)

SPONSOR: ENAR
CHAIR: KENNETH RICE, UNIVERSITY OF WASHINGTON

8:30 Tree-structured Gatekeeping Procedures for Multiple Endpoints
Ajit C. Tamhane*, Northwestern University, Alex Dmitrienko, Eli Lilly and Company, Brian Wiens, Myogen, Xin Wang, Sanofi-Aventis

8:45 Gatekeeping Testing Procedures Based on the Simes Test with Clinical Trial Applications
Alex Dmitrienko*, Eli Lilly and Company, Ajit C. Tamhane, Northwestern University, Xing Wang, Sanofi-Aventis

9:00 Simultaneous Inference for Multiple Testing and Clustering via a Dirichlet Process Mixture Model
David B. Dahl*, Qianxing Mo and Marina Vannucci, Texas A&M University

9:15 Bonferroni-based Correction Factor for Multiple, Correlated Endpoints
Rickey E. Carter*, Amy E. Herrin and Sharon D. Yeatts, Medical University of South Carolina

9:30 ***Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control***
Wenguang Sun* and Tony Cai, University of Pennsylvania

9:45 Quick Calculation for Sample Size while Controlling False Discovery Rate with Application to Microarray Analysis
Peng Liu*, Iowa State University, J.T. G. Hwang, Cornell University

10:00 On Generalized FDR
Fei Zou*, Fred A. Wright and Jianhua Hu, University of North Carolina-Chapel Hill

61. CONTRIBUTED PAPERS: SPATIAL/TEMPORAL METHODOLOGY AND APPLICATIONS

Hanover D (Exhibit Level)

SPONSORS: ENAR, ASA SECTION ON STATISTICS AND THE ENVIRONMENT
CHAIR: RONALD E. GANGNON, UNIVERSITY. WISCONSIN

8:30 Spatially Adaptive Intrinsic Gaussian Markov Random Fields
Yu Yue*, Paul L. Speckman and Dongchu Sun, University of Missouri at Columbia

8:45 Decomposition of Regression Estimators to Explore the Influence of Unmeasured Confounders in Space and Time
Yun Lu* and Scott L. Zeger, Johns Hopkins University

9:00 ***Hierarchical Multiresolution Approaches for Dense Point-Level Breast Cancer Treatment Data***
Shengde Liang*, Sudipto Banerjee, Bradley P. Carlin, University of Minnesota

9:15 A Disaggregation Approach to Bayesian Spatial Modeling of Categorical Data
Eric C. Tassone*, Alan E. Gelfand and Marie L. Miranda, Duke University

9:30 Smoothing Approaches for Evaluating Ecological Bias: An Application to Racial Disparity in Mortality Rates in the U.S. Medicare Population
Yijie Zhou*, Francesca Dominici and Thomas A. Louis, Johns Hopkins University

9:45 ***Mixtures of Polya Trees for Flexible Spatial Frailty Survival Modeling***
Luping Zhao*, Timothy E. Hanson and Bradley P. Carlin, University of Minnesota

10:00 Spatial Modeling of the Relationship between Air Quality and Myocardial Infarction Adjusted for Socio-Demographic Status
Jie Yang* and Linda J. Young, University of Florida, Carol A.G. Crawford, National Center for Environmental Health, CDC, Greg Kearney and Chris Duclos, Division of Environmental Health, Florida Department of Health

TUESDAY, MARCH 13
10:15–10:30 A.M.

BREAK

Grand Foyer (Exhibit Level)

TUESDAY, MARCH 13
10:30 A.M.–12:15 P.M.

PRESIDENTIAL INVITED ADDRESS

SPONSOR: ENAR
ORGANIZER/CHAIR: LISA LAVANGE, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Regency VI-VII (Ballroom Level)

10:30 Introduction: Lisa LaVange, The University of North Carolina at Chapel Hill
10:35 Distinguished Student Paper Awards: Peter Imrey, Cleveland Clinic Foundation
10:45 Success, Change, and the Future: The Evolving Role of a Statistician in Research and Development
Frank Rockhold, GlaxoSmithKline

TUESDAY, MARCH 13
12:15–1:30 P.M.

LUNCH ON YOUR OWN

TUESDAY, MARCH 13
1:45–3:30 P.M.

63. IMS MEDALLION LECTURE

Regency V (Ballroom Level)

SPONSOR: IMS
ORGANIZER: ANDREW NOBEL, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAIR: DAVID BANKS, DUKE UNIVERSITY

1:45 Prediction by Supervised Principal Components
Robert Tibshirani, Stanford University

64. COVARIANCE SELECTION AND VARIANCE COMPONENT TESTING IN MODELING LONGITUDINAL DATA

Dunwoody (ACC Level)

SPONSOR: ENAR
ORGANIZER: BO CAI, UNIVERSITY OF SOUTH CAROLINA
CHAIR: BO CAI, UNIVERSITY OF SOUTH CAROLINA

1:45 Bayesian Variable Selection Under Various Model Settings
Mahlet G. Tadesse*, University of Pennsylvania, Sinae Kim, University of Michigan, Naijun Sha, University of Texas El Paso, Marina Vannucci, Texas A&M University
2:15 Likelihood Ratio Tests for Zero Variance in Linear Mixed Models
Ciprian M. Crainiceanu*, Johns Hopkins University
2:45 Regularized Estimation of Covariance Matrices for Longitudinal Data Through Shrinkage and Smoothing
Jianhua Huang*, Texas A&M University, Linxu Liu, Columbia University
3:15 Floor Discussion

65. DIAGNOSTIC MEDICAL IMAGING

Piedmont (ACC Level)

SPONSORS: ENAR, ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: GENE PENNELLO, U.S. FOOD AND DRUG ADMINISTRATION
CHAIR: GENE PENNELLO, U.S. FOOD AND DRUG ADMINISTRATION

1:45 Thresholds on Test Positivity and Reference Standard: Effects on Test Predictive Value and Accuracy
Constantine Gatsonis*, Brown University, Shang Ying Shiu, Brown University and Academia Sinica
2:10 Evaluating the Potential of Novel Phase II Imaging Endpoints
Lori E. Dodd*, National Cancer Institute
2:35 Assessing Computer-aided Diagnosis Algorithms: Algorithm Architecture, Statistical Tools and Study Designs
Nicholas Petrick*, Brandon D. Gallas, Frank W. Samuelson and Sophie Paquerault, U.S. Food and Drug Administration
3:00 Overview of Contrast Agents and Specific Statistical Challenges for Medical Imaging
Kohkan Shamsi, Symbiotic Pharma Research, Suming W. Chang*, Berlex Inc., Joerg Kaufmann, Schering AG
3:25 Floor Discussion

66. NEW STRATEGIES IN DESIGNING COMBINED-PHASE CLINICAL TRIALS

Courtland (ACC Level)

SPONSOR: ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: YU SHEN, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER
CHAIR: YU SHEN, THE UNIVERSITY OF TEXAS M.D. ANDERSON CANCER CENTER

1:45 A Bayesian Approach to Jointly Modeling Toxicity and Biomarker Expression in a Phase I/II Dose-Finding Trial
B. Nebiyou Bekele* and Yu Shen, The University of Texas M. D. Anderson Cancer Center

2:10 Practical Recommendations for Phase I/II Designs
Rick Chappell*, University of Wisconsin-Madison, Ken YK Cheung, Columbia University

2:35 Seamless Phase II & III: Smaller, Stronger, Quicker
Scott M. Berry*, Berry Consultants, Donald A. Berry, M.D. Anderson Cancer Center

3:00 A Bayesian Decision-theoretic Approach Based Adaptive Designs for Clinical Trials
Yi Cheng*, Indiana University-South Bend, Yu Shen, M.D. Anderson Cancer Center

3:25 Floor Discussion

67. RECENT INNOVATIONS IN DYNAMIC TREATMENT REGIMES

Hanover AB (Exhibit Level)

SPONSORS: ENAR, ASA SECTION ON BIOPHARMACEUTICAL STATISTICS
ORGANIZER: DANIEL SCHARFSTEIN, JOHNS HOPKINS UNIVERSITY
CHAIR: DANIEL SCHARFSTEIN, JOHNS HOPKINS UNIVERSITY

1:45 An Introduction to Dynamic Treatment Regimes
Marie Davidian*, North Carolina State University

2:15 Estimating Mean Response as a Function of Treatment Duration in an Observational Study, Where Duration May Be Informatively Censored
Anastasios A. Tsiatis*, North Carolina State University

2:45 Hypothesis Testing & Dynamic Treatment Regimes
Susan Murphy*, Lacey Gunter and Bibhas Chakraborty, University of Michigan

3:15 Floor Discussion

68. STATISTICAL MODELING IN ECOLOGY

Inman (ACC Level)

SPONSOR: ASA SECTION ON STATISTICS AND THE ENVIRONMENT
ORGANIZER: JUN ZHU, UNIVERSITY OF WISCONSIN-MADISON
CHAIR: JUN ZHU, UNIVERSITY OF WISCONSIN-MADISON

1:45 Modelling Non-Linear Mixture Experiments in Stream Ecology
Mary C. Christman*, University of Florida, Christopher Swan, University of Maryland

2:10 Estimating Rare Butterfly Abundance by Combining Multiple Data Types with Simple Population Models
Kevin Gross* and Nick M. Haddad, North Carolina State University, Brian R. Hudgens, Institute for Wildlife Studies

2:35 Invasions, Epidemics, and Binary Data in a Cellular World
Mevin B. Hooten*, Utah State University, Christopher K. Wikle, University of Missouri

3:00 Hierarchical Modeling and Inference in Metacommunity Systems
Jeffrey A. Royle*, USGS Patuxent Wildlife Research Center

3:25 Floor Discussion

69. CONTRIBUTED PAPERS: MICROARRAY ANALYSIS II

Hanover C (Exhibit Level)

SPONSOR: ENAR
CHAIR: AIXIANG JIANG, VANDERBILT UNIVERSITY

1:45 Evaluating The Position Effect Of Single-Base-Pair Mismatch Probes In Affymetrix Genechip®
Fenghai Duan*, University of Nebraska Medical Center

2:00 Enhanced Quantile Approach for Assessing Differential Gene Expression
Huixia Wang*, North Carolina State University, Xuming He, University of Illinois at Urbana-Champaign

2:15 Meta-Analysis for Microarray Experiments
Shiju Zhang* and Grier P. Page, University of Alabama at Birmingham

2:30 Minimax Estimation of Means with Application to Microarray Experiments
Tiejun Tong*, Yale University, Liang Chen, University of Southern California, Hongyu Zhao, Yale University

2:45 Integrating Quantitative Information from ChIP-chip Experiments into Motif Finding
Heejung Shim* and Sunduz Keles, University of Wisconsin-Madison

3:00 Bayesian Change Point Analysis of Genetic Instability in High-throughput Experiments
David L. Gold* and Choa Sima, Texas A&M University, Daniel P. Gaile, The State University of New York, Norma Nowak, Roswell Park Cancer Institute, Bani Mallick, Texas A&M University

3:15 Dynamic Network Analysis of Time Course Microarray Experiments
Donatello Telesca* and Lurdes YT Inoue, University of Washington

70. CONTRIBUTED PAPERS: FUNCTIONAL DATA ANALYSIS

Hanover E (Exhibit Level)

SPONSORS: ENAR, IMS
CHAIR: LI QIN, FRED HUTCHINSON CANCER RESEARCH CENTER

1:45 CARDS: Classification After Reduction of Dimension with SCAD Penalty
Woncheol Jang*, Jeonyoun Ahn and Cheolwoo Park, University of Georgia

2:00 Regularized Functional Linear Models
Qi Long*, Emory University, Ming Yuan, Georgia Institute of Technology

2:15 Multiscale Analysis on fMRI Data
Cheolwoo Park*, University of Georgia

2:30 Diagnostics for Nonlinear Dynamics
Giles Hooker*, Cornell University

2:45 Smoothing Parameter Selection in Penalized Signal Regression
Philip T. Reiss*, New York University

3:00 Nonparametric Estimation for Relating Dose Distributions to Outcome Measures
Matthew J Schipper* and Jeremy MG Taylor, University of Michigan

3:15 Estimating Coefficients and Linearly Independent Solutions of a Linear Differential Operator with Covariates for Functional Data
Seoweon Jin* and Joan G. Staniswalis, The University of Texas at El Paso

71. CONTRIBUTED PAPERS: LATENT VARIABLE APPLICATIONS, INCLUDING STRUCTURAL EQUATIONS AND FACTOR ANALYSIS

Baker (ACC Level)

SPONSOR: ENAR
CHAIR: LIMIN PENG, EMORY UNIVERSITY

1:45 Semiparametric Relationship among Latent Variables in the Structural Equation Models
Bo Ma* and Andrew B. Lawson, University of South Carolina

2:00 Residual-Based Diagnostics for Structural Equation Models
Brisa N. Sanchez*, University of Michigan, E. Andres Houseman and Louise M. Ryan, Harvard School of Public Health

2:15 Application of Covariance Shrinkage to Factor Analysis
Sock-Cheng Lewin-Koh* and Nicholas J. Lewin-Koh, Eli Lilly and Company

2:30 L^2-Based Homogeneity Tests for Mixtures with Structural Parameters
Hongying Dai*, Columbus State University, Richard Charnigo, University of Kentucky

2:45 Supervised Bayesian Latent Class Models for High-Dimensional Data
Stacia M. DeSantis*, E. Andres Houseman, Brent A. Coull and Rebecca A. Betensky, Harvard University

3:00 ***Bayesian Modeling of Embryonic Growth using Latent Variables***
James C. Slaughter* and Amy H. Herring, University of North Carolina at Chapel Hill, Katherine E. Hartmann, Vanderbilt University Medical Center

3:15 Bayesian Multivariate Growth Curve Latent Class Models for Mixed Outcomes
Benjamin E. Leiby*, Thomas Jefferson University, Mary D.Sammel,, Thomas R. Ten Have and Kevin G. Lynch, University of Pennsylvania

72. CONTRIBUTED PAPERS: GENETIC EPIDEMIOLOGY/ STATISTICAL GENETICS

Hanover F (Exhibit Level)

SPONSORS: ENAR, ASA SECTION ON STATISTICS IN EPIDEMIOLOGY
CHAIR: RUTH PFEIFFER, NATIONAL CANCER INSTITUTE

1:45 ***Efficient Association Mapping of Quantitative Trait Loci with Selective Genotyping***
Bevan E. Huang* and Danyu Lin, University of North Carolina at Chapel Hill

2:00 A General Quantitative Genetic Model for Constructing the Network of Haplotype-haplotype Interactions in Genetic Association Studies
Song Wu*, Jie Yang and Rongling Wu, University of Florida

2:15 Estimating a Multivariate Familial Correlation using Joint Models for Canonical Correlations
Hye-Seung Lee*, University of South Florida, Myunghee Cho Paik and Joseph H. Lee, Columbia University

2:30 Performance of Statistical Procedures for the Detection of Interlocus Interactions in Genome-wide Association Studies: Statistical Power and Type I Error Rates
Solomon K. Musani*, Amit Patki, and Hemant K. Tiwari, University of Alabama at Birmingham

2:45 Using Duplicate Genotyped Data in Genetic Analyses: Testing Association and Estimating Error Rates
Nathan L. Tintle*, Hope College, Stephen J. Finch, Stony Brook University, Derek Gordon, Rutgers University

3:00 Tagging SNP Selection with Supervised Support Vector Score Test and Recursive Feature Addition Algorithm in Case-Control Association Studies
Yulan Liang* and Arpad G. Kelemen, University at Buffalo, Qingzhong Liu, New Mexico Institute of Mining and Technology

3:15 Identification of Differentially Expressed Gene Categories in Microarray Studies Using Multivariate Nonparametric Analysis
Dan Nettleton*, Iowa State University, Justin Recknor, Eli Lilly and Company, James M. Reecy, Iowa State University

73. CONTRIBUTED PAPERS: SURVIVAL DATA: VARIABLE SELECTION AND COMPETING RISKS

Hanover D (Exhibit Level)

SPONSOR: ENAR
CHAIR: BIN WANG, UNIVERSITY OF SOUTH ALABAMA

1:45 Validation of Clinical Prognostic Models for Suicide Attempt
Hanga Galfalvy*, Maria A. Oquendo and John J. Mann, Columbia University

2:00 ***Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates***
Sijian Wang* ***(Van Ryzin Award Winner)***, Bin Nan, Ji Zhu and David Beer, University of Michigan

2:15 Robust Regression for Censored Response for High Dimensional Covariates
Huiliang Xie* and Jian Huang, University of Iowa

2:30 Combining Multiple Biomarker Models in Survival Analysis
Zheng Yuan* and Debashis Ghosh, University of Michigan

2:45 Selection of Experiments for Weighted Distributions
Broderick O. Oluyede*, Georgia Southern University

3:00 Semiparametric Analysis of Mixture Regression Models with Competing Risks Data
Wenbin Lu*, North Carolina State University, Limin Peng, Emory University

3:15 Cumulative Incidence Function under the Semiparametric Additive Risk Model
Seunggeun Hyun*, National Institute of Child Health and Human Development, Yanqing Sun, University of North Carolina at Charlotte, Rajeshwari Sundaram, National Institute of Child Health and Human Development

74. CONTRIBUTED PAPERS: BAYESIAN METHODS

Spring (ACC Level)

SPONSOR: ENAR
CHAIR: JU-HYUN PARK, UNIVERSITY OF NORTH CAROLINA-CHAPEL HILL

1:45 Collapsed Stick-Breaking Processes for Semiparametric Bayes Hierarchical Models
Mingan Yang* and David B. Dunson, National Institute of Environmental Health Sciences, NIH

2:00 Constructing Dependent Polya Trees with Copulas
Song Zhang* and Peter Müller, University of Texas M.D. Anderson Cancer Center

2:15 ***The Nested Dirichlet Process***
Abel Rodriguez*, Duke University, David B. Dunson, National Institute of Environmental Health Sciences, NIH, Alan E. Gelfand, Duke University

2:30 Smoothing ANOVA for General Designs
Yue Cui* and James S. Hodges, University of Minnesota

2:45 Bayesian Dynamic Latent Class Models for Multivariate Longitudinal Categorical Data
Bo Cai*, University of South Carolina, David B. Dunson, National Institute of Environmental Health Sciences, NIH, Joseph B. Stanford, University of Utah

3:00 Bayesian Methods for Highly Correlated Data
Richard F. MacLehose* and David B. Dunson, National Institute of Environmental Health Sciences, NIH, Amy H. Herring, University of North Carolina, Jane A. Hoppin, National Institute of Environmental Health Sciences

3:15 Bayesian Hierarchical Models for the Two-Component Memory Process
Xiaoyan Lin* and Dongchu Sun, University of Missouri-Columbia

TUESDAY, MARCH 13
3:30–3:45 P.M.

BREAK

Grand Foyer (Exhibit Level)

TUESDAY, MARCH 13
3:45–5:30 P.M.

75. ANALYSIS OF VERY LARGE GEOSTATISTICAL DATASETS

Courtland (ACC Level)

SPONSOR: ASA SECTION ON STATISTICS AND THE ENVIRONMENT
ORGANIZER: BRAD CARLIN, UNIVERSITY OF MINNESOTA
CHAIR: BRAD CARLIN, UNIVERSITY OF MINNESOTA

3:45 Gaussian Predictive Process Models for Large Spatial Datasets
Sudipto Banerjee*, University of Minnesota, Alan E. Gelfand, Duke University, Andrew O. Finley, University of Minnesota, Huiyang Sang, Duke University

4:15 Approximate Likelihood for Large Irregularly Spaced Spatial Data
Montserrat Fuentes*, North Carolina State University

4:45 Gaussian Process Models for High Dimensional Spaces
Dave Higdon* and Brian J. Williams, Los Alamos National Laboratory

5:15 Discussant: Alan Gelfand, Duke University

76. DIAGNOSTICS FOR MIXED MODELS

Regency V (Ballroom Level)

SPONSORS: ASA BIOMETRICS SECTION, ENAR
ORGANIZER: GEERT VERBEKE, CATHOLIC UNIVERSITY OF LEUVEN
CHAIR: GEERT VERBEKE, CATHOLIC UNIVERSITY OF LEUVEN

3:45 The Linear Model has Three Basic Types of Residual
John Haslett*, Trinity College, Dublin, Ireland, Steve J. Haslett, Massey University, Palmerston North-New Zealand

4:10 Extending the Box-Cox Transformation to the Linear Mixed Model
Matthew J. Gurka*, University of Virginia, Lloyd J. Edwards, University of North Carolina at Chapel Hill, Keith E. Muller, University of Florida, Lawrence L. Kupper, University of North Carolina at Chapel Hill

4:35 Residual Diagnostics for Latent Variable Mixture Models
Chen-Pin Wang*, University of Texas Health Science Center at San Antonio, C. H. Brown, University of South Florida, Karen Bandeen-Roche, Johns Hopkins University

5:00 Formal and Informal Model Selection and Assessment when Data are Incomplete
Geert Molenberghs*, Hasselt University-Diepenbeek, Belgium

5:25 Floor Discussion

77. DISCOVERING STRUCTURE IN MULTIVARIATE DATA USING LATENT CLASS AND LATENT FEATURE MODELS

Dunwoody (ACC Level)

SPONSOR: ENAR
ORGANIZER: DAVID B. DUNSON, NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES, NIH
CHAIR: DAVID B. DUNSON, NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES, NIH

3:45 Latent Class Measurement of Health States Lacking a Gold Standard
Karen Bandeen-Roche*, Jeannie-Marie Sheppard and Jing Ning, Johns Hopkins Bloomberg School of Public Health

4:15 Functional Data Analytic Approach To Processing Mass Spectrometry Data With An Application to SELDI-TOF MS analysis
Jaroslaw Harezlak* and Xihong Lin, Harvard School of Public Health, Shan Jiang, Tsinghua University-Beijing

4:45 Modeling Multiple Latent Classes
Melanie M. Wall*, University of Minnesota

5:15 Floor Discussion

78. PANEL DISCUSSION: RETHINKING THE FDA

Hanover AB (Exhibit Level)

SPONSORS: IMS, ENAR, ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: DAVID BANKS, DUKE UNIVERSITY
CHAIR: LISA LAVANGE, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Panelists: Susan Ellenberg, University of Pennsylvania
Gary Koch, The University of North Carolina at Chapel Hill
Steve Snapinn, Amgen
Dalene Stangl, Duke University

79. STATISTICAL METHODS FOR INTERPRETING AND ANALYZING PROTEIN MASS-SPECTROMETRY DATA

Baker (ACC Level)

SPONSORS: ENAR, ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: JEANETTE ECKEL-PASSOW, MAYO CLINIC
CHAIR: ANN OBERG, MAYO CLINIC

3:45 Introduction to Mass Spectrometry Based Proteomics
Christopher J. Mason*, Mayo Clinic College of Medicine
4:15 Spot Detection and Quantification for 2-D Proteomic Data
Jeffrey S. Morris*, The University of Texas M. D. Anderson Cancer Center
4:45 An Application of Bi-linear Models to a Proteomics Problem
Terry M. Therneau*, Mayo Clinic
5:15 Floor Discussion

80. SURVIVAL ANALYSIS AND ITS APPLICATIONS IN GENETICS/GENOMICS

Hanover F (Exhibit Level)

SPONSOR: ENAR
ORGANIZER: YI LI, HARVARD UNIVERSITY
CHAIR: YI LI, HARVARD UNIVERSITY

3:45 Modeling Association of Bivariate Competing Risks
Yu Cheng*, University of Pittsburgh, Jason P. Fine, University of Wisconsin-Madison
4:10 Semiparametric Variance-Component Models for Linkage and Association Analysis of Censored Trait Data
Guoqing Diao*, George Mason University, Danyu Lin, University of North Carolina at Chapel Hill
4:35 Multivariate Survival Analysis for Case-Control Family Data
Li Hsu*, Fred Hutchinson Cancer Research Center
5:00 Regularized Estimation in Pathway-Based Censored Data Regression Modeling of Genomic Data
Hongzhe Li* and Zhi Wei, University of Pennsylvania
5:25 Floor Discussion

81. CONTRIBUTED PAPERS: DENSITY ESTIMATION AND EMPIRICAL LIKELIHOOD

Piedmont (ACC Level)

SPONSOR: ENAR
CHAIR: PENG ZHANG, HARVARD SCHOOL OF PUBLIC HEALTH

3:45 Problems Related to Efficacy Measurement and Analysis
Sibabrata Banerjee* and Sunil Dhar, New Jersey Institute of Technology
4:00 Nonparametric Bayes Structural Equation Models for Multivariate Data
Ju-Hyun Park* and David B. Dunson, National Institute of Environmental Health Sciences
4:15 Estimating the Slope in Linear Regression Models Using Kernel Densities
Thomas Jaki*, Cleveland State University
4:30 Empirical Likelihood Inference in Presence of Nuisance Parameters
Mi-Ok Kim*, Cincinnati Children's Hospital Medical Center
4:45 Bayesian Smoothing of Density Estimation via Hazard Rates
Luyan Dai* and Dongchu Sun, University of Missouri-Columbia
5:00 Density Estimation in Informative Censoring
Bin Wang*, University of South Alabama
5:15 Diagnostic Measures for Empirical Likelihood of General Estimating Equations
Hongtu Zhu*, University of North Carolina at Chapel Hill, Niansheng Tang, Yunnan University, Kunming People's Republic of China, Joseph G. Ibrahim, University of North Carolina at Chapel Hill, Heping Zhang, Yale University School of Medicine

82. MEASUREMENT ERROR AND SURROGATE ENDPOINTS

Hanover E (Exhibit Level)

SPONSOR: ENAR
CHAIR: CIPRIAN CRAINICEANU, JOHNS HOPKINS UNIVERSITY

3:45 Estimation of Generalized Partially Linear Models with Measurement Error Using Sufficiency Scores
Lian Liu*, Texas A&M University
4:00 Restricted Maximum Likelihood Estimation of a Measurement Error Model via the Monte Carlo EM Algorithm: ARE-Analysis of Enzyme Activity Data
Antara Majumdar* and Randy L. Carter, University at Buffalo

4:15 Flexible Evaluation of Mixed Continuous-Binary Surrogate Endpoints Using Information Theory and Simplified Modeling Strategy Based on Meta-Analytic Approach
Pryseley N. Assam*, Abel Tilahun, Ariel Alonso and Geert Molenberghs, Hasselt University

4:30 Flexible Surrogate Marker Evaluation from Several Randomized Clinical Trials with Binary Endpoints Using SAS
Abel Tilahun*, Pryseley N. Assam, Ariel Alonso and Geert Molenberghs, Hasselt University

4:45 On Some Aspects of Covariates' Models with Measurement Errors in a Completely Randomized Design
Karabi Sinha*, University of Illinois at Chicago

5:00 Analysis of Tissue Microarray Data using Measurement Error Models
Ronglai Shen*, Debashis Ghosh and Jeremy MG Taylor, University of Michigan

5:15 Improving Postural Instability Onset Time Measure for Parkinson's Disease
Peng Huang*, Medical University of South Carolina, Ming Hui Chen, University of Connecticut, Debajyoti Sinha, Medical University of South Carolina

83. CONTRIBUTED PAPERS: SURVIVAL DATA: FRAILTY MODELS & CURE RATES

Spring (ACC Level)

SPONSOR: ENAR
CHAIR: JAYAWANT MANDREKAR, MAYO CLINIC

3:45 Gauss Quadrature Estimation in Frailty Proportional Hazards Models
Lei Liu*, University of Virginia

4:00 Frailty Model with Spline Estimated Baseline Hazard
Pang Du*, Virginia Tech

4:15 Bayesian Case Influence Diagnostics for Survival Data
Hyunsoon Cho* and Joseph G. Ibrahim, University of North Carolina at Chapel Hill, Debajyoti Sinha, Medical University of South Carolina, Hongtu Zhu, University of North Carolina at Chapel Hill

4:30 Shared Frailty Models for Jointly Modelling Grouped and Continuous Survival Data
Denise A. Esserman*, University of North Carolina, Andrea B. Troxel, University of Pennsylvania School of Medicine

4:45 A New Latent Cure Rate Marker Model for Survival Data
Sungduk Kim, Yingmei Xi* and Ming-Hui Chen, University of Connecticut

5:00 A New Threshold Regression Model for Survival Data with a Cure Fraction
Sungduk Kim*, Ming-Hui Chen and Dipak K. Dey, University of Connecticut

5:15 ***Analysis of Smoking Cessation Patterns Using a Stochastic Mixed Effects Model with a Latent Cured State***
Sheng Luo*, Ciprian M. Crainiceanu and Thomas A. Louis, Johns Hopkins University, Nilanjan Chatterjee, National Cancer Institute, National Institutes of Health

84. CONTRIBUTED PAPERS: CANCER APPLICATIONS, INCLUDING SPATIAL CLUSTER DETECTION

Inman (ACC Level)

SPONSORS: ASA BIOMETRICS SECTION, ENAR
CHAIR: ANURADHA ROY, UNIVERSITY OF TEXAS AT SAN ANTONIO

3:45 Introducing Spatial Correlation in the Spatial Scan Statistics
Zhengyuan Zhu*, University of North Carolina at Chapel Hill, Ji Meng Loh, Columbia University

4:00 Evaluating Spatial Methods for Cluster Detection of Cancer Cases
Lan Huang*, Barnali Das and Linda Pickle, National Cancer Institute

4:15 Radon Leukemia Study: A Hierarchical Population Risk Model for Spatially Correlated Exposure Measured with Error
Brian J. Smith*, The University of Iowa, Lixun Zhang, Yale University, William R. Field, The University of Iowa

4:30 A Weighted Kaplan-Meier Approach for Estimation of Recurrence of Colorectal Adenomas
Chiu-Hsieh Hsu*, University of Arizona, Jeremy Taylor, University of Michigan, Qi Long, Emory University, David Alberts, University of Arizona, Patricia Thompson, University of Michigan

4:45 Natural History Model of Metastatic Progression Applied to Lung Cancer
Maksim A. Pashkevich*, Bronislava M. Sigal and Sylvia K. Plevritis, Stanford University

5:00 'Smooth' Regression Analysis of Arbitrarily Censored Cluster-correlated Time-to-event Data
Lihua Tang* and Marie Davidian, North Carolina State University

5:15 Floor Discussion

85. CONTRIBUTED PAPERS: VARIABLE SELECTION METHODS AND APPLICATIONS

Hanover C (Exhibit Level)

SPONSOR: ENAR
CHAIR: DAVID M. SHERA, UNIVERSITY OF PENNSYLVANIA

3:45 Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis
Hongmei Yang*, Daowen Zhang and Hao Helen Zhang, North Carolina State University

4:00 Bridge Logistic Regression Based on ROC Criterion
Guoliang Tian*, Zhenqiu Liu, Hongbin Fang and Ming Tan, University of Maryland Greenebaum Cancer Center

4:15 Variable Selection using Random Forests
Andrejus Parfionovas* and Adele Cutler, Utah State University

4:30 Model Selection for Multivariate Smoothing Splines with Correlated Random Errors
Eren Demirhan* and Hao Helen Zhang, North Carolina State University

4:45 Nonparametric Bayes Local Regression and Variable Selection
Yeonseung Chung*, University of North Carolina at Chapel Hill, David B. Dunson, NIEHS

5:00 Variance Component Selection for Multilevel Partially-reduced Dose-response Curves in Cell-culture Bioassay
Carrie G. Wager*, Lansky Consulting

5:15 Correction for Model Selection Bias Using a Modified Model Averaging Approach for Supervised Learning Method Applied to EEG Experiments
Kristien Wouters*, José Cortiñas and Geert Molenberghs, Universiteit Hasselt-Belgium, Abdellah Ahnaou, Wilhelmus Drinkenburg and Luc Bijnens, Johnson & Johnson Pharmaceutical Research and Development-Belgium

86. CONTRIBUTED PAPERS: GENERAL METHODS AND APPLICATIONS

Hanover D (Exhibit Level)

SPONSOR: ENAR
CHAIR: FREDA COONER, FOOD AND DRUG ADMINISTRATION

3:45 Utilizing Unscheduled Reports of Disease in Estimating Rates: A Conceptual Framework for Use in Evaluating Health of Workers in the World Trade Center Cleanup
Sylvan Wallenstein* and Carol A. Bodian, The Mount Sinai School of Medicine, Jeanne M. Stellman, Mailman School of Public Health, Columbia University and The Mount Sinai School of Medicine

4:00 Calibrated Spatial Stratified Estimator in Spatial Population
Jong-Seok Byun*, Hanshin University-Korea, Chang-Kyoon Son, Korea Institute for Health and Social Affairs, Jong-Min Kim, University of Minnesota

4:15 Tests to Identify Cases Requiring Proportional Hazards Models with Estimated Inverse-Selection-Probability Weights
Qing Pan* and Douglas E. Schaubel, University of Michigan

4:30 Imputation in a Multimode Multi-instrument Study of Cancer Care
Yulei He* and Alan M. Zaslavsky, Harvard Medical School

4:45 Analysis of Effect of Side Collision Avoidance Signal on Simulated Driving with a Navigation System
Yi Ye* and Dongyuan Wang, University of North Florida

5:00 Evaluating Early Proteinuria Changes as a Surrogate Endpoint for Renal Disease Outcomes: a Bayesian Approach
Qian Shi* and Mary K. Cowles, University of Iowa

5:15 A New Multilevel Nonlinear Mixture Dirichlet Model for EST Data
Fang Yu*, Ming-Hui Chen and Lynn Kuo, University of Connecticut, Wanling Yang, The University of Hong Kong

Scientific Program

WEDNESDAY, MARCH 14
8:30–10:15 A.M.

87. BIOSURVEILLANCE AND ANOMALY DETECTION
Hanover C (Exhibit Level)

SPONSOR: ASA SECTION ON STATISTICS IN DEFENSE AND NATIONAL SECURITY
ORGANIZER: MYRON J. KATZOFF, NATIONAL CENTER FOR HEALTH STATISTICS
CHAIR: MYRON J. KATZOFF, NATIONAL CENTER FOR HEALTH STATISTICS

8:30 Biosurveillance and the BioSense Program
Henry R. Rolka*, Centers for Disease Control and Prevention
8:55 Automated Time Series Forecasting for Biosurveillance
Galit Shmueli*, University of Maryland, Howard S. Burkom and Sean P. Murphy, Johns Hopkins Applied Physics Laboratory
9:20 How to Lie with ROC Curves and Run AMOC
Howard S. Burkom* and Sean P. Murphy, Johns Hopkins Applied Physics Laboratory
9:45 Evaluation of the DC Department of Health's Syndromic Surveillance System
Michael A. Stoto*, Georgetown University School of Nursing and Health Studies, Beth Ann Griffin and Arvind K. Jain, RAND Corporation, John Davies-Cole, Chevelle Glymph, Garret Lum, Gebreyesus Kidane and Sam Washington, Department of Health
10:10 Floor Discussion

88. INNOVATIONS IN SURVIVAL ANALYSIS METHODOLOGY FOR PUBLIC HEALTH PROBLEMS
Baker (ACC Level)

SPONSOR: ENAR
ORGANIZER: ALEXANDER TSODIKOV, UNIVERSITY OF MICHIGAN
CHAIR: DEBASHIS GHOSH, UNIVERSITY OF MICHIGAN

8:30 Time-varying Cross-ratio Estimation for Bivariate Survival Data
Bin Nan*, University of Michigan, Xihong Lin and James M. Robins, Harvard University
9:00 Estimating the Effect of a Time-Dependent Therapy in Observational Studies
Douglas E. Schaubel* and John D. Kalbfleisch, University of Michigan
9:30 Generalized Shared Frailty Models
Alex Tsodikov*, University of Michigan, Szu-Ching Tseng, University of California-Davis
10:00 Floor Discussion

89. INSTRUMENTAL VARIABLE METHODS FOR CAUSAL INFERENCE
Hanover AB (Exhibit Level)

SPONSORS: IMS, ASA BIOMETRICS SECTION
ORGANIZER: DYLAN SMALL, UNIVERSITY OF PENNSYLVANIA
CHAIR: DYLAN SMALL, UNIVERSITY OF PENNSYLVANIA

8:30 Regression and Weighting Methods using Instrumental Variables
Zhiqiang Tan*, Johns Hopkins University
8:55 Structural Proportional Hazards Models for Causal Inference in Randomized Trials
Els J. Goetghebeur*, Ghent University-Belgium, An Vandebosch, Janssen Pharmaceutica-Belgium
9:20 Threshold Crossing Models and Bounds on Treatment Effects
Edward J. Vytlacil*, Columbia University, Azeem Shaikh, University of Chicago
9:45 Efficient Nonparametric Estimation of Causal Effects in Randomized Trials with Noncompliance
Jing Cheng*, University of Florida College of Medicine, Dylan S. Small, University of Pennsylvania, Zhiqiang Tan, Johns Hopkins University, Thomas R. Ten Have, University of Pennsylvania
10:10 Floor Discussion

90. DOSE-FINDING IN CLINICAL TRIALS
Regency V (Exhibit Level)

SPONSOR: ASA BIOPHARMACEUTICAL SECTION
ORGANIZERS: KEN CHEUNG, COLUMBIA UNIVERSITY; DARRYL DOWNING, GLAXOSMITHKLINE
CHAIR: DARRYL DOWNING, GLAXOSMITHKLINE

8:30 Model-based Designs for Drug Combinations
Valerii V. Fedorov*, GlaxoSmithKline
8:55 Two-stage Phase I Designs in Cancer Trials
Tze L. Lai*, Stanford University
9:20 Monitoring Late Onset Toxicities in Phase I Trials Using Predicted Risks
Neby Bekele, Yuan Ji*, Yu Shen and Peter F. Thall, University of Texas-M. D. Anderson Cancer Center

9:45 Patient-Specific Dose-Finding Based On Bivariate Outcomes With Covariates
Peter F. Thall* and Hoang Nguyen, University of Texas-M. D. Anderson Cancer Center
10:10 Floor Discussion

91. NON-STATIONARY TIME SERIES ANALYSIS WITH APPLICATIONS TO BIOMEDICAL DATA

Hanover E (Exhibit Level)

SPONSOR: ASA BIOMETRICS SECTION
ORGANIZER: WENSHENG GUO, UNIVERSITY OF PENNSYLVANIA
CHAIR: WENSHENG GUO, UNIVERSITY OF PENNSYLVANIA

8:30 Discrimination of Brain Signals Using Localized Higher Order Spectra
Hernando Ombao*, University of Illinois at Urbana-Champaign
8:55 Local Spectral Envelope
David S. Stoffer*, University of Pittsburgh
9:20 Multivariate Time-dependent Spectral Analysis Using Cholesky Decomposition
Ming Dai*, University of North Carolina-Charlotte, Wensheng Guo, University of Pennsylvania
9:45 Time-Frequency Functional Model
Li Qin*, Fred Hutchinson Cancer Research Center, Wensheng Guo and Brian Litt, University of Pennsylvania
10:10 Floor Discussion

92. INTRODUCTORY LECTURE SESSION: SOFTWARE PACKAGES FOR HANDLING MISSING DATA

Regency VI (Ballroom Level)

SPONSORS: ENAR, ASA SECTION ON TEACHING STATISTICS IN THE HEALTH SCIENCES
ORGANIZER: DIANE CATELLIER, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAIR: DIANE CATELLIER, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

8:30 XMISS in Cytel Studio 7 for Regression Models with Missing Covariates
Ming-Hui Chen*, University of Connecticut
9:00 Software Packages for Handling Missing Data
Nicholas J. Horton*, Smith College, Ken P. Kleinman, Harvard Medical School and Harvard Pilgrim Health Care
9:30 Missing Data Libraries for S-Plus
Joseph L. Schafer*, The Pennsylvania State University
10:00 Discussant: Donald Rubin, Harvard University

93. CONTRIBUTED PAPERS: ANALYSIS OF LABORATORY EXPERIMENTS

Inman (ACC Level)

SPONSOR: ENAR
CHAIR: KAREN CHISWELL, NORTH CAROLINA STATE UNIVERSITY

8:30 Designs for Modern Bioassay: Exploiting the Combined Strengths of Lab Robots and Multilevel Models
David M . Lansky*, Lansky Consulting, LLC
8:45 The Assessment of Antitumor Activity for Solid Tumor Xenograft Studies
Jianrong Wu*, St Jude Children's Research Hospital
9:00 Application of a Four Parameter Logistic Model for Estimating Titers of Functional Multiplexed Pneumococcal Opsonophagocytic Killing Assay (MOPA)
Deli Wang*, The Comprehensive Cancer Center-The University of Alabama at Birmingham, Robert L. Burton and Moon H. Nahm, The University of Alabama at Birmingham, Seng-jaw Soong, The Comprehensive Cancer Center-The University of Alabama at Birmingham
9:15 Binary Time Series Modeling with Application to Kinetic Studies in Micropipette Experiments
Ying Hung* and Chien-Fu Jeff Wu, Georgia Institute of Technology
9:30 The Exchangeable Logistic Regression in Correlated Data
Xin Dang* and Hanxiang Peng, University of Mississippi
9:45 Risk Analysis Using Generalized Linear Mixed Effects Models
Matthew W. Wheeler*, National Institute for Occupational Safety and Health, A. John Bailer, Miami University
10:00 Comparison of Designs for Response Surface Models with Random Block Effects
Sourish C. Saha* and Andre I. Khuri, University of Florida

94. CONTRIBUTED PAPERS: LONGITUDINAL COUNT DATA

Spring (ACC Level)

SPONSOR: ENAR
CHAIR: SUJATA PATIL, MEMORIAL SLOAN-KETTERING CANCER CENTER

8:30 Modeling Delirium Progression using a Non-homogeneous Markov Process
Rebecca A. Hubbard*, Jesse R. Fann and Lurdes YT Inoue, University of Washington

8:45 Modeling Longitudinal Count Data with the Possibility of Dropouts Mohamed A. Alosh*, Food and Drug Administration

9:00 Nonparametric Models for Multivariate Panel Count Data
Li-Yin Lee* and KyungMann Kim, University of Wisconsin, Madison

9:15 ***Regression Analysis of Multivariate Panel Count Data***
Xin He*, University of Missouri-Columbia, Xingwei Tong, Beijing Normal University, Jianguo Sun, University of Missouri-Columbia, Richard Cook, University of Waterloo

9:30 Covariance Structures for a Mixed Effect Markov Model for Repeated Binary Outcomes
Robert J. Gallop*, West Chester University

9:45 ***Semiparametric Estimation Methods for Panel Count Data Using Montone Polynomial Splines***
Minggen Lu*, Ying Zhang and Jian Huang, University of Iowa

10:00 Marginalized Random Effects Models for Longitudinal Ordinal Data
Keunbaik Lee* and Michael J. Daniels, University of Florida

95. CONTRIBUTED PAPERS: MICROARRAY ANALYSIS III

Dunwoody (ACC Level)

SPONSOR: ENAR
CHAIR: DONGXIAO ZHU, STOWERS INSTITUTE FOR MEDICAL RESEARCH

8:30 Applications of Spacings in Genomics
Stanley B. Pounds* and Cheng Cheng, St. Jude Children's Research Hospital

8:45 A Method for Gene Set Enrichment Analysis of Toxicogenomics Data
Rongheng Lin*, National Institute of Environmental Health Sciences, NIH, Shuangshuang Dai, Alpha-Gamma Technologies, Inc., Richard D Irwin, Alexandra N. Heinloth, Gary A. Boorman, Bhanu P. Singh and Leping Li, National Institute of Environmental Health Sciences, NIH

9:00 Feature Selection with a Supervised Partitioning Scheme
Yaomin Xu*, Case Western Reserve University and The Cleveland Clinic Foundation, Jiayang Sun, Case Western Reserve University

9:15 Analyzing Gene Expression Data from Some Non-microarray Transcription Profiling Techniques
Sujay Datta*, Texas A&M University

9:30 Clustering Threshold Gradient Descent Regularization: with Applications to Microarray Studies
Shuangge Ma*, Yale University, Jian Huang, University of Iowa

9:45 A Copula Approach to Missing Data in Genetic Association Studies
Gina M. D'Angelo*, Eleanor Feingold and Lisa A. Weissfeld, University of Pittsburgh

10:00 Floor Discussion

96. CONTRIBUTED PAPERS: QUANTITATIVE TRAIT LOCI

Courtland (ACC Leverl)

SPONSOR: ENAR
CHAIR: JAE K. YOO, LOUISVILLE UNIVERSITY

8:30 A Likelihood-based Method for Mapping Quantitative Trait Loci that Control Longitudinal Binary Responses
Hongying Li*, Ramon C. Littell and Rongling Wu, University of Florida

8:45 Nonparametric Modeling of Longitudinal Covariance Structure in Functional Mapping of Quantitative Trait Loci
John Stephen F. Yap* and Rongling Wu, University of Florida

9:00 Modeling of Multiple Traits for Genome-wide Epistatic QTL Mapping
Samprit Banerjee* and Nengjun Yi, University of Alabama at Birmingham

9:15 A Score Test for Linkage Analysis of Ordinal Traits
Rui Feng*, University of Alabama at Birmingham, Heping Zhang, Yale University

9:30 Analyzing and Modeling Dichotomous Traits in Large Complex Pedigrees
Charalampos Papachristou*, Carole Ober and Mark Abney, University of Chicago

9:45 An Approximate Bayesian Approach for Quantitative Trait Loci Estimation
Yu-Ling Chang*, Fred A. Wright and Fei Zou, University of North Carolina - Chapel Hill

10:00 QTL Detection for Zero-inflated Poisson Traits with Random Effects
Rhonda R. DeCook*, University of Iowa, Dan Nettleton, Iowa State University

97. CONTRIBUTED PAPERS: NON- AND SEMI-PARAMETRICS

Piedmont (ACC Level)

SPONSOR: ENAR
CHAIR: LIANMING WANG, NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES

8:30 Estimating Linear Functionals of Indirectly Observed Input Functions
Eun-Joo Lee*, Millikin University

8:45 The Matched Pairs Sign Test Using Bivariate Ranked Set Sampling
Hani M. Samawi*, Georgia Southern University, Mohammad F. Al-Saleh, Yarmouk University, Obaid A. Al-Saidy, Sultan Qaboos University

9:00 Estimation in Nonparametric Models with Missing Outcome at Random
Lu Wang*, Xihong Lin and Andrea Rotnitzky, Harvard University School of Public Health

9:15 On Robustness of Classification Based on Depth Transvariation
Nedret Billor*, Ash Abebe, Asuman S. Turkmen and Nudurapati Sai, Auburn University

9:30 Testing the Equality of Mean Functions for Continuous Time Stochastic Processes
Yolanda Munoz Maldonado*, University of Texas-Houston, School of Public Health

9:45 Semiparametric Varying Coefficient Periodic Regression Model
Yingxue Cathy Liu* and Naisyin Wang, Texas A&M University

10:00 Efficient Estimation of Population Quantiles in General Semiparametric Regression Models
Arnab Maity*, Texas A&M University

98. CONTRIBUTED PAPERS: KINETIC AND OTHER MODELING, INCLUDING PK/PD

Hanover D (Exhibit Level)

SPONSOR: ENAR
CHAIR: BRIAN CAFFO, JOHNS HOPKINS UNIVERSITY

8:30 Viral Kinetic Modeling of HBV DNA
Larry F. Leon* and Anne Cross, Bristol-Myers Squibb Pharmaceutical Research Institute

8:45 A New Steady State Assessment Method Based on Nonlinear Mixed Effect Modeling
Quan Hong*, Eyas Abu-Raddad and Didier Renard, Eli Lilly and Company

9:00 Bayesian Semi-parametric Analysis of PK/PD Mechanistic Models
Michele Guindani*, Peter Müller and Gary Rosner, University of Texas-M.D. Anderson Cancer Center

9:15 Covariate Model for Studying the Pharmacogenetic Architecture of Drug Response by Allometric Scaling
Wei Hou* and Rongling Wu, University of Florida

9:30 ***A Bayesian Approach for Differential Equation Models with Application to HIV Dynamic Studies***
Tao Lu* and Yangxin Huang, University of South Florida

9:45 A Parametric Model for a Choice RT Experiment
Jennifer L. Asimit*, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto, Willard J. Braun, University of Western Ontario, William A. Simpson, Simulation and Modelling Section, Defence Research and Development Canada

10:00 Clustered Data Models for the Infarcted Heart
Raymond G. Hoffmann*, Nicholas Pajewski, Xiaoguang Zhu, Ming Zhao and Raymond Migrino, Medical College of Wisconsin

WEDNESDAY, MARCH 14
10:15–10:30 A.M.

BREAK

Grand Foyer (Exhibit Level)

WEDNESDAY, MARCH 14
10:30 A.M.–12:15 P.M.

99. ERROR PROBABILITIES, SAMPLE SIZE, AND MULTIPLICITY: LIKELIHOOD METHODS

Regency VI (Ballroom Level)

SPONSOR: ASA BIOPHARMACEUTICAL SECTION
ORGANIZER: JAY HERSON, JOHNS HOPKINS UNIVERSITY
CHAIR: JAY HERSON, JOHNS HOPKINS UNIVERSITY

10:30 Follow the Likelihood Principle and Observe Misleading Evidence Less Often: Implications for Studies with Multiple Endpoints
Jeffrey D. Blume*, Brown University

11:00 Evidential Likelihood Approach to Multiple Test Adjustments: Application to Linkage Analysis
Lisa J. Strug*, Columbia University, Susan E. Hodge, Columbia University and New York State Psychiatric Institute

11:30 Likelihood-based Inference for Early Stopping in Single Arm Clinical Trials with Time-to-event Outcomes
Elizabeth Garrett-Mayer*, Sidney Kimmel Comprehensive Cancer Center-Johns Hopkins University
12:00 Discussant: Sue Jane Wang, Food and Drug Administration

100. INTRODUCTORY LECTURE SESSION: GROUP RANDOMIZED TRIALS

Courtland (ACC Level)

SPONSORS: ENAR, ASA SECTION ON TEACHING STATISTICS IN THE HEALTH SCIENCES
ORGANIZER: JI-HYUN LEE, H. LEE MOFFITT CANCER CENTER & RESEARCH INSTITUTE
CHAIR: JI-HYUN LEE, H. LEE MOFFITT CANCER CENTER & RESEARCH INSTITUTE

10:30 Power in Group-Randomized Trials
David M. Murray*, The Ohio State University
10:55 Dynamic Block-randomization in Group-randomized Trials when the Composition of Blocking Factors is not Known in Advance
Scarlett L. Bellamy*, University of Pennsylvania
11:20 The Use of Group Sequential Designs in Group Randomized Trials
Ziding Feng*, Fred Huchinson Cancer Research Center, Charles E. McCulloch, University of California at San Francisco
11:45 The Merits of Breaking the Matches; A Cautionary Tale
Allan Donner*, University of Western Ontario
12:10 Floor Discussion

101. MARGINALIZED MODELS

Regency V (Ballroom Level)

SPONSOR: ASA BIOMETRICS SECTION
ORGANIZER: BAHJAT QAQISH, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAIR: JOHN PREISSER, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

10:30 Marginalized Models for Longitudinal Ordinal Data
Michael J. Daniels* and Keunbaik Lee, University of Florida
11:00 Mixtures of Marginalized Models for Binary Process Data with Continuous-time Dropout
Li Su and Joseph W. Hogan*, Brown University
11:30 Marginalized Model Estimation under a Biased Sampling Study Design: A Conditional Likelihood Approach
Jonathan S. Schildcrout*, Vanderbilt University, Patrick J. Heagerty, University of Washington
12:00 Discussant: Patrick J. Heagerty, University of Washington

102. NEW METHODS USING STATISTICAL GRAPHICS AS TOOLS IN ADDRESSING RESEARCH QUESTIONS

Inman (ACC Level)

SPONSORS: ASA SECTION ON STATISTICAL EDUCATION, ASA SECTION ON TEACHING STATISTICS IN THE HEALTH SCIENCES
ORGANIZER: LINDA J. YOUNG, UNIVERSITY OF FLORIDA
CHAIR: LINDA J. YOUNG, UNIVERSITY OF FLORIDA

10:30 Visualizing Geospatial Time Series Using Micromaps and Conditioned Comparisons
Daniel B. Carr* and Chunling Zhang, George Mason University
11:00 Graph-theoretic Scagnostics for Projection Pursuit
Heike Hofmann*, Dianne H. Cook and Hadley Wickham, Iowa State University
11:30 Graphical Displays for the Analysis and Presentation of Complex Statistical Data
Linda W. Pickle*, StatNet Consulting, LLC
12:00 Floor Discussion

103. PANEL DISCUSSION: ROLE OF BIOSTATISTICIANS IN POLICY ISSUES

Hanover AB (Exhibit Level)

SPONSORS: ENAR, IMS
ORGANIZER: LISA LAVANGE, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAIR: MARIE DAVIDIAN, NORTH CAROLINA STATE UNIVERSITY

Panelists:	David Banks, Duke University
	Barry Graubard, National Cancer Institute
	Tom Louis, Johns Hopkins University
	Sally C. Morton, RTI International

104. CONTRIBUTED PAPERS: ENVIRONMENTAL STATISTICS

Dunwoody (ACC Level)

SPONSORS: ASA SECTION ON STATISTICS AND THE ENVIRONMENT, ENAR
CHAIR: BRENT JOHNSON, EMORY UNIVERSITY

10:30 Space-time Bayesian Survival Modeling of Chronic Wasting Disease in Deer
Hae Ryoung Song* and Andrew Lawson, University of South Carolina, Dennis Heisey and Erik Osnas, University of Wisconsin-Madison, Damien Joly, Fish and Wildlife Division, Alberta Sustainable Resource Development, Julie Langenberg, Wisconsin Department of Natural Resources

10:45 Projected Multivariate Linear Mixed-effects Models for Clustered Angular Data
Daniel B. Hall*, Lewis Jordan and Jinae Lee, University of Georgia, Jing Shen, IBM Thomas J. Watson Research Center

11:00 A Distance-based Classifier with Application to Microbial Source Tracking
Jayson D. Wilbur*, Worcester Polytechnic Institute

11:15 A Copula Model for Spatial Process with Air Pollution Application
Engin A. Sungur, University of Minnesota, Yoonsung Jung*, Kansas State University, Jong-Min Kim, University of Minnesota

11:30 Clustering with Logistic Mixture Model
Yongsung Joo*, Keunbaik Lee and Joonghyuk Kim, University of Florida, Sungtaek Yun, Korea University

11:45 On Comparison of Estimation Methods in Capture-Recapture Studies
Chang Xuan Mao and Na You*, University of California-Riverside

12:00 Comparisons of Sets of Multivariate Time Series
Jaydip Mukhopadhyay*, Nalini Ravishanker, University of Connecticut

105. CONTRIBUTED PAPERS: EPIDEMIOLOGY USING BAYESIAN AND EMPIRICAL BAYES METHODS

Baker (ACC Level)

SPONSORS: ASA SECTION ON STATISTICS IN EPIDEMIOLOGY, ENAR
CHAIR: LAURA GUNN, GEORGIA SOUTHERN UNIVERSITY

10:30 Semiparametric Bayesian Models for Short-term Prediction of Cancer Mortality Series
Kaushik Ghosh*, New Jersey Institute of Technology, Ram C. Tiwari, National Cancer Institute

10:45 Bayesian Analysis of the 1918 Pandemic Influenza in Baltimore, MD and Newark, NJ
Yue Yin*, Johns Hopkins University, Donald Burke, University of Pittsburgh, Derek Cummings, Johns Hopkins University and University of Pittsburgh, Thomas A. Louis, Johns Hopkins University

11:00 Constrained Bayes Prediction of Left-Censored HIV RNA Levels at a Meaningful Time Point
Reneé H. Moore*, University of Pennsylvania School of Medicine, Robert H. Lyles, Amita K. Manatunga and Kirk A. Easley, Rollins School of Public Health-Emory University

11:15 The Use of Hierarchical Models to Study Genetic Risk Factors
Marinela Capanu* and Colin Begg, Memorial Sloan-Kettering Cancer Center

11:30 Validating Risk Prediction Models using Family Registries
Wenyi Wang*, Johns Hopkins Bloomberg School of Public Health, Alison P. Klein, Johns Hopkins School of Medicine-Johns Hopkins Bloomberg School of Public Health, Brian Caffo, Johns Hopkins Bloomberg School of Public Health, Giovanni Parmigiani, Johns Hopkins Bloomberg School of Public Health-Johns Hopkins School of Medicine

11:45 On an Empirical Bayes Estimator for the Blue of HIV Population Based on CD-4 Cell Count
Suddhendu Biswas *, Sejong Bae, and Karan P. Singh, University of North Texas Health Science Center

12:00 The Dirichlet Process Prior for Choosing the Number of Latent Classes of Disability and Biological Topics
Tanzy M. Love*, Carnegie Mellon University, Cyrille Joutard, GREMAQ , University Toulouse 1, Edoardo Airoldi and Stephen Fienberg, Carnegie Mellon University

106. CONTRIBUTED PAPERS: METHODS FOR HIGH DIMENSIONAL DATA

Piedmont (ACC Level)

SPONSOR: ENAR
CHAIR: EUNHEE KIM, UNIVERSITY OF NORTH CAROLINA-CHAPEL HILL

10:30 Robust Partial Least Squares Regression
Asuman S. Turkmen* and Nedret Billor, Auburn University

10:45 Canonical Parallel Direction for Paired High Dimensional Low Sample Size Data
Xuxin Liu* and Steve Marron, University of North Carolina at Chapel Hill

11:00 Robust Tests for Detecting a Signal in a High Dimensional Sparse Normal Vector
Eitan Greenshtein, SAMSI, Junyong Park*, University of Maryland

11:15 Sufficient Dimension Reduction with Missing Predictors
Lexin Li* and Wenbin Lu, North Carolina State University

11:30 Improving Optimal Sufficient Dimension Reduction with Symmetric Predictors in Multivariate Regression
Jae Keun Yoo*, University of Louisville

11:45 Cancer Outlier Differential Gene Expression Detection
Baolin Wu*, University of Minnesota

12:00 Generalized Latent Variable Models for Spatial Correlated Binary Data with Applications to Dental Outcomes
Yanwei Zhang* and David Todem, Michigan State University, KyungMann Kim, University of Wisconsin-Madison

107. CONTRIBUTED PAPERS: MULTIVARIATE AND CORRELATED DATA

Hanover C (Exhibit Level)

SPONSOR: ENAR
CHAIR: BEN SAVILLE, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

10:30 Multivariate Clustered Non-normal Data Modeling: with Applications to Periodontal Research
Bin Cheng*, Columbia University

10:45 A General Method of Constructing a Test of Multivariate Normality With Application to Longitudinal Data Analysis
Tejas A. Desai*, The Indian Institute of Management-Ahmedabad, India

11:00 A Generalization of Z Test to Multivariate Normality
Haiyan Su*, Changyong Feng, Hongyue Wang, Xin Tu and Wan Tang, University of Rochester

11:15 Correlated Bivariate Continuous and Binary Outcomes: Issues and Applications
Armando Teixeira-Pinto*, Harvard University, Faculdade de Medicina da Universidade do Porto, Sharon-Lise Normand, Harvard Medical School-Harvard School of Public Health

11:30 A Normal-mixture Model with Random-Effects Method for Analyzing RR-Interval Data
Jessica M. Ketchum*, Virginia Commonwealth University

11:45 An R-square Statistic for Fixed Effects in the Gaussian Linear Model with Structured Covariance
Lloyd J. Edwards*, University of North Carolina at Chapel Hill, Keith E. Muller, University of Florida, Russell D. Wolfinger, SAS Institute, Bahjat F. Qaqish, University of North Carolina at Chapel Hill, Oliver Schabenberger, SAS Institute

12:00 An Improved Genetic Algorithm Using a Derivative-free Directional Search
Wen Wan* and Jeffrey B. Birch, Virginia Polytechnic Institute and State University

108. CONTRIBUTED PAPERS: DIAGNOSTICS II

Spring (ACC Level)

SPONSOR: ENAR
CHAIR: NUSRAT RABBEE, GENENTECH, INC.

10:30 A Probit Latent Class Model with General Correlation Structures for Evaluating Accuracy of Diagnostic Tests
Huiping Xu* and Bruce A. Craig, Purdue University

10:45 ROC Analysis for Longitudinal Disease Diagnostic Data Without a Gold Standard Test
Chong Wang*, Bruce W. Turnbull and Yrjo T. Grohn, Cornell University, Soren S. Nielsen, The Royal Veterinary and Agricultural University-Denmark

11:00 Bayesian Sample Size Determination with Two Possibly Correlated Imperfect Diagnostic Tests
Dunlei Cheng*, Baylor University

11:15 Comparing Multiple Sensitivities and Specificities with Different Diagnostic Criteria: Applications to Sexual Abuse Research and Studies of High-Risk Sexual Behavior
Qin Yu*, Wan Tang and Xin Tu, University of Rochester

11:30 Direct Estimation of the Area Under the Receiver Operating Characteristic Curve in the Presence of Verification Bias
Hua He* and Michael P. McDermott, University of Rochester

11:45 Application of Latent Class Models to Diagnostic Test Data
Afisi S. Ismaila*, McMaster University, Canada

12:00 Evaluation of Agreement between Observers Making Replicated Binary Assessments
Michael Haber and Jingjing Gao*, Rollins School of Public Health-Emory University, Huiman Barnhart, Duke Clinical Research Institute-Duke University

109. CONTRIBUTED PAPERS: STATISTICAL MODELS FOR GENETIC DATA

Hanover E (Exhibit Level)

SPONSOR: ENAR
CHAIR: SEOUNG BUM KIM, UNIVERSITY OF TEXAS AT ARLINGTON

10:30 Which Missing Value Imputation Method to Use in Expression Profiles: a Comparative Study and Two Selection Schemes
Guy N. Brock*, University of Louisville, John R. Shaffer, Richard E. Blakesley-Ball, Meredith J. Lotz and George C. Tseng, University of Pittsburgh

10:45 High-dimensional Model for Understanding the Genetic Network of Ontogenetic Allometry
Chenguang Wang*, Qin Li and Rongling Wu, University of Florida

11:00 Time Squared: Repeated Measures on Phylogeny
Hua Guo*, Robert E. Weiss and Marc A. Suchard, UCLA

11:15 Improving Identification of Regulatory Elements by using Context Dependent Markov Background Models
Nak-Kyeong Kim*, Kannan Tharakaraman and John L. Spouge, NCBI, NLM, NIH

11:30 Modelling and Estimating Differences in Allele Frequencies using Multiple SNPs
Nicholas J. I. Lewin-Koh*, Eli Lilly and Company, Lang Li, Indiana University School of Medicine

11:45 Random Forests and Multiple Imputation for Uncovering Haplotype Associations
B. Aletta S. Nonyane* and Andrea S. Foulkes, University of Massachusetts School of Public Health

12:00 Multivariate Approaches to Analyzing Gene expression Data Enhanced with the Domain Knowledge Daniel C. Parks*, Xiwu Lin and Kwan R. Lee, GlaxoSmithKline Pharma R&D

110. CONTRIBUTED PAPERS: LOG-RANK OR OTHER COMPARISONS OF SURVIVAL CURVES IN INDEPENDENT OR MATCHED SAMPLES

Hanover D (Exhibit Level)

SPONSOR: ENAR
CHAIR: XUEFENG LIU, WAYNE ST. UNIVERSITY

10:30 A Weighted Log-rank Test to Detect Early Difference in Censored Survival Distributions
Qing Xu* and Jong-Hyeon Jeong, University of Pittsburgh

10:45 Improving the Efficiency of the Logrank Test Using Auxiliary Covariates
Xiaomin Lu* and Anastasios Tsiatis, North Carolina State University

11:00 ***A Supremum Log-Rank Test for Adaptive Two-Stage Treatment Strategies and Corresponding Sample Size Formula***
Wentao Feng* and Abdus S. Wahed, University of Pittsburgh

11:15 Exact, Distribution Free Confidence Intervals for Late Effects in Censored Matched Pairs
Shoshana R. Daniel* and Paul R. Rosenbaum, University of Pennsylvania

11:30 Checking the Censored Two-Sample Accelerated Life Model using Integrated Cumulative Hazard Difference
Seung-Hwan Lee*, Illinois Wesleyan University

11:45 Inference for Survival Curves with Informatively Coarsened Discrete Event-Time Data: Application to ALIVE
Michelle D. Shardell*, University of Maryland School of Medicine, Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health, David Vlahov, Center for Urban Epidemiologic Studies-New York Academy of Medicine, Noya Galai, Johns Hopkins Bloomberg School of Public Health, Samuel R. Friedman, National Development and Research Institutes

12:00 Estimating Cumulative Treatment Effects in the Presence of Non-Proportional Hazards
Guanghui Wei* and Douglas E. Schaubel, University of Michigan

NOTES

ENAR WEBSITE

WWW.ENAR.ORG

POSTER SESSION

SELECTING AN ACCEPTABLE POPULATION FROM K MULTIVARIATE NORMAL POPULATIONS

Weixing Cai*, Syracuse University

We want to achieve the selection goal of selecting an acceptable population from k multivariate normal populations.

email: wcai01@syr.edu

A GENERALIZED LINEAR MODEL BASED APPROACH FOR ESTIMATING CONDITIONAL CORRELATIONS

Xueya Cai*, The State University of New York at Buffalo
Gregory E. Wilding, The State University of New York at Buffalo
Alan Hutson, The State University of New York at Buffalo

There exists an extensive literature on proposed methods for testing the equality of the population correlation coefficients between two variables across a third variable of interest. Currently available tests are specific to independent or dependent correlation coefficients, and include likelihood ratio tests and test statistics based on Fisher's z-transformation. Use of these testing procedures is limited to the case where the third variable is categorical. We present a generalized linear model approach for estimation and hypothesis testing about the correlation coefficient, where the correlation itself is modeled as a function of a numeric covariate. The strengths of this approach are (1) the tests are more statistically powerful than existing methods in certain situations, (2) complex trends of the correlation coefficients can be examined, (3) the approach can be expanded to include more than a single covariate, (4) the approach is flexible in that random effects may be considered, and (5) the model reduces to that corresponding to standard comparisons of correlations as a special case. The method is illustrated by analyzing a real life example.

email: xueyacai@buffalo.edu

RESAMPLING DEPENDENT CONCORDANCE CORRELATION COEFFICIENTS

John M. Williamson, Centers for Disease Control and Prevention
Sara B. Crawford, Centers for Disease Control and Prevention
Hung-Mo Lin*, Penn State College of Medicine

The concordance correlation coefficient (CCC) is a popular index for measuring the reproducibility of continuous variables. We examine two resampling approaches, permutation testing and the bootstrap, for conducting hypothesis tests on dependent CCCs obtained from the same sample. Resampling methods are flexible, require minimal marginal and joint distributional assumptions, and do not rely on large sample theory. However, the permutation test requires a restrictive assumption (exchangeability) which limits its applicability in this situation. Simulation results indicate that inference based on the bootstrap is valid, although type-I error rates are inflated for small sample sizes (30). For illustration we analyze data from a carotid stenosis screening study.

email: hlin@hes.hmc.psu.edu

CATEGORICAL DATA

A PREDICTIVE MODEL FOR BINARY PAIRED OUTCOME WITH TWO-STAGE SAMPLING

Jieqiong Bao*, Emory University
Ming Yuan, Georgia Institute of Technology
Jose N.G. Binongo, Emory University
Andrew Taylor, Emory University
Amita Manatunga, Emory University

We are interested in developing a model for predicting two different outcomes (kidney obstruction and need for diuretic) for subjects referred for Tc-99m MAG3 renal scans because of suspected urinary tract obstruction. A complication in constructing the likelihood arises because the MAG3 data are collected in a two-stage sampling scheme where the second outcome (obstruction) is obtained conditioning on the positive status of the first outcome (need for diuretic). We propose a maximum likelihood approach for the two outcomes while accounting for the correlation between responses from left and right kidneys and the two-stage sampling scheme. We model the marginal distributions using logistic regression at the first stage and an odds ratio representing the association between left and right kidneys. The second stage outcome is modeled conditioning on the response of the first outcome using a logistic model and we propose a maximum likelihood approach for estimating the parameters. The performance of the proposed model is examined by simulation studies and is applied to the nuclear medicine data from MAG3 renography studies.

email: jbao@sph.emory.edu

ESTIMATION USING THE NONCENTRAL HYPERGEOMETRIC DISTRIBUTION FOR COMBINING 2X2 TABLES

Kurex Sidik, Pfizer, Inc.
Jeffrey N. Jonkman*, Mississippi State University

We consider estimation of the overall log odds ratio and the between-study variance (or heterogeneity variance) from a series of independent 2x2 tables. The estimates are obtained using the conditional distributions of the 2x2 tables with fixed marginal totals, specifically the noncentral hypergeometric distribution, along with an assumed normal distribution of the true study-specific log odds ratios. The method presented differs from the traditional random effects meta-analysis approach for 2x2 tables in two respects. First, it does not assume a normal distribution for the observed within-study log odds ratio statistics; and second, it does not require estimation of the heterogeneity variance as a first step for estimation of the overall log odds ratio. Furthermore, the method directly provides an estimate of the heterogeneity variance and an associated standard error.

jonkman@math.msstate.edu

HIERARCHICAL BAYESIAN ANALYSIS OF BIVARIATE BINARY DATA

Ananya Roy*, University of Florida
Malay Ghosh, University of Florida
Ming-Hui Chen, University of Florida

We are interested in developing a model for predicting two different outcomes (kidney obstruction and need for diuretic) for subjects referred for Tc-99m MAG3 renal scans because of suspected urinary tract obstruction. A complication in constructing the likelihood arises because the MAG3 data are collected in a two-stage sampling scheme where the second outcome (obstruction) is obtained conditioning on the positive status of the first outcome (need for diuretic). We propose a maximum likelihood approach for the two outcomes while accounting for the correlation between responses from left and right kidneys and the two-stage sampling scheme. We model the marginal distributions using logistic regression at the first stage and an odds ratio representing the association between left and right kidneys. The second stage outcome is modeled conditioning on the response of the first outcome using a logistic model and we propose a maximum likelihood approach for estimating the parameters. The performance of the proposed model is examined by simulation studies and is applied to the nuclear medicine data from MAG3 renography studies.

email: aroy2@stat.ufl.edu

CALIBRATION OF THE CONTINUAL REASSESSMENT METHOD AND IMPLEMENTATION IN R

Ken Cheung*, Columbia University

Here is an increasing use and willingness to use outcome-adaptive designs by clinicians in dose-finding clinical trials. While there are a large number of statistical methods for dose-finding trials, it is equally important to have the tools to calibrate and modify a method so that it will yield good operating characteristics in practice. In this poster presentation, I will review several practical aspects directly related to the continual reassessment method (CRM), including the TITE-CRM for late onset toxicities, choosing dose-toxicity models in the CRM, and calibration of two-stage CRM. This poster also features numerical implementation of these techniques in R.

email: yc632@columbia.edu

CHALLENGES IN APPLYING THE INTENT-TO-TREAT PRINCIPLES TO A SCHOOL-BASED GROUP RANDOMIZED TRIAL

Kimberly Drews*, The George Washington University Biostatistics Center
Laure El ghormli, The George Washington University Biostatistics Center

The HEALTHY study is a multi-center primary prevention group randomized trial (GRT) designed to reduce or moderate risk for type 2 diabetes. This 3-year study is conducted in 42 middle schools evenly randomized to intervention or control. Enrolled students in the cohort are followed from 6th grade to 8th grade. Intent to treat (ITT) design is one approach where all subjects are followed until the end of the trial, irrespective of whether the subject is still receiving/complying with the assigned treatment. In a GRT, the 'group' (in this case school) is the unit of randomization, adherence, and analysis. Typically, when ITT principles have been applied it has been at the group level. Recently, however, there is an increasing interest to mimic ITT in the usual clinical trial fashion and apply at the subject level, which is not the unit of randomization. Dynamic settings such as the school environment limit the extent to which those principles can be applied. We present anticipated obstacles to ITT for GRT in a school-based setting. We highlight specific concerns to the HEALTHY study using data collected during an 8-week pilot study. The impact of ignoring the ITT principles is discussed from an analysis perspective.

email: kdrews@biostat.bsc.gwu.edu

A STEP-DOWN PROCEDURE WITH FEEDBACK FOR ELIMINATING INFERIOR TREATMENTS IN CLINICAL TRIALS

Chen-ju Lin*, Georgia Institute of Technology
Anthony J. Hayter, University of Denever

The problem of detecting which treatments or drugs are strictly inferior to the best ones has received a lot of attention in biostatistics and clinical trials. This study focuses on developing a new methodology for this problem, extending current methodologies in subset selection and multiple comparison procedures. Specifically, a new two-step procedure with feedback is investigated for the comparison of three unknown treatments. The feedback component utilizes information from the first stage when the second stage is implemented. While controlling the overall error rate of falsely detecting inferior treatments, the new procedure can provide greater power for eliminating the inferior treatments.

email: chenju.lin@gatech.edu

CALCULATING NUMBER NEEDED TO TREAT FOR CONTINUOUS OUTCOMES USING RECEIVER-OPERATOR CHARACTERISTIC ANALYSES

David R. Nelson*, Eli Lilly & Company
Haitao Gao, Eli Lilly & Company

Number needed to treat (NNT) is a measure of the number of patients required to be treated with a superior therapy to, on average, improve the outcome in one patient. NNT is the inverse of the difference of two proportions, and has been limited to dichotomous endpoints. We propose that NNT can be calculated for continuous endpoints using Receiver-Operator Characteristic (ROC) analyses. Drug-placebo and Drug-Drug response curves have been proposed as methods to examine treatment effects using ROC curves produced by reversing the typical analysis; therapy (e.g., A, B) as the “dependent variable,” the outcome (Yi) is the “independent” variable. The area under the curve (c-statistic) is Pr (Ya > Yb), assuming A is superior and increasing values of Y are desirable. For binary endpoints, the difference in proportions between therapies is directly related to AUC for drug response curves, equal to 2c-1. In the case of continuous endpoints, 2c-1 is Pr (Ya > Yb) - Pr (Ya < Yb). Assuming the probability of a patient responding to treatment A is Pr (Ya > Yb), and the probability of patients responding to treatment B is Pr (Ya < Yb), then the inverse of 2c-1 is the NNT to find one more “A” responder than “B” responder, a “generic” NNT for a continuous endpoint. Plotting drug response curves provides all “specific” NNTs for each possible cutoff.

email: nelsondr@lilly.com

SOME CONSIDERATIONS FOR ITT AND TREATMENT EMERGENT ADVERSE EVENT ANALYSES

Hui Quan*, Sanofi-Aventis
Qiankun Sun, Sanofi-Aventis
Ji Zhang, Sanofi-Aventis
Weichung J. Shih, University of Medicine and Dentistry of New Jersey

Different from the use of Intention-To-Treat (ITT) analysis for efficacy evaluation, for adverse event safety analysis, many pharmaceutical companies currently use Treatment Emergent Adverse Event (TEAE) analysis. In the analysis, period and adverse events occurring after a pre-specified post-treatment window will not be included. One consideration for using TEAE analysis is that including substantial off-drug period and events in the analysis might dilute the power for detecting safety signal especially if after discontinuation residual treatment effect diminishes quickly. Quantitative analyses will be performed to compare the power of ITT and TEAE analyses for several different settings and metrics (difference of rates and relative risk). The choice of approach for the analysis has an impact not just on analysis but also on study design. If ITT analysis is the choice, it should be specified at design stage in the protocol that all patients including those who discontinue from the study would be followed until the end of the study for all types of AEs. This sometimes may not be realistic particularly if an AE needs laboratory values to confirm.

email: hui.quan@sanofi-aventis.com

JUST SAY NO TO CHANGE FROM BASELINE ANALYSES

Timothy W. Victor*, Endo Pharmaceuticals
Richard E. White, Endo Pharmaceuticals

Measuring the magnitude of change of some attribute due to an intervention is the focus of many clinical trials. The ‘change from baseline’ strategy is often employed, whereby variates are computed by subtracting pretest scores from post-test scores to measure the effect of some intervention. Although the unsuitability of such scores has long been discussed in the measurement literature, they are still commonly used even by some otherwise sophisticated investigators. Several strategies for analyzing change will be presented and discussed. Finally, the interpretation of clinical trial data and the subsequent decisions made based on these analyses shall be explored from the perspective of measurement and statistical conclusion validity.

email: victor.timothy@endo.com

ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

EFFECT OF AIR POLLUTION (PM2.5 & PM10) ON LOW BIRTHWEIGHT IN NORTH CAROLINA

Simone Gray*, Duke University
Kerry Williams, Duke University
Sharon Edwards, Duke University
Eric Tassone, Duke University
Geeta Swamy, Duke University
Alan Gelfand, Duke University
Marie Lynn Miranda, Duke University

This study aims to determine whether airborne particulate matter (PM) exposure during pregnancy is associated with birthweight (BWT). Using North Carolina (NC) Detailed Birth Records for 2000–2003, USEPA PM2.5 & PM10 monitoring data were geographically linked to pregnant women living within 20km of a monitor. Analysis was restricted to singleton first births with no congenital anomalies, maternal medical complications, or alcohol use. Multivariate linear modeling was used to determine the association between PM and BWT controlling for gestation, race, education, marital status, and infant sex for N=55,246 (44,960) for PM2.5 (PM10) analyses. PM10 or PM2.5 exposure in the 2nd trimester was negatively associated with BWT ($p<.02$). Tobacco use was associated with 150 g lower BWT in both PM models ($p<.00001$). The combination of living in areas with 90%ile PM2.5 levels and tobacco use was associated with an additional 28 g decrease in BWT ($p<.005$). Maternal exposure to PM during the 2nd trimester is negatively associated with BWT. These effects are further compounded by maternal co-exposure to tobacco. Spatial mappings of PM levels would allow obstetricians to identify at-risk women and provide strategies to reduce PM exposure.

email: simone@stat.duke.edu

ADJUSTING FOR GENOTYPING ERROR IN NON-INVASIVE DNA-BASED MARK-RECAPTURE POPULATION STUDIES

Shannon M. Knapp*, Purdue University
Bruce A. Craig, Purdue University
Katy Simonsen, Purdue University and Bristol-Myers Squibb

DNA from non-invasive sources is increasingly being used as molecular tags for mark-recapture population estimation. These sources, however, provide small quantities of often contaminated DNA which can lead to genotyping errors that will bias the resulting population estimate. In this talk, we describe a novel approach, called GUAVA, to address this problem. GUAVA combines an explicit model of genotyping errors with an MCEM algorithm that both reduces the bias of the population estimate caused by genotyping errors and incorporates the additional uncertainty due to genotyping errors into the variance of the estimate. We demonstrate this approach on four different sets of genetic markers, with average Probability of Identities (PIDs) ranging from 6.3 x 10^-10 to 2.6 x 10^-5, and compare the results to both the uncorrected method and a 'Biologist-Corrected' method. GUAVA was found to be consistently less biased and to have consistently smaller variance than both the uncorrected and Biologist-Corrected methods. Additionally, GUAVA captured the true population size in the 95% confidence interval between 40 and 95.9% of the time with a Lincoln-Peterson estimator and between 93.4 and 96.8% with Bailey's Binomial estimator, compared to 0% for the other two methods with either estimator in all simulations.

email: knappsm@purdue.edu

ESTIMATING THE NUMBER OF SPECIES FROM A CENSORED SAMPLE

Chang Xuan Mao, University of California-Riverside
Junmei Liu*, University of California-Riverside

Estimating the number of species in a population based on a random sample of individuals has enormous important applications. For a variety of reasons, data are often censored in the sense that the exact numbers of individuals from some species are not recorded once they exceed a certain threshold. In this study, we use doubly-truncated Poisson mixture model as an alternative to the existing approximate procedures to estimate a sequence of lower bounds to the number of species and the effect of data censoring is investigated. Nonparametric likelihood estimation is used and the computational issues are considered. The results show that the data censoring did not cause big difference to the lower bounds to the number of species. Comparing the lower bounds to the number of species estimated by the existing procedures, the lower bounds estimated by our approach are much less sensitive to the data censoring threshold. Simulation study and real examples are provided to evaluate the proposed methods.

email: jliu005@student.ucr.edu

ARCHAEOLOGICAL APPLICATION OF GENERALIZED ADDITIVE MODELS

Yuemei Wang*, Rollins School of Public Health, Emory University
Lance A. Waller, Rollins School of Public Health, Emory University
Zev Ross, ZevRoss Spatial Analysis

We investigate the problem of cluster detection when adjusting for covariates using Generalized Additive Models (GAM). GAMs are extensions of generalized linear models by replacing partial or all parametric terms with smooth functions. In addition, GAMs smooth on 2-dimensional data by using thin-plate regression. We apply GAMs to the Black Mesa archaeological project to identify clusters of early versus late Anasazi settlement sites when adjusting for exposure to rivers around those sites. We compare the GAM results with those based on kernel density estimation of the early-to-late relative risk surface, SaTscan statistics and upper lever scan statistics. We also examine the definition and application of simultaneous confidence bands for the estimated probability surfaces using bootstrap simulation.

email: ywang27@emory.edu

EPIDEMIOLOGIC METHODS

THE DANGERS OF CATEGORIZING BMI IN STUDIES INVESTIGATING IN-HOSPITAL MORTALITY FOLLOWING CARDIAC SURGERY

Giovanni Filardo* Institute for Health Care Research and Improvement, Baylor Research Institute and Southern Methodist University
Cody Hamilton, Institute for Health Care Research and Improvement, Baylor Research Institute
Baron Hamman, Baylor University Medical Center
Hon KT Ng, Southern Methodist University
Paul Grayburn, Baylor University Medical Center

Aims: To investigate how categorizing body mass index (BMI) into weight classes can impact the assessment of the relationship between BMI and in-hospital mortality following coronary artery bypass graft surgery (CABG). Methods/Results: BMI-mortality (in-hospital) relationship was assessed in 5,762 patients who underwent isolated CABG at Baylor University Medical Center (Dallas, TX) between 1/1/1997 and 11/30/2003. Different ways of modeling BMI were used to investigate this association in a propensity-adjusted model, controlling for in-hospital mortality risk factors identified by the Society of Thoracic Surgeons and other clinical/non-clinical details. A highly significant ($p=0.003$) association between BMI (modeled with a restricted cubic spline) and mortality was found. Risk of in-hospital mortality was pronounced for subjects with BMI values in the mid-20s or over 40 kg/m^2. Study results were strongly affected by the way BMI was specified in the multivariable model. Only 4 of the 11 considered BMI categorizations produced significant results, and these results did not fully determine the effect of BMI on mortality. Conclusions: Study results are critically affected by the way BMI is treated in the analysis. Conceivably, findings of other studies investigating the relationship between BMI and adverse outcomes following CABG may be similarly affected.

email: giovanni.filardo@aya.yale.edu

LOCAL MULTIPLICITY ADJUSTMENTS FOR SPATIAL CLUSTER DETECTION

Ronald E. Gangnon*, University of Wisconsin-Madison

The spatial scan statistic is a widely applied tool for cluster detection. TThe spatial scan statistic evaluates the significance of a series of potential circular clusters using Monte Carlo simulation to account for the multiplicity of comparisons. The adjustment is analogous to a Bonferroni adjustment. In most settings, the extent of the multiplicity problem varies across the study region. For example, urban areas will have many overlapping clusters, while rural areas have few. The spatial scan statistic does not account for these local variations in the multiplicity problem.
We propose two new spatially-varying multiplicity adjustments for spatial cluster detection, one based on a nested Bonferroni adjustment and one based on local averaging. Geographic variations in power for the spatial scan statistics and the two new statistics are explored through simulation studies, and the methods are applied to both the well-known New York leukemia data and data from a case-control study of breast cancer in Wisconsin.

email: ronald@biostat.wisc.edu

CAUSAL INTERMEDIATE EFFECTS IN HIV PREVENTION RESEARCH

Giovanni Filardo, Institute for Health Care Research and Improvement
Cody Hamilton*, Institute for Health Care Research and Improvement

Research regarding HIV prevention frequently focuses on enhancing health literacy and behavior through interventions aimed at increasing HIV-related knowledge. Analyses describing the association between HIV-related knowledge and behavior should account for the casual intermediate attributes of HIV-related knowledge. Such an analysis may be beyond the reach of traditional regression techniques. In the present study a Bayesian modeling framework based on a previous published work is proposed to investigate the impact of HIV-related knowledge on willingness to get HIV testing in 300 educated pregnant women seen at 2 clinics in Urumqi, Xinjiang (China). The framework presented allows HIV-related knowledge to act both as a direct effect on willingness to get HIV testing and as an intermediary for other factors including age, ethnicity, occupation, and education while simultaneously accounting for any confounding effect of these factors. This framework provided an improved estimate of the association between HIV-related knowledge and willingness to get HIV testing. Related future research should account for the dual nature of HIV-related knowledge as intermediary and as direct causal effect on behavior, which can be achieved through the use of the proposed Bayesian hierarchical modeling coupled with Markov Chain Monte Carlo (MCMC) sampling.

email: giovanni.filardo@aya.yale.edu

STATISTICAL METHODS FOR ASSOCIATIONS BETWEEN EXPOSURE PROFILES AND RESPONSE

Amy H. Herring*, University of North Carolina at Chapel Hill
David A. Savitz, Mount Sinai School of Medicine

Exposures of interest in environmental epidemiology and other fields are often time-varying exposure profiles. In some settings, detailed information about these exposure profiles is collected but then summarized and included in a model as a predictor of a health outcome of interest. We compare a number of methods for analyzing exposure profiles, ranging from including simple mean exposures over time to using latent exposure profile class indicators as predictors of response to fitting joint models for exposure profiles and response. We provide guidance concerning settings in which simple summaries may be adequate and when more sophisticated strategies may be needed.

email: amy_herring@unc.edu

EVALUATING SPATIAL METHODS FOR CLUSTER DETECTION OF CANCER CASES

Lan Huang*, National Cancer Institute
Barnali Das, National Cancer Institute
Linda Pickle, National Cancer Institute

Power and sample size requirements have been developed for independent observations but not for spatially correlated data. We are developing such requirements for a test of spatial clustering and cluster detection for cancer cases with Poisson distribution. We compared global clustering methods including Morani¯s I, Tango's MEET, extended Tango's MEET and Besag-Newell R, and cluster detection methods including circular and elliptic spatial scan statistic (SaTScan), flexibly shaped spatial scan statistics, Turnbull's cluster evaluation permutation procedure (CEPP), local indicator of spatial autocorrelation (LISA) and upper level set scan statistics (ULS). We identified eight geographic patterns that are representative of patterns of mortality due to various types of cancer in the United States from 1995-2000. We then evaluated the selected spatial methods based on county level data simulated from these different spatial patterns in terms of geographic locations and relative risks, and varying samples using the 2000 population in each county. The comparison provides insight into the power, precision of cluster detection and computing cost of the spatial methods when applying the cancer count data.

email: huangla@mail.nih.gov

MODELING SURVIVAL: AN ALTERNATIVE TO PROPORTIONAL HAZARDS IN ALZHEIMER'S DISEASE

Elizabeth A. Johnson*, Johns Hopkins University
Kathryn Ziegler-Graham, St. Olaf College
Ron Brookmeyer, Johns Hopkins University

There is considerable variation in the literature of the effects of Alzheimer's disease on longevity. The relative risks of mortality associated with Alzheimer's appear to vary inversely with age of diagnosis, and are higher for females. We developed an alternative to the traditional proportional hazards model to help explain these findings. We used a multistate model in which persons progress from early to advanced disease and are at risk of death in all states. The model assumed that the effect of Alzheimer's disease on mortality was to add a constant amount to the background death rates once the disease became advanced. We used vital statistics for background mortality rates by age and gender, and then calibrated the model to the published literature on survival among Alzheimer's patients. We performed sensitivity analyses to different assumptions about the stages of Alzheimer's disease and compared our results to those obtained from the proportional hazards model. We found that the additive multistate model provides a more parsimonious and clinically interpretable description of the effects of Alzheimer's disease on mortality than the proportional hazards model. The proposed model is useful for forecasting disease prevalence, evaluating the effects of new treatments, and may also have application to other chronic diseases in the elderly.

email: ejohnson@jhsph.edu

STRUCTURAL EQUATION MODELING OF GENOTYPE BY ENVIRONMENT INTERACTION IN CORONARY HEART DISEASE

Xiaojuan Mi*, University of Nebraska-Lincoln
Kent M. Eskridge, University of Nebraska-Lincoln

Coronary Heart Disease (CHD) is a complex disease involving multiple genetic and environmental risk factors. CHD is the result of a number of interrelated physiological traits involving multiple causal relationships. Efforts to identify GEI have focused on two approaches: treating each exposure variables independently, and including all exposure in a single equation model. These efforts have not accounted for the complex interrelationships among variables and neglected the indirect effects of gene and risk factor through intermediate traits. Structural equation modeling (SEM) allows one to account for the effects of intermediate variables by simultaneously analyzing a system of equations where each equation describes a causal relationship among variables considered in the system. The purposes of this study are to use structural equation modeling methodology to account for these interrelationships among traits, to examine the effects of genotype-by-environment interaction on the incidence of CHD, and to evaluate the advantages of the proposed model over the single equation approach in terms of interpretation and fit. A cohort study will be carried out based on Framingham Heart Study offspring cohort data set.

email: xjmixu@yahoo.com

SENSITIVITY ANALYSIS FOR THE NESTED CASE CONTOROL DATA

Kenichi Yoshimura*, National Cancer Center-Japan

Cohort sampling is an attractive cost-reducing approach for the situation where the anticipated costs of the exposures measurements are high, especially for the genetic epidemiology. In the settings of measuring time-to-event endpoints, we widely use nested case-control studies with risk-set sampling in practice and Thomas proposed the widely-used estimator of the parameters of Cox's model. My motivating example is the nested case control study evaluating the relationship between the Helicobacter pylori infection and the gastric cancer incidence, in which we should take account of the potential confounders, e.g. poor hygiene situations. However, the sensitivity analysis is merely used for evaluating the confounders in this setting. The aim of this presentation is to propose the marginal structural models that are applicable to the sensitivity analysis of nested case-control design. For the estimation of the parameters, we proposed several estimators, which are consistent to the design. The results of simulation study showed favorable natures of the proposed estimator in comparison with conventional estimator. The proposed estimators can easily be applied to the more flexible sampling designs.

email: keyoshim@gan2.res.ncc.go.jp

INTERNAL PILOT DESIGN WITH INTERIM ANALYSIS

John A. Kairalla*, University of North Carolina-Chapel Hill
Keith E. Muller, University of Florida
Christopher S. Coffey, University of Alabama-Birmingham

Internal pilot studies for Gaussian data use an updated variance estimate at an interim stage to re-estimate needed sample size for a study. Allowing early stopping ability for efficacy or futility at the interim stage would increase the benefits of the study design. The proposed model for an internal pilot with interim analysis (IPIA) will be introduced and discussed. Exact forms for the joint distribution of the two stage test statistics can be used to derive exact forms for study power, type I error rate, and expected sample size. Results encompass a wide range of studies in the general linear univariate model (GLUM) setting with Gaussian errors and fixed predictors. This includes 1 and 2 group t-tests as special cases. The non-asymptotic nature of the theory makes it useful in small sample studies.

email: johnk@unc.edu

UNDERREPORTING IN A ZERO-INFLATED GENERALIZED POISSON REGRESSION MODEL

Mavis Pararai*, Georgia Southern University

Statisticians often make the assumption that the zeros in a zero-inflated regression model are correctly reported. Some of the zeros could be due to underreporting especially if the counts are self-reported. A zero-inflated generalized Poisson regression model for underreported counts is developed. Illustrative data is used to show how the model can capture the underreported zeros.

email: mpararai@georgiasouthern.edu

A UNIFIED APPROACH FOR COMPUTING POWER AND SAMPLE SIZE ESTIMATES FOR THE DORFMAN-BERBAUM-METZ AND OBUCHOWSKI-ROCKETTE METHODS FOR ANALYZING MULTIREADER ROC STUDIES

Stephen L. Hillis*, VA Iowa City Health Care System

The Dorfman-Berbaum-Metz (DBM) method is the most frequently used method for analyzing multireader ROC studies. Recently it has been shown that the DBM method and the Obuchowski-Rockette (OR) method yield identical results when based on the same procedure parameters. I discuss a unified approach for computing power and sample size estimates based on variance component estimates from pilot data analyzed using either either analysis method; this unified approach yields identical results for the two methods. Properties of the power and sample size estimates are examined in a simulation study and potential problem areas are discussed.

email: steve-hillis@uiowa.edu

NEW LARGE-SAMPLE CONFIDENCE INTERVALS FOR A BINOMIAL CONTRAST

Joshua M. Tebbs*, University of South Carolina
Scott A. Roths, Penn State University

We consider the problem wherein one desires to estimate a contrast (or a general linear combination) of binomial probabilities from $k>2$ independent populations. In particular, we create a new family of asymptotic confidence intervals, extending the approach taken by Beal (1987, Biometrics 43, 941-950) in the two sample case. One of our new intervals is shown to perform very well when compared to the best available intervals documented in Price and Bonett (2004, Computational Statistics and Data Analysis 45, 449-456). Furthermore, our interval estimation approach is quite general and could be extended to handle more complicated parametric functions and even to other discrete probability models, such as the Poisson or geometric, in stratified settings. We illustrate our new intervals using data from an ecological study recently conducted in Mississippi.

email: tebbs@stat.sc.edu

A NONPARAMETRIC MEAN ESTIMATOR FOR JUDGMENT POST-STRATIFIED DATA

Xinlei Wang*, Southern Methodist University
Johan Lim, Texas A & M University
Lynne Stokes, Southern Methodist University

MacEachern, Stasny and Wolfe (2004) introduced a data collection method, called judgment post-stratification (JP-S), based on ideas similar to those in ranked set sampling, and proposed methods for mean estimation from JP-S samples. In this paper we propose an improvement to their methods, which exploits the fact that the distributions of the judgment post-strata are often stochastically ordered, so as to form a mean estimator using isotonized sample means of the post-strata. This new estimator is strongly consistent with similar asymptotic properties to those in MacEachern, Stasny and Wolfe (2004). It is shown to be more efficient for small sample sizes, which appears to be attractive in applications requiring cost efficiency. Further, we extend our method to JP-S samples with imprecise ranking or multiple rankers. The performance of the proposed estimators is examined on three data examples through simulation.

email: swang@smu.edu

DISTRIBUTIONS FOR SUMS OF EXCHANGEABLE BERNOULLI RANDOM VARIABLES

Chang Yu*, Vanderbilt University Medical Center
Daniel Zelterman, Yale University

We describe new families of discrete distributions that can be used to model sums of exchangeable Bernoulli random variables. The distributions focus on the tail behavior of the sum of clustered exchangeable Bernoullis. We demonstrate binomial and beta-binomial are special cases of the new distributions. These discrete distributions can be parameterized in terms of their range, mean, variance, and the higher order correlations among the exchangeable Bernoullis. These models are fitted to examples involving mortality rates for children in a survey of families in Brazil and fetal deaths and malformations of a developmental toxicology study conducted at the NTP.

email: chang.yu@vanderbilt.edu

TRANSCRIPTION FACTOR ANALYSIS IN GENE EXPRESSION DATA

William T. Barry*, University of North Carolina
Fred A. Wright, University of North Carolina
Mayetri Gupta, University of North Carolina

DNA microarrays allow researchers to measure the coexpression of thousands of genes, and are commonly used to identify changes in expression either across experimental conditions or in association with some clinical outcome. With increasing availability of gene annotation, researchers have begun to ask global questions of functional genomics including the co-regulation of genes by transcription factors (TF). We present how models for the discovery of TF binding sites can be adapted to scoring upstream gene sequences for the presence of known motifs. We further expand the models of motif occurrences to look for the joint presence of TFs potentially interacting in cis-regulating manner. Next, a resampling-based hypothesis test is described that looks for increased differential expression in microarray experiments as evidence of transcriptional regulation. Analyses have been performed on both simulated and real datasets in order to demonstrate the validity and capability of this model-based framework.

email: wbarry@bios.unc.edu

DECTECTING QTLS BY BAYESIAN HIERARCHICAL REGRESSION MODEL

Susan J. Simmons, The University of North Carolina at Wilmington
Yi Chen*, The University of North Carolina at Wilmington

The problem of identifying the genetic loci contributing to variation in a quantitative trait (called QTL, for short) has been researched for a number of years. Most research focus on the problem with only one observation per genotype. For years, plant biologists have condensed replicates within lines to one genotype to use these conventional methods. We propose a hierarchical regression model that incorporates replicates within each line. To identify QTLs, we perform a stochastic search through the model space.

email: yc4918@uncw.edu

MAXIMUM LIKELIHOOD FACTOR ANALYSIS WHEN N < P

Karen E. Chiswell*, GlaxoSmithKline
John F. Monahan, North Carolina State University

Many implementations of maximum likelihood factor analysis (e.g., SAS PROC FACTOR, R's factanal() or Matlab's factoran) require that the number of observations, n, be greater than the number of variables, p. We have revisited Jöreskog's 1967 algorithm for fitting the k-factor model by maximum likelihood. Although the likelihood function is well defined when n < p, the objective function in Jöreskog's original algorithm is a function of log|S| (where S is the sample covariance matrix). When S is singular (e.g., when n <= p), this objective function is undefined. We define an equivalent objective function that does not involve the troublesome log|S|, and have implemented an algorithm for fitting the k-factor model by maximum likelihood when n < p. The dimensions of the data restrict the number of factors, and some care must be taken when computing eigenvalues of large symmetric matrices. We present results of simulation studies illustrating the performance of maximum likelihood factor analysis when n < p. In the applications that interest us, n is approximately 10 and p ranges from 25 to 50. We are most interested in the case where the true factor loadings are smooth functions of an ordered variable such as time.

email: karenc2204@yahoo.com

LATENT VARIABLE APPROACH FOR META-ANALYSIS OF GENE EXPRESSION DATA FROM MULTIPLE MICROARRAY EXPERIMENTS

Hyungwon Choi*, University of Michigan
Ronglai Shen, University of Michigan
Debashis, Ghosh, University of Michigan
Arul M. Chinnaiyan, University of Michigan

There are numerous challenges in combining analyses from multiple gene expression studies using microarray technology. Diverse array platforms not only harbor genes that appear in some studies only, but also produce heterogeneity in expression patterns indicating poor reproducibility across studies. Here we present a data integration method for meta-analysis of microarray data based on a fixed scale transformation of expression measurements, which potentially addresses the latter concern mentioned above through denoising effect of the data transformation. We also propose a computationally inexpensive algorithm for data transformation based on the expectation maximization (EM) algorithm, and compare it to full Bayesian modeling. Both data transformation algorithms have been implemented in the Bioconductor package metaArray. We demonstrate that the two transformations yield similar conclusions in terms of gene selection and biological interpretation through a meta-analysis of three gene expression studies from liver, lung, and prostate. We compare this method with other meta-analytic approaches in order to characterize the signature, and validate it on an independent breast cancer study with respect to its predictive ability of clinical outcomes.

email: hwchoi@umich.edu

ABOUT DICHOTOMIZED CONTINUOUS TRAITS IN FAMILY-BASED ASSOCIATION TESTS: DO YOU REALLY NEED QUANTITATIVE TRAITS?

David W. Fardo*, Harvard School of Public Health
Juan C. Celedon, Channing Laboratory-Brigham and Women's Hospital-Harvard Medical School
Benjamin A. Raby, Channing Laboratory-Brigham and Women's Hospital-Harvard Medical School
Scott T. Weiss, Channing Laboratory-Brigham and Women's Hospital-Harvard Medical School
Christoph Lange, Harvard School of Public Health-Channing Laboratory-Brigham and Women's Hospital-Harvard Medical School

In family-based association studies, quantitative traits are thought to provide higher statistical power than dichotomous traits. It is standard practice to collect quantitative traits and to analyze them as such. However, in many situations, continuous measurements are more difficult to obtain and/or need to be adjusted for other factors/confounding variables which also have to be measured. In such scenarios, it might be advantageous to record and analyze a 'simplified/dichotomized' version of the original trait. We will assess here the effects of using such a 'dichotomized' trait on the power of the test statistic. Using simulation studies, we derive rules for the dichotomization of quantitative traits under fairly general circumstances. The proposed rules achieve power levels that are comparable to the analysis of the original quantitative trait. This methodology can be utilized to increase the power of a study with a fixed budget. The guidelines are illustrated by an application to an asthma study.

email: dfardo@hsph.harvard.edu

AN ASSOCIATION-BASED METHOD TO DETECT EPISTASIS FOR QUANTITATIVE TRAITS IN FAMILY DATA

Guimin Gao*, University of Alabama at Birmingham
Hua Li, Stowers Institute for Medical Research

For quantitative traits in family data, although epistatsis testing has been incorporated into variance component linkage analysis method (Almasy and Blangero 1998), epistatsis testing based on association study is not well developed. We propose a linear mixed model approach based on association study to test epistasis for quantitative traits in family data. We implement the proposed model by a two stage approach using existing software. The first stage of analysis is typical polygenic analysis and can be accomplished by software SOLAR. The second stage of analysis is multiple linear regression analysis which can be implemented by any standard statistical analysis software. We apply the proposed method to a CEPH family gene expression dataset where we did detect epistasic effects.

email: guiming@uab.edu

A MIXED EFFECTS MODEL IMPLEMENTATION OF THE S-SCORE ALGORITHM

Richard E. Kennedy*, Virginia Commonwealth University
Kellie J. Archer, Virginia Commonwealth University

Analysis of Affymetrix GeneChip data is a complex, multistep process. Most often, methods that condense the multiple probe level intensities into single probe set level measures are applied, followed by application of statistical tests to determine which genes are differentially expressed. An alternative approach is a probe-level analysis, which tests for differential expression directly using probe-level data. Probe-level models offer the potential advantage of more accurately capturing sources of variation in microarray experiments. We present recent work examining a novel probe-level analysis algorithm for performing two-chip comparisons, the S-Score. We describe extensions to the S-Score that incorporates mixed effects modeling of probe-level microarray data. These extensions permit the analysis of experiments with multiple chips and conditions, and increase accuracy by using more rigorous estimates of the parameters for the S-Score statistic. Initial studies have shown that the performance of the S-Score compares favorably to other methods that use probeset expression summaries, and we present results of the assessment of the extended S-Score using spike-in datasets.

email: rkennedy@vcu.edu

NETWORK NEIGHBORHOOD ANALYSIS WITH THE MULTI-NODE TOPOLOGICAL OVERLAP MEASURE

Ai Li*, University of California-Los Angeles
Steve Horvath, University of California-Los Angeles

The goal of neighborhood analysis is to find a set of genes (the neighborhood) that is similar to an initial `seed' set of genes. Neighborhood analysis methods for network data are important in systems biology. If individual network connections are susceptible to noise, it can be advantageous to define neighborhoods on the basis of a robust inter-connectedness measure, e.g. the topological overlap measure. Since the use of multiple nodes in the seed set may lead to more informative neighborhoods, it can be advantageous to define multi-node similarity measures. The pairwise topological overlap measure is generalized to multiple network nodes and subsequently used in a recursive neighborhood construction method. A local permutation scheme is used to determine the neighborhood size. Using four network applications and a simulated example, we provide empirical evidence that the resulting neighborhoods are biologically meaningful, e.g. we use neighborhood analysis to identify brain cancer related genes.

email: aili@ucla.edu

DOMAIN ENHANCED ANALYSIS OF MICROARRAY DATA USING THE GENE ONTOLOGY

Jiajun Liu*, North Carolina State University
Jacqueline M. Hughes-Oliver, North Carolina State University
Alan J. Menius, Glaxo Smith Kline Inc.

New biological systems technologies give scientists the ability to measure thousands of bio-molecules including genes, proteins, lipids and metabolites. We use domain knowledge, e.g., the Gene Ontology, to guide analysis of such data. By focusing on domain-aggregated results at, say the molecular function level, increased interpretability is available to biological scientists beyond what is possible if results are presented at the gene level. We use a ``top-down'' approach to perform domain aggregation by first combining gene expressions before testing for differentially expressed patterns. This is in contrast to the more standard ``bottom-up'' approach where genes are first tested individually then aggregated by domain knowledge. The benefits are greater sensitivity for detecting signals. Our method, domain enhanced analysis (DEA) is assessed and compared to other methods using simulation studies and analysis of two publicly available leukemia data sets. Our DEA method uses functions available in R (http://www.r-project.org/) and SAS (http://www.sas.com/). The two experimental data sets used in our analysis are available in R as Bioconductor packages, ``ALL'' and ``golubEsets'' ({{http://www.bioconductor.org/}})

email: jliu6@stat.ncsu.edu

A FAST BAYESIAN METHOD FOR EQTL LINKAGE ANALYSIS IN EXPERIMENTAL CROSSES

Jinze Liu*, University of North Carolina-Chapel Hill
Fred Wright, University of North Carolina-Chapel Hill
Fei Zou, University of North Carolina-Chapel Hill
Yu-Ling Chang, University of North Carolina-Chapel Hill

We apply a recently-developed approximate Bayesian linkage analysis approach to the expression quantitative trait loci (eQTL) problem, in which microarray measurements of thousands of transcripts are examined for linkage to genomic regions. The approach uses the Laplace approximation to integrate over genetic model parameters (not including genomic position), and has been fully developed for backcross, F2 intercross, doubled haploid, and brother-sister mating recombinant inbred crosses. The method is much faster than commonly-used Monte Carlo approaches, and thus suitable for the extreme computational demands of eQTL analysis. We have formulated biologically attractive priors involving explicit hyperparameters for probabilities of cis-acting and trans-acting QTLs. The high-throughput nature of microarray data enables precise likelihood estimation of these hyperparameters, with Bayesian inference performed at the level of individual transcripts. The approach offers highly interpretable direct posterior densities for linkage for each transcript at each genomic position, although Bayes factors may also be easily computed. The speed of our approach enables posterior estimation even down to the resolution of individual gene positions, although in practice the accuracy of positional inference will be limited by the recombination resolution of the experimental cross. Simulation studies verify the applicability of the likelihood approach, with increasing precision for larger sample sizes. We present the model, simulation results, and analysis of real datasets.

email: liuj@cs.unc.edu

GENE PROFILING DATA OF THE RED LIGHT SIGNALING PATHWAYS IN ROOTS

Xiaobo Li, University of Florida
Richard Lee, University of Florida
Xueli Liu*, University of Florida
Melanie J. Correll, University of Florida
Gary F. Peter, University of Florida

The purpose of this study is to illustrate different statistical tools (packages) that can be used in affymetrix data analysis to reveal some hidden interesting genes for gene profiling data of the red light signaling pathways in roots. The data are first normalized using the RMA algorithm. Based on the normalized data, we filter 3707 genes out of 22410 by above 2-fold changes and some other criteria. A limma linear model is then applied to fit the gene expression data. Among these 3707 genes, we selected the top 200 differentially expressed genes. Kmeans and Pam are used to cluster these 200 genes and they produce essentially very similar results. A hierarchical clustering is then applied to order these genes into different groups which might function under the same biological pathways. A heatmap is constructed to see the clustering of the differentially expressed genes. Further validation of some of the interesting new genes revealed by this study will be investigated.

email: xbli@ufl.edu

HIERARCHICAL BAYESIAN MODEL FOR QTL DETECTION

Caroline A. Pearson*, AAIPharma Inc./University of North Carolina at Wilmington

Plant QTL experiments involve replicates within lines. In order to use conventional software programs, most plant biologists summarize line information into a single number and disregard variability within lines. We illustrate via a simulation study the increase in power by using a hierarchical Bayesian model that incorporates line variability over the available conventional methods for plant QTL experiments.

email: caroline.pappas@aaipharma.com

WEIGHTED RANK AGGREGATION OF CLUSTER VALIDATION MEASURES: A MONTE CARLO CROSS-ENTROPY APPROACH

Vasyl Pihur*, University of Louisville
Susmita Datta, University of Louisville
Somnath Datta, University of Louisville

Biologists often employ clustering techniques in the explorative phase of microarray data analysis to discover relevant biological groupings. In many cases, a thorough visual inspection of the results is performed to assess their significance and validity. The inadequacies of the visual approach are apparent and numerous objective validation measures have been proposed over the years to remedy the situation. Unfortunately, a given clustering algorithm can perform poorly under one validation measure while outperforming many other algorithms under a different one. In some studies, mostly visual attempts were made to simultaneously judge the relative performance of a clustering method under more than one validation measures. The success of the visual assessment, however, is hampered by the subjective nature of the procedure and the complexity that may arise when the number of validation measures to be considered is large. Using a weighted rank aggregation approach, we uccessfully combined the results from numerous cluster validation measures via Monte Carlo cross-entropy algorithm. We illustrate our procedure using one simulated and three real gene expression data sets for various platforms where we rank a total of twelve clustering algorithms using ten different validation measures.

email: vpihur@hotmail.com

STATISTICAL ISSUES AND ANALYSES OF IN VITRO GENOMIC DATA IN ORDER TO IDENTIFY CLINICALLY RELEVANT PROFILES IN VIVO

Laila M. Poisson*, University of Michigan
Debashis Ghosh, University of Michigan

Functional genomics studies are beginning to utilize controlled in vitro experiments to test specific hypotheses about gene expression that would be difficult to test in a human population. The information from gene lists obtained in these in vitro studies must then be applied to independent human (in vivo) samples, e.g. tumor samples in cancer patients. Though graphical models, such as dendrograms on heatmaps, are commonly used only the gene list is retained between experiments and quantification of the list's predictive ability is difficult. Here we explore a method for using the in vitro data to classify in vivo samples based on the correlation of the in vitro eigenarrays and the in vivo samples. Classification is based on the experimental design of the in vitro study and does not require that all human samples are classified. Clinical relevance and utility of the classification in the human samples is quantified by permutation testing, where the permutation sample size is adjusted to account for correlation between genes in the expression signature. The appropriate null hypothesis is discussed in this context We demonstrate these analyses using the Core Serum Response signature of Chang et al (2004) as an example signature.

email: lpoisson@umich.edu

A WAVELET-BASED METHOD FOR DETERMINING PERSISTENT DNA COPY NUMBER VARIATION ACROSS MULTIPLE SUBJECTS

William R. Prucka*, University of Alabama at Birmingham
Christopher S. Coffey, University of Alabama at Birmingham

Microarray-based comparative genomic hybridization (array-CGH) can assay DNA copy number variation (CNV) across the entire genome. CNV, insertion or deletion of genomic regions, has been implicated in cancer susceptibility and progression of diseases such as AIDS. For diseases with putative CNV causes, regions persistently varying in copy number across multiple affected individuals may serve as candidates for confirmatory analysis. We propose a new method for testing persistent CNV across multiple array-CGH signals. The significance of the candidate CNV is tested using “wavestrapping”, a wavelet-based bootstrapping procedure. Wavestrapping uses the discrete wavelet packet transform to resample the observed signals and provide a nonparametric assessment of the significance of CNV. This method obviates the need to “call” gains and losses, or make restrictive assumptions on the distribution of array probe levels. Preliminary results on simulated data show that the method is capable of resolving persistent CNV. Additional analyses will investigate the performance of the method on a breast cancer dataset and compare the results against other segmentation based procedures.

email: prucka@uab.edu

STATISTICAL METHODS FOR THE DISCOVERY OF GENE REGULATORY NETWORKS

Pingping Qu*, University of North Carolina at Chapel Hill
Mayetri Gupta, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill

Transcription factors (TFs) play a crucial role in gene regulation by binding to transcription factor binding sites (TFBSs) on the genome usually upstream of the genes. Gene regulation in comparatively simpler species like yeast has been studied extensively, and large databases of experimentally determined TFBSs now exist. However, with vast and complex genomes such as the human genome, similar methods for deciphering gene regulation and determining transcription factors do not perform as well. Here we considered a human time-course microarray gene expression data set under two biological conditions. As transcription factors directly affect the downstream gene transcription, one possible approach is to jointly model the gene expression data and the upstream sequence data. The statistical challenge here is to try to find from a vast set of motif candidates the possible different transcription factors binding to the genome under different biological conditions. We present a Bayesian hierarchical model and a Monte Carlo method to tackle this problem.

email: pqu@bios.unc.edu

A LIKELIHOOD RATIO TEST OF INCOMPLETE DOMINANCE VERSUS OVERDOMINANCE AND/OR UNDERDOMINANCE WITH APPLICATION TO GENE EXPRESSION LEVELS

Kai Wang*, University of Iowa

For genetically inherited phenotypes, the value of the phenotype value depends on the underlying genotypes. The effect of the genotype at a locus on a phenotypc can be classified as overdominance, underdominance or incomplete dominance. Distinction of these effects has importnat implicaitons to natural selection, imbreding depression and gene mapping. In this report, I develop two likelihood ratio test procedures, one for testing incomplete dominance against overdominance (one-sided test) and the other for testing incomplete dominance against overdominance and underdominance (two-sided test). These testing procedures are applied to a data set that contains gene expression levels and genotype data from a sample consisting of 60 B6 $\times$ BTBR F_2 mouse.

email: kai-wang@uiowa.edu

MULTIVARIATE CORRELATION ESTIMATOR FOR REPLICATED OMICS DATA

Dongxiao Zhu*, Stowers Institute for Medical Research
Youjuan Li, University of Michigan

Estimating pairwise correlation from replicated omics data is fundamental to the active area of pattern discovery research. The existing approach estimates bivariate correlation by averaging over replicates. It is not completely satisfactory since it introduces strong bias while reducing variance. We propose a new multivariate correlation estimator that models all replicates as samples from the multivariate normal distribution. We derive the estimator by maximizing a likelihood function. For small-sample data, we provide a re-sampling based statistical inference procedure; For larger-sample data, we provide an asymptotic statistical inference procedure based on the likelihood ratio test. We demonstrate the advantages of the new multivariate estimator over the traditional bivariate estimator using simulations and real data analysis examples. The proposed estimation and inference procedures have been implemented in an R package ``GeneNT' that is available from http://cran.r-project.org/.

email: doz@stowers-institute.org

A HIDDEN MARKOV MODEL FOR INFERRING HAPLOTYPE STRUCTURE FROM MOUSE SNP DATA

Jin P. Szatkiewicz*, The Jackson Laboratory
Glen L. Beane, The Jackson Laboratory
Gary A. Churchill, The Jackson Laboratory

Single nucleotide polymorphisms (SNPs) are the most abundant type of genetic variation in mammals. Recent studies of sequence variation using whole genome SNP data suggest that genetic variation is organized in haplotype blocks. In particular, it has been suggested that the mouse genome has a mosaic structure of polymorphisms attributable to recent descent from a limited number of genetically diverse founders. Understanding the haplotype structure of the laboratory mouse will accelerate the identification of genes associated with complex phenotypes. However, missing data, genotyping errors, mutations and the absence of clear haplotype structure in some genomic regions present analytic challenges that must be addressed. We have developed a Hidden Markov Model and a software tool that assigns individual strains to local haplotypes and imputes missing SNP alleles. An ad-hoc state-trimming approach improves the performance of the model, resulting in a more concise and interpretable haplotype reconstruction. We illustrate the method using a publicly available dataset of 130,000 SNPs collected on 42 inbred mouse strains.

e-mail: jin.szatkiewicz@jax.org

HEALTH POLICY APPLICATIONS

IMPLICATIONS OF JOINT EFFECTS WITHIN THE FAMILY FOR TREATMENT CHOICES

John A. Myers*, University of Louisville

Purpose: Cost-effectiveness analyses traditionally treat patients as isolated individuals and ignore the effects of improvement in patients' health on the welfare of closely linked individuals. Spillover effects deal with this issue by However, the closely linked individual does not receive any health intervention. A novel measurement, joint effects, is introduced which measures the direct or indirect effect on closely linked individual's utility as well as the indirect effect on the patient's utility through the closely linked individual's utility who both undergo a health intervention. Methods: We focus on a two-person family to model how these joint effects might affect treatment choices by using a model based on a family utility function with altruistic linkages to show that there can be direct and indirect effects on the welfare of closely linked individuals. Results: Our results are consistent with the model's predictions. By using the collective model to represent the family's well being, we show how the optimality condition changes in the presence of altruistic linkages. Conclusions: The concept of spillover effects is extended to include a special case, joint effects.

email: john.myers@louisville.edu

SUPPRESSION RULES FOR UNRELIABLE ESTIMATES OF HEALTH INDICATORS

Jennifer D. Parker*, National Center for Health Statistics/CDC
Diane M. Makuc National Center for Health Statistics/CDC

Health statistics for small groups defined by such characteristics as age, race/ethnicity, socioeconomic position, and geography are of interest to policymakers. However, some estimates for small groups and rare health conditions are unreliable, whether based on surveys or complete counts such as death certificates. Federal reports and online tables use different criteria for flagging or suppressing estimates failing to meet specific reliability standards. Suppression criteria vary across data systems and may be expressed in terms of a relative standard error (for example >30%), number of events (for example < 20), width of the confidence interval, and/or sample size. Criteria can lead to different decisions. For example, suppressing rates with < 20 events always leads to rates with a low relative standard error (< 25%) but may not meet usual confidence interval rules. This presentation compares suppression rules used for tabular data dissemination at the National Center for Health Statistics. Using the minimum sample size to express the different criteria in a common unit, the comparison shows how rules can be relevant for some scenarios but not for others. With increasing demand for dissemination of small group estimates, the need to clearly convey data limitations also increases.

email: jdp3@cdc.gov

IMAGING

BAYESIAN ANALYSIS OF A BIVARIATE AUTORADIOGRAPHIC IMAGE OF TUMORS: LOCAL VS. GLOBAL CORRELATION

Timothy D. Johnson*, University of Michigan

We present a bivariate imaging model to investigate local versus global correlation of two radiolabeled compounds (FAZA and RGD). FAZA is a compound that is selectively taken up by hypoxic cells while RGD peptides are implicated in the inhibition of angiogenesis. Conventional wisdom dictates that hypoxic regions of tumors and antiangiogenesis are positively correlated. A correlation analysis of dual autoradiographs of tumors harvested from mice that ignores the spatial correlation within the radiographs corroborates this wisdom. However, when we account for the spatial correlation of these images a new picture emerges: one that suggest local regions of negative correlation. This has important implications in the testing of molecularly targeted therapies. Our model is a hidden Markov random field model. The radiographic intensities, conditional on class membership, are modeled as a bivariate normal distribution. The prior on class membership is a Gibbs distribution that induces spatial correlation in the posterior. We use RJMCMC to estimate the number of classes. The primary outcome of interest is the correlation of FAZA and RGD intensities at each pixel in the image. We show that there are localized regions of negative correlation suggesting a reevaluation of convential wisdom.

email: tdjtdj@umich.edu

EXAMINING MODULATION OF FUNCTIONAL NETWORKS IN MULTI-SESSION MULTI-SUBJECT FMRI

Rajan S. Patel*, Amgen, Inc.

Recent work regarding the analysis of brain imaging data has focused on examining functional relationships among spatially distributed brain regions. To date, emphasis has been placed on determining functionally connected areas under certain conditions (including the resting state condition) using various methods such as seed voxel correlation approaches and independent components analysis. We developed a hypothesis-unconstrained data-driven method for multi-session multi-subject fMRI data to examine and visualize inter-session modulation of functional networks using a hybrid anatomical and functional parcellation of the brain. We construct two-level mixed effects models of a normalizing transformation of the correlation of the fMRI time courses of each pair of the anatomical volumes of interest. With permutation test p-values, we are able to control family-wise error rates and determine statistically significant modulations of functionally connected networks. We apply this method to multi-session multi-subject fMRI data to examine the effect of methylphenidate on functional connectivity under various tasks.

email: rajan@alumni.rice.edu

TEST-STATISTICS BY SHRINKING VARIANCE COMPONENTS WITH AN APPLICATION TO FMRI

Shuchih Su*, Johns Hopkins Bloomberg School of Public Health
Brian Caffo, Johns Hopkins Bloomberg School of Public Health
Elizabeth Garrett-Mayer, Johns Hopkins Kimmel Cancer Center
Susan Spear Bassett, Johns Hopkins University

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive technique that is commonly used to quantify changes in blood oxygenation and coupled to neuronal activation. Group fMRI studies aim to study neuronal activity using blood oxygenation as a proxy. Single voxel t-tests have been commonly used to determine whether activation related to the protocol differs across groups. Due to the limited number of subjects within each study, efficient estimation of residual variance at each voxel is difficult. Thus combining information across voxels in the statistical analysis of fMRI data is desirable. We construct a hierarchical model and apply a Bayesian framework on the analysis of group fMRI data, employing techniques used in high throughput genomic studies. The key idea is to shrink residual variances by combining information across voxels, and subsequently to construct an improved test statistic. This hierarchical model results in a shrinkage of voxel-wise residual sample variances towards a common value. In simulation studies, the shrunken estimator for voxel-specific variance components on the group analyses outperforms (in the terms of MSE) the classical residual variance estimator which and leads to a more powerful and robust test statistic. Data from an experiment on auditory word-pair-associates paradigm is explored.

email: shsu@jhsph.edu

COMPARISON OF MULTIVARIATE STRATEGIES FOR THE ANALYSIS OF SKEWED PSYCHOMETRIC DATA WITH MANY ZEROS

G.K. Balasubramani*, Epidemiology Data Center, GSPH, University of Pittsburgh
Stephen R. Wisniewski, Epidemiology Data Center, GSPH, University of Pittsburgh

Available literature on analyzing skewed data is inadequate and an optimal method is yet to be identified. An ordinal data measurement always produces skewed distributions. We discuss methods of analyzing data that are positively skewed and with many zero values. The methods can be used to model the skewed data as a response variable in terms of one or more independent variables. The statistical analysis consists of five stages. The first stage involves modeling the skewed data using ordinary regression. The second and third stage involves modeling the presence data using logistic regression with two different thresholds. The fourth and final stage involves the use of ordinal regression with the ordered measurements as groupings as well as the original ordered value itself. Our objectives are (i) to compare the five different multivariate strategies for analyzing skewed psychometric outcome data through measure of parameter stability and (ii) to promote wider application of this approach for ordinal data measurements. We report our findings and recommend approaches for analyzing skewed data for large samples through simulation and the empirical study using the FIBSER1 outcome data.

email: balagk@edc.pitt.edu

USING A BIVARIATE BINOMIAL MIXTURE MODEL TO ESTIMATE BIVARIATE COMPONENT DISTRIBUTIONS IN A CONTINUOUS MIXTURE MODEL

Tatiana A. Benaglia*, Pennsylvania State University
Thomas Hettmansperger, Pennsylvania State University

Suppose that a group of subjects performs a series of similar tasks (repeated measures) at a certain stage in life and these repeated measures can be assumed independent. Usually, subjects are not homogeneous with respect to their responses, that is, there are multiple distinct solution strategies and each strategy is characterized by a component distribution in a mixture model. Considering observations taken at a certain stage, it is possible to estimate each component cdf without assuming any parametric shape for the distributions. Using a cut point model we estimate the posterior probabilities and use these as weights to get a weighted kernel density estimate. We extend this idea to the bivariate case where observations are taken in two different stages. In this case, observations are not independent over time and our interest is not only to describe each component group in each stage but also to model the correlation between these stages. We use the cut point model and apply a bivariate binomial mixture to estimate the posterior probabilities. We show that the model works for estimating the moments for each component and also gives a good estimate of the correlation. Using the posterior probabilities we can estimate the bivariate pdfs for each component.

email: tab321@stat.psu.edu

FACTORS AFFECTING VEHICULAR LATERAL CONTROL MEASURES IN DRIVING SIMULATORS

Jeffrey D. Dawson*, University of Iowa
Joshua D. Cosman, University of Iowa
Yang Lei, University of Iowa
Elizabeth Dastrup, University of Iowa
JonDavid Sparks, University of Iowa
Matthew Rizzo, University of Iowa

In simulated studies of driving, information on driver performance is recorded at high frequencies and then reduced into measures of vehicular control. Boer (2000) proposed a measure called 'steering entropy' to quantify variability in the position of the steering wheel, and Dawson et al (2006) have applied this method to the lateral position of a simulated vehicle within the driving lane. We recently performed a factorial experiment in a driving simulator known as SIREN to assess whether speed, road curvature, and/or steering strategies affect the entropy measures of steering wheel and lane position. We found that curved roads are associated with higher values of entropy, while a steering strategy based on sudden jerks of the steering wheel is associated with lower values of entropy. Speed had very little effect on entropy in the range we examined, although may interact with road curvature. Surprisingly, the number of times that the car crossed the traffic lanes did not seem to be correlated with entropy. These results suggest that entropy may be a useful tool in quantifying vehicular control, but caution should be exercised when interpreting the results, as the directions of association are not always in the anticipated direction.

email: jeffrey-dawson@uiowa.edu

HIGH BREAKDOWN INFERENCE FOR THE CONSTRAINED MIXED-VARX MODEL

Mark A. Gamalo*, University of Missouri-Kansas City

The proliferation of many clinical studies obtaining multiple biophysical signals from several individuals repeatedly in time is increasingly recognized, a recognition generating growth in statistical models that analyze cross-sectional time series data. In general, these statistical models try to explain intra-individual dynamics of the response and its relation to some covariates; and how this dynamics can be aggregated consistently in a group. However, one problem inherent in these signals is the spurious occurrence of observations beyond what is physically reasonable which could be due to error in data acquisition. This, if not taken cared of, can have deleterious effects in a statistical model e.g. biased parameter estimates. To address these problems, we propose a covariate-adjusted constrained Vector Autoregressive model, a technique similar to the STARMAX model (Stoffer, {\it JASA} {\bf 81}, 762-772), to describe serial dependence of observations. Estimation of the resulting model is carried through the use of robust procedures that control effects of outliers. Usually, inference on these estimates is done using bootstrap. However, in higher dimensions convergence of the bootstrap approximation is a problem. We use instead a high breakdown inference based on the asymptotic distribution of the robust estimates.

email: gamalom@umkc.edu

MULTIPLE INDICATOR HIDDEN MARKOV MODEL WITH AN APPLICATION TO MEDICAL UTILIZATION DATA

Ran Li*, University of Minnesota
Melanie Wall, University of Minnesota
Tianming Gao, University of Minnesota

Many medical data are longitudinal and hence involve repeated observations on each of many individuals. When the probability of observing changes from one time to the next is the major interest of the researcher, Markov models may be suitable for the data analysis. Furthermore, rather than considering the probability of moving from observation to observation, the concept of underlying “health states” can be introduced and consequently how subjects move from one health state to another and whether this dynamic moving pattern of the subjects is affected by certain treatment can be investigated. Hidden Markov models are introduced for modeling these unobserved hidden states. Since a simple univariate hidden Markov model is not sufficient for longitudinal data when there are more than one outcome variable measured at each time on each individual, multiple indicator hidden Markov models are instead applied treating the multiple outcomes as reflections of one common underlying “health state”. The motivating dataset in this talk postulates a hidden state process underlying multiple types of medical encounters collected monthly within an insurance claims database where the states characterize individuals’ medical use prior to and after alcoholism treatment.

email: lixxx311@umn.edu

A NONLINEAR LATENT CLASS MODEL FOR JOINT ANALYSIS OF MULTIVARIATE LONGITUDINAL DATA AND A TIME-TO-EVENT

Cecile Proust-Lima*, University of Michigan
Helene Jacqmin-Gadda, University of Michigan

We propose a joint model for exploring the association between correlated longitudinal markers and time-to-event in the context of cognitive ageing and occurrence of dementia. A longitudinal latent class model describes the different latent classes of evolution of the latent process representing cognition, and a proportional hazard model describes the risk of dementia according to the latent classes. The latent process representing cognition is linked to the markers of cognition by using nonlinear transformations which include parameters to be estimated. An asset of this method is that by introducing flexible nonlinear transformations between the markers and the latent process, non Gaussian continuous markers can be considered. Estimation of the parameters is achieved by maximizing the observed log-likelihood and some methods are proposed to evaluate the goodness-of-fit of the joint model and particularly the assumption of independence between the markers and the time-to-event conditionally on the latent classes. This methodology is applied to data from PAQUID, a French prospective cohort study of ageing, in order to investigate the different profiles of cognitive decline and their association with dementia.

email: cecile.proust@isped.u-bordeaux2.fr

AN RKHS FORMULATION OF DISCRIMINATION AND CLASSIFICATION FOR STOCHASTIC PROCESSES

Hyejin Shin*, Auburn University

Modern data collection methods are now frequently returning observations that should be viewed as the result of digitized recording or sampling from stochastic processes rather than vectors of finite length.In spite of great demands, only a few classification methodologies for such data have been suggested and supporting theory is quite limited. Our focus in this talk is on discrimination and classification in the infinite dimensional setting. Specially, we have developed a theoretical framework for Fisher's linear discriminant analysis of sample paths from stochastic processes through use of the Lo\`{e}ve-Parzen isomorphism that connects a second order process to the reproducing kernel Hilbert space generated by its covariance kernel. This approach provides a seamless transition between finite and infinite dimensional settings and lends itself well to computation via smoothing and regularization.

email: hjshin@auburn.edu

MISSING DATA, MEASUREMENT ERROR, AND CAUSAL INFERENCE

PROPENSITY SCORE SUBCLASSIFICATION FOR THE EFFECTS OF TIME-DEPENDENT TREATMENTS: APPLICATIONS TO DIABETES

Constantine Frangakis, Johns Hopkins University
Aristide Achy-Brou*, Johns Hopkins University
Michael Griswold, Johns Hopkins University

Setting: Time-Dependent (longitudinal) treatment regimes with long term residual effects. Assumption: Perfect knowledge of the variables used in the assignment mechanism. Infinite sample: Single-time (Rubin83) and longitudinal (Robins87). Practical constraint: Dimension of the variables used in the assignment mechanism is huge. Key contribution: We offer a likelihood method for using propensity scores with time-dependent treatments.

email: aachybro@jhsph.edu

ESTIMATION OF THE MEAN RESPONSE IN SAMPLE SELECTION MODELS WITH ENDOGENOUS COVARIATES

Joseph Gardiner*, Michigan State University
Zhehui Luo, Michigan State University

The presence of endogenous covariates and sample selection need to be addressed in analyses of discrete response and censored regression models of observational data. We consider the estimation of the mean of a continuous response in a linear model for the log-transformed response with sample selection where the joint model shares endogenous covariates. The QLIM procedure in SAS/ETS and the NLMIXED procedure in SAS/STAT software are harnessed for maximum likelihood estimation using a classic empirical example of wages of women participants in the labor force.

email: jgardiner@epi.msu.edu

A COMPARISON OF TWO PROCEDURES FOR ACCOMMODATING ATTRITION IN SUBSTANCE ABUSE CLINICAL TRIALS

Sarra L. Hedden*, Medical University of South Carolina
Robert F. Woolson, Medical University of South Carolina
Robert J. Malcolm, Medical University of South Carolina

Drop-out due to attrition may affect the validity of a clinical trial. When extensive drop-out occurs, specific statistical hypothesis tests of the treatment effect may not have the appropriate size or power. The major aims of this project are to assess size and power of two specific statistical hypothesis tests of the treatment effect: Stratified Summary Statistics (SSS) and Last Observation Carried Forward (LOCF). Methods included a Monte Carlo simulation study incorporating the general design of substance abuse clinical trials. Preliminary results of the simulation study demonstrate that SSS have varying Type I error rates dependent on both sample size and number of stratum factors. Type I error rates were >.05 given smaller sample sizes and greater stratum factors; Type I error rates for LOCF were less than or equal to .05. Type I error rates of SSS were robust to missing data rates/mechanisms. SSS was a more powerful test of the treatment effect compared to LOCF under a variety of missing data rates and mechanisms. The use of any missing data method must consider the assumptions/design of a clinical trial and the missing data mechanism as well as their possible effects on both size and power of the test.

email: heddens@musc.edu

REGRESSION ANALYSIS FOR GROUP TESTING DATA WITH COVARIATE MEASUREMENT ERROR

Xianzheng Huang*, University of South Carolina
Joshua M. Tebbs, University of South Carolina

When modeling the prevalence of a rare trait, the use of group testing can provide significant reduction in cost. To incorporate the effects of covariates on the underlying prevalence in group-testing settings, Vansteelandt et al. (2000, Biometrics) consider a family of binary regression models. In this talk, we extend this regression approach to the case wherein covariates are measured with error. In the presence of measurement error and when the model for the unobservable covariate is misspecified, we demonstrate that the regression parameter estimates based on group-testing data are less sensitive to measurement error than those based on individual-testing data.

email: huang@stat.sc.edu

GENERALIZED RIDGE REGRESSION MODELS FOR ESTIMATING A TREATMENT EFFECT USING SURROGATE MARKER DATA

Yun Li*, University of Michigan
Jeremy M.G. Taylor, University of Michigan
Roderick J.A. Little, University of Michigan

Surrogate markers are potentially useful for predicting the treatment effect on the true endpoint in settings where the true endpoint is rare or costly to obtain. As defined by Prentice (1989), a perfect surrogate fully captures the treatment effect on the true endpoint, while a partial surrogate captures some of the treatment effect. We examine the use of a surrogate marker in increasing efficiency of the treatment effect estimator when the true endpoint is partially missing. When the perfect surrogacy assumption holds, use of the marker results in a substantial efficiency gain; when it fails, the resulting estimator is often seriously biased. When a marker is treated as a partial surrogate in a fixed-effects linear regression model, the efficiency gain is generally small. Hence, there is a trade-off between efficiency gain and bias, depending on whether partial or perfect surrogacy is assumed. We propose a generalized ridge regression approach that compromises between the perfect and partial surrogacy models. In simulations, the proposed method retains most of the efficiency gain when the marker is close to a perfect surrogate, while limiting the bias that results when the marker is incorrectly assumed to be perfect.

email: yunlisph@umich.edu

A COMPARISON OF METHODS FOR CORRELATION ANALYSIS OF BIOMARKER DATA WHEN SOME OBSERVATIONS ARE BELOW THE ANALYTIC LIMIT OF DETECTION

Hongwei Wang, Louisiana State University Health Sciences Center
Stephen W. Looney*, Medical College of Georgia
Siuli Mukhopadhyay, Medical College of Georgia

A problem frequently encountered in the correlation analysis of data from two biomarkers is that there are samples for which the concentration of one biomarker or the other is below the analytic limit of detection (LOD). These samples have been handled in a variety of ways in the biomarker literature; a common method is to replace the missing value with either LOD or LOD/2 and then calculate the correlation. Others have simply ignored values below the LOD when performing correlational analyses of the data. In this presentation, we review these and other methods that have been used in dealing with observations that are below the LOD and present results of a simulation study examining the performance of each of the various methods. Recommendations are made concerning preferred methods for handling observations below the LOD when correlating biomarker data.

email: slooney@mcg.edu

ESTIMATING THE CAUSAL EFFECT OF RACE ON STROKE OUTCOME

Megan E. Price*, Emory University
Vicki S. Hertzberg, Emory University
Kerrie Krompf, Emory University
Michael R. Frankel, Emory University

Background: In non-randomized studies, direct attribution of causality to a given factor is not possible. A propensity score, the probability of factor status conditional on observed covariates, can be used as an additional covariate in regression analyses, allowing for the causal inference of the factor's effect. Objective: To determine effect of race using propensity scores in an observational study of patients hospitalized for stroke. Methods: The effect of race was considered on the change in Barthel Index (BI) from hospital discharge to 90 day follow-up. Propensity scores were calculated and subsequently included in linear regression models estimating the difference in BI change, whites versus blacks. These results were compared to traditional linear regression models controlling for covariates. Results: Covariates identified by traditional model building were discharge Rankin score and smoking status with effect size estimated as -13.4. Covariates identified via the propensity score method included those covariates as well as medically indigent status, diabetes, low density lipoprotein, and peripheral vascular disease, with effect size estimated as -16.9. Conclusion: Propensity score analyses do not change the observed relationship between outcome and factor. However, estimates of effect size are increased, and reveal additional mediating covariates that were not identified using traditional tools.

email: meprice@emory.edu

BAYESIAN APPROACH FOR ANALYZING CLUSTER RANDOMIZED TRIAL AND ADJUSTING FOR MISCLASSIFICATION IN GLMM

Dianxu Ren*, University of Pittsburgh

Bayesian hierarchial modelling techniques have some advantages over classic methods for the analysis of cluster randomzied trial. Bayesian approach is also becoming more popular to deal with measurement error and misclassification problems. We propose a Bayesian approach to analyze a cluster randomized trial with adjusting for misclassification in a binary covariate in the random effect logistic model when a gold standard is not available. This Markov Chain Monte Carlo (MCMC) approach uses two imperfect measures of a dichotomous exposure under the assumptions of conditional independence and non-differential misclassification. Both simulated numerical example and real clinical example are given to illustrate the proposed approach. The Bayesian approach has great potential to be used in misclassification problem in Generalized Linear Mixed Model (GLMM) since it allow us to fit complex models and identify all the parameters. Our results suggest that Bayesian approach for analyzing cluster randomized trial and adjusting for misclassification in GLMM is flexible and powerful.

email: dir8@pitt.edu

BAYESIAN MODEL AVERAGING IN LATENT VARIABLE MODELS

Benjamin R. Saville*, University of North Carolina at Chapel Hill
Amy H. Herring, University of North Carolina at Chapel Hill

Latent variables are used in a variety of settings to model variables that are not directly observed. Given the unobservable nature of latent variable models, model selection strategies with latent variables can be particularly challenging. The number of latent variables may be unknown, or there may be uncertainty about the relationships among the latent variables or the distribution of latent variables. Common approaches to model selection such as likelihood ratio tests and deviance tests are often not applicable. Bayesian model averaging provides an attractive alternative to model selection strategies, but can be particularly challenging in latent variable models. Issues include obtaining accurate posterior model probabilities as well as defining effects that can be averaged across different models. Latent variables are not always measuring the same effects, nor are they always measured on the same scale. In addition, constraints on parameters can vary across models in order to retain identifiability. We propose strategies for implementing Bayesian model averaging in latent variable models and apply them to a study of drinking water disinfection by-products and pregnancy outcomes.

email: bsaville@bios.unc.edu

INFERENCE AND SENSITIVITY ANALYSIS FOR THE MALARIA ATTRIBUTABLE FRACTION

Dylan Small*, University of Pennsylvania

The cardinal symptom of malaria is fever. The malaria attributable fraction (MAF) in an area is the proportion of fevers in an area that are attributable to malaria. The MAF is an important epidemiological quantity for measuring the burden of malaria. The difficulty in estimating the MAF is that many other diseases besides malaria cause fever and children living in areas of high malaria endemicity often tolerate malaria parasites without developing any signs of disease; consequently, a fever episode may not be attributable to malaria even if the child has malaria parasites in his or her blood. We present a potential outcomes framework for the MAF, and analyze previously proposed estimators for the MAF under this framework. We show that previously proposed estimators depend on an assumption that parasite levels among individuals are effectively randomly assigned, and present evidence that this assumption does not hold for a data set. We develop a sensitivity analysis that assesses the sensitivity of inferences to departures from a random assignment of parasites assumption.

email: dsmall@wharton.upenn.edu

TESTING THE MEDIATING EFFECT IN MEDIATIONAL MODELS WITH MULTIPLE OUTCOMES

Kang Sun*, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh
Wesley Thompson, University of Pittsburgh
Patricia R. Houck, University of Pittsburgh Medical Center

Mediational analysis is used to explain how an antecedent predictor affects the consequent outcome through an intervening variable called a mediator. The assessment of mediational status is often accomplished through the use of Sobel and Clogg tests, which are the standard tools in a single outcome scenario. We have extended these tests to mediational models with multiple outcomes. The extensions include both parametric methods for assessing mediational status and, additionally, a bootstrapping approach. Using Monte Carlo simulations with two outcomes and changing sample and effect sizes systematically, Type I error rates and statistical powers of these inferential procedures are compared. Results show that in the presence of moderate correlation between the predictor and the mediator, the extended Clogg parametric test has the most accurate type I error rate and the highest power. Sample size and power estimates for varying strengths of correlation among the predictor and the mediator are presented. Implications of these results in clinical research are discussed.

email: kas34@pitt.edu

PREDICTING THE TREATMENT EFFECT FROM A SURROGATE MARKER USING A POTENTIAL OUTCOMES APPROACH

Jeremy MG Taylor*, University of Michigan
Yun Li, University of Michigan
Michael Elliott, University of Michigan

A possible use of a surrogate marker is to predict the actual treatment effect on the true endpoint. An effective surrogate marker occurs along the causal pathway between treatment and the true endpoint and hence can be used for prediction of the treatment effect. To assess the degree to which the treatment effect occurs through a surrogate marker, Frangakis and Rubin (2002) offered a criterion for surrogacy evaluation based on principal stratification and principal causal effects. When the surrogate marker, true endpoint and treatment assignment are binary, we carry out the evaluation by estimating the probabilities associated with the combinations of different sequences of potential outcomes for the surrogate marker and the true endpoint. To address nonidentifiability, we explore the role of assumptions that are plausible in the context of a surrogate marker in a clinical trial. Estimation is carried out using Bayesian methods. The assumptions are incorporated into the prior distributions. The estimated probabilities from the counterfactual model are later used to calculate the causal treatment effect in situations where the surrogate marker is fully observed but the true endpoints is partially missing. We extend the approach to multiple trial settings using hierarchical modeling.

email: jmgt@umich.edu

THE EFFECT OF INTRAVENOUS LEVOCARNITINE AMONG HEMODIALYSIS PATIENTS: MARGINAL STRUCTURAL MODELING IN A VERY LARGE SAMPLE

Eric D. Weinhandl*, Minneapolis Medical Research Foundation
David T. Gilbertson, Minneapolis Medical Research Foundation
Allan J. Collins, Minneapolis Medical Research Foundation

Marginal structural modeling (MSM) permits estimation of causal effects when exposure is highly confounded by time-dependent factors. MSM depends upon assumptions that are difficult to satisfy, and requisite calculations that can be computationally expensive. We applied MSM to a series of retrospective cohorts of hemodialysis patients, to estimate the causal effect of intravenous levocarnitine upon subsequent inpatient days per month. We encountered multiple difficulties. Lack of clinical knowledge of an appropriate parameterization for the counterfactual model of levocarnitine necessitated extensive exploratory work. However, large sample sizes (N = 120-150,000, with 12 observations per subject), high dimensional estimation spaces for both exposure and outcome (Dim = 200-300), and the utilization of generalized estimating equations created significant computational challenges. Finally, knowledge that the Experimental Treatment Assignment (ETA) assumption had been violated in clinical settings necessitated further data inspection. We present novel algorithms for empirical selection of predictors of exposure, for reduction of computational demands, and for ETA sensitivity analyses.

email: eweinhandl@cdrg.org

A COMPARISON OF MISSING DATA METHODS FOR SF36 QUALITY OF LIFE DATA

Liping Zhao*, Christiana Care Health System
Paul Kolm, Christiana Care Health System
William S. Weintraub, Christiana Care Health System

There is a growing literature of methods for analysis of non-ignorable missing data. The choice of methods is not easily made because of the assumptions made and/or difficulty of implementing the methods. In this study, we analyze longitudinal SF36 quality of life data of mitral valve (MV) repair and MV replacement patients before surgery, and 1, 3 and 12 months following surgery using different methods appropriate to the assumptions of MAR or MNAR. Methods include complete case, last value carried forward, maximum likelihood, pattern mixture, shared parameter and selection models. Results indicate differences in parameter estimates among the methods. Because assumptions regarding missing data mechanisms are usually not testable, we conclude that the different methods serve as sensitivity analyses.

email: lzhao@christianacare.org

STATISTICAL ANALYSIS OF STANDARDIZED UPTAKE VALUES WITH NEGATIVE PET SCANS

Qin Zhou*, Memorial Sloan-Kettering Cancer Center
Richard Wong, Memorial Sloan-Kettering Cancer Center
Steven M. Larson, Memorial Sloan-Kettering Cancer Center
Mithat Gönen, Memorial Sloan-Kettering Cancer Center

Positron emission tomography (PET) is now a popular tool in the diagnosis of cancers and many other diseases. Standardized uptake value (SUV) is a quantitative measure of the tracer uptake in PET scans and commonly used to summarize the results. SUV is measured on a region of interest (ROI), an area marked on the scan by the radiologist. In negative scans, it is difficult to choose an ROI and for expedience, SUVs are not reported by most investigators, thus generating missing values and creating a challenge in statistical analysis of such data. In this study, several statistical methods are applied for tackling this problem. The explored methods include case deletion, single value imputation, single imputation and multiple imputation. We also consider imputation from an external reference distribution derived from specific normal organs. A data set of head/neck squamous cell carcinoma is used as an example for comparing these imputing methods and demonstrating their advantages and disadvantages.

email: zhouq@mskcc.org

SURVIVAL ANALYSIS

EVALUATION OF RECURRENT EVENT ANALYSES IN PEDIATRIC FIREARM VICTIMS' EMERGENCY DEPARTMENT VISITS

Hyun J. Lim*, University of Saskatchewan-Canada
Marlene Melzer-Lange, Medical College of Wisconsin

The evaluations of covariate effects for the recurrent events data in longitudinal studies have received considerable attention. We review the existing models - multiple failure time models and frailty models, to estimate the relative risks of recurrences in a given dataset. We use simulated data to illustrate applicability of the models. Relevance and applicability of the models and interpretation of the estimates of the relative risk for each recurrence in the pediatric firearm victims emergency department visit database will also be investigated.

email: hyun.lim@usask.ca

COMPARING TREATMENTS FOR TWIN-TWIN TRANSFUSION SYNDROME: AN APPLICATION OF SURVIVAL ANALYSIS

David M. Shera*, The Children's Hosptial of Philadelphia and The University of Pennsylvania
Timothy Crombleholme, Cincinatti Children's Hospital

In any study involving prenatal treatment of twins, there are three individuals, the mother, and each of the twins, with separate but potentially correlated outcomes. In the case of twin-twin transfusion syndrome, the biology is asymmetric and so the prospects for outcome, primarily survival, are not exchangeable. Furthermore, a simple analysis of success/failure does not account for the timing of events. This report estimates the the association among outcomes by using time varying covariates for co-twin status in a survival anlysis. Results will inform future treatment decisions and study designs.

email: shera@email.chop.edu

VARIABLE SUBSET SELECTION AND MODEL SELECTION

A SIMULATION STUDY OF A MODEL SELECTION PROCEDURE FOR NONLINEAR LOGISTIC REGRESSION

Scott W. Keith*, University of Alabama at Birmingham
David B. Allison, University of Alabama at Birmingham

Splines are useful and exceptionally flexible tools for modeling nonlinear relationships in study variables. Depending on the complexity of the splines used, they are capable of underfitting, which results in excessive residual variance, or overfitting, which results in excessive bias. Many methods have been developed to control and optimize spline model fitting including penalized splines, generalized additive models, and free-knot splines. We have developed a model selection procedure for adjusting the complexity of free-knot splines. This involves fitting a free-knot spline to the data by nonlinear least squares first with k-1 and then with k knots and using our parametric bootstrap test of significant contribution to model fit from adding the kth knot and adjoining spline segment. Our objective is to evaluate the properties of our model selection methods by conducting a simulation study. By simulating data under a wide variety of conditions we will determine if, and under which conditions, the procedure works well. We will focus on simulating design matrices and modeling simulated binary outcome data conditional on having a nonlinear relationship with at least one independent variable. The method will be assessed in each simulated dataset by comparing the selected nonlinear logistic regression model to the true model.

email: swkeith@uab.edu

VARIABLE SELECTION IN CLUSTERING VIA DIRICHLET PROCESS MIXTURE MODELS

Sinae Kim*, The University of Michigan
Mahlet G. Tadesse, University of Pennsylvania School of Medicine
Marina Vannucci, Texas A&M University

The increased collection of high-dimensional data in various fields has raised a strong interest in clustering algorithms and variable selection procedures. In this poster, I propose a model-based method that addresses the two problems simultaneously. I use Dirichlet process mixture models to define the cluster structure and introduce in the model a latent binary vector to identify discriminating variables. I update the variable selection index using a Metropolis algorithm and obtain inference on the cluster structure via a split-merge Markov chain Monte Carlo technique. I evaluate the method on simulated data and illustrate an application with a DNA microarray study.

email: sinae@umich.edu

IDENTIFYING BIOMARKERS FROM MASS SPECTROMETRY DATA WITH ORDINAL OUTCOMES

Deukwoo Kwon*, National Cancer Institute
Mahlet G. Tadesse, University of Pennsylvania
Naijun Sha, University of Texas at El Paso
Ruth M. Pfeiffer, National Cancer Institute
Marina Vannucci, Texas A&M University

In recent years, there has been an increased interest in using protein mass spectroscopy to identify molecular markers to discriminate diseased from healthy individuals. Existing methods are tailored towards classifying observation into nominal categories. Sometimes, however, the outcome of interest may be measured ordered scale. Ignoring this natural ordering may result in loss of information. In this paper, we propose a Bayesian model for the analysis of mass spectrometry data with ordered outcome. The method provides a unified approach for identifying relevant markers and predicting class membership. This is accomplished by building a stochastic search variable selection method with an ordinal outcome model. We apply the methodology to mass spectrometry data on ovarian cancer cases and healthy individuals. We also utilize wavelet-based techniques to denoise the mass spectra prior to analysis. We identified protein markers associated with being healthy, a low grade case, or a high grade case. For comparison, we repeated the analysis using conventional classification procedures and found improved predictive accuracy with our method.

email: kwonde@mail.nih.gov

REGRESSIONS BY ENHANCED LEAPS-AND-BOUNDS VIA ADDITIONAL OPTIMALITY TESTS (LBOT)

Xuelei (Sherry) Ni*, Kennesaw State University
Xiaoming Huo, Georgia Institute of Technology

In exhaustive subset selection in regressions, the leaps-and-bounds algorithm by Furnival and Wilson (1974) is the current state-of-the-art. It utilizes a branch and bound strategy. We improve it by introducing newly designed optimality tests, retaining the original general framework. Being compared with the original leaps-and-bounds algorithm, the proposed method further reduces the number of subsets that are needed to be considered in the exhaustive subset search. Simulations demonstrate the improvements in numerical performance. Our new description of the leaps-and-bounds algorithm, which is based on our newly designed pair tree, is independent of programming languages, and therefore is more accessible.

email: xni2@kennesaw.edu

VARIABLE SELECTION AND ESTIMATION IN THE PARTIALLY LINEAR AFT MODEL WITH HIGH-DIMENSIONAL COVARIATES

Huaming Tan*, The University of Iowa
Jian Huang, The University of Iowa

We consider two regularization approaches, the LASSO and the threshold gradient directed regularization, for variable selection and estimation in the partially linear accelerated failure time (PL-AFT) model with high-dimensional covariates. The PL-AFT model has two regression components: a linear component for high-dimensional covariates such as gene expression data, and a nonparametric component for other low-dimensional covariates. Our study is motivated by studies that investigate the relationship between survival and genomic measurements and other variables such as clinical or environmental covariates. To obtain unbiased estimates of genomic effects, it is necessary to take into account these covariates, whose effects on the survival can be highly nonlinear and are often best to be modeled in a nonparametric way. We use the Stute's weighted least squares method to construct the loss function, which uses the Kaplan-Meier weights to account for censoring. The weighted least squares loss function makes the adaptation of this approach to high-dimensional settings computationally feasible. We use V-fold cross validation for tuning parameter selection. The proposed methods are evaluated using simulations and demonstrated on a real data example.

email: huaming-tan@uiowa.edu

NOTES

ENAR WEBSITE

WWW.ENAR.ORG

ABSTRACTS

2. DYNAMIC NETWORK MODELS

ANALYZING BRAIN NETWORKS WITH GRANGER CAUSALITY

Mingzhou Ding*, University of Florida

Commonly used interdependency measures such as cross correlation and spectral coherence do not yield directional information. Phase spectra may be used for that purpose only under very ideal conditions. Recent work has begun to explore the use of causal measures to further dissect the interaction patterns among neural signals. In this talk I will describe the concept of Granger Causality and introduce Geweke's causality spectra. The technique will then be applied to the analysis of multichannel local field potentials recorded from behaving monkeys performing sensorimotor and selective attention tasks.

email: mding@bme.ufl.edu

UNDERSTANDING PROTEIN FUNCTION ON A GENOME-SCALE USING NETWORKS

Mark Gerstein*, Yale University

My talk will be concerned with topics in proteomics, in particular predicting protein function on a genomic scale. We approach this through the prediction and analysis of biological networks -- both of protein-protein interactions and transcription-factor-target relationships. I will describe how these networks can be determined through integration of many genomic features and how they can be analyzed in terms of various simple topological statistics. I will discuss the accuracy of various reconstructed quantities. Further information at http://topnet.gersteinlab.org

email: Mark.Gerstein@yale.edu

GRAPHICAL MODELS FOR TEMPORAL DATA

Paola Sebastiani*, Boston University

Modelling longitudinal data is becoming an important task in today biomedical research. The advance of technology allows for the parallel measurement of gene products such as RNA abundance and gives the opportunity to dissect a biological system by observing its performance over time. An important task is to model gene products measured over time. This talk will describe the use of graphical models for this task. The focus will be on Dynamic Bayesian Networks and the issue of modelling relations from data that are not normally distributed and do not have a linear dependency structure using approximate learning algorithms and stochastic algorithms for inference.

email: sebas@bu.edu

NETWORK BICLUSTERING: IDENTIFY CONDITION-SPECIFIC NETWORK MODULES ACROSS MASSIVE BIOLOGICAL NETWORKS

Haiyan Hu, University of Southern California
Yu Huang, University of Southern California
Xianghong J. Zhou*, University of Southern California

Current approaches for biological network analysis focus mainly on the analysis of a single biological network, providing static descriptions of network features. With the rapid accumulation of biological networks generated under different conditions, there is an urgent need to analyze the dynamics of biological networks, i.e. how network features change with conditions. Such dynamic descriptions will provide deeper insights into cellular organization and pathway coordination. We developed a novel algorithm, NETBICLUSTER, to identify condition-specific network modules from a large number of biological networks. The NETBICLUSTER algorithm not only identifies network modules corresponding to particular biological processes, but also mines non-dense network subgraph patterns and more importantly, provides network dynamics information such as condition-specific activation and pathway cross-talking. Applying NETBICLUSTER to 97 mouse gene co-expression networks derived from microarray datasets, we discovered a large number of functionally homogeneous network modules and dataset conditions under which they are activated and deactivated.

email: xjzhou@usc.edu

3. GENETICAL GENOMICS: COMBINING EXPRESSION AND ALLELIC VARIATION DATA FOR COMPLEX DISEASES

DIMENSION REDUCTION METHODS IN THE STUDY OF THE GENETICS OF GENE EXPRESSION

Stephanie A. Monks*, Oklahoma State University
Qiang Guo, Oklahoma State University
Kathy Hanford, University of Nebraska

Several groups have recently suggested a novel approach for combining many types of genomic information in order to better understand the underlying causes of disease. These papers provided the initial data needed to determine if such an approach is feasible along with surveys of the genetic effects on the expression of genes (how genes affect other genes) in yeast, mouse, humans and corn. One of the challenges of such an approach is how to study the genetics of a vast set of highly interrelated measures which likely represent a much smaller set of truly meaningful responses. In other words, how do we separate the wheat from the chaff to find those genetic “kernels” that impact disease and quantitative traits. The objective of this proposal is to use currently available data to apply several statistical methods of dimension reduction as a means for reducing the number of phenotypes to be tested for linkage.

email: stephanie.monks@okstate.edu

COMBINED CORRELATION, LINKAGE, AND ENRICHMENT ANALYSIS IN EQTL STUDIES

Christina Kendziorski*, University of Wisconsin

A general goal common to many expression QTL (eQTL) mapping studies is inferring the genetic basis of gene function. The identification of genes coordinately regulated across a series of experimental conditions can provide valuable information about function; however, correlation alone does not provide a basis for localizing common regulatory regions. For this, genetic information is required. We have developed statistical methods that combine correlation, genetic linkage, and annotation analysis in experimental crosses to identify regions controlling shared functions. We illustrate the approach with an application to an F2 popluation segregating for diabetes. Using the combined analysis, we are able to identify the function of previously uncharacterized genes, identify novel members of known pathways, and predict master regulatory regions. This is joint work with Brian Yandell and Alan Attie.

email: kendzior@biostat.wisc.edu

USING WEIGHTED GENE CO-EXPRESSION NETWORKS FOR INTEGRATING GENE EXPRESSION, GENETIC MARKER AND COMPLEX TRAIT DATA

Steve Horvath*, University of California-Los Angeles

One promising approach to understand the genetics of complex traits involves the integration of genetic marker data with gene expression data. Here, we describe how weighted gene co-expression network analysis can be used to study the genetics of complex traits. To construct a weighted gene co-expression network on the basis of microarray data, we apply a soft thresholding approach to the absolute value of the Pearson correlation matrix. We define network modules as clusters of genes with high topological overlap. Instead of focusing on individual genes, a module based analysis provides a systems-level view of genes related to a complex trait. We use genetic marker data to identify chromosomal loci (referred to as module quantitative trait loci, mQTL) that perturb trait related modules. We find that intramodular network connectivity can be an important complementary screening variable for identifying complex trait related genes. Multivariate regression models allow one to integrate network concepts (intramodular connectivity) and genetic concepts (single point LOD scores, SNP significance) to determine the factors that affect the relationship between gene expression data and the complex trait. The resulting integrated approach suggests screening criteria for identifying the genetic drivers of the complex trait.

email: shorvath@mednet.ucla.edu

4. ITEM RESPONSE THEORY

NONPARAMETRIC BAYESIAN ITEM RESPONSE THEORY

Kristin A. Duncan, San Diego State University
Steven N. MacEachern*, The Ohio State University

Item response theory is a body of research directed at the assesment of an undelying ability or achievement level. In its simplest form, it allows one to use the results of a multiple choice exam both to rank-order a batch of candidates and to decide which candidates have mastered the material covered by the exam. In this talk, we will present a nonparametric Bayesian approach to item response theory. Under this approach, a Dirichlet process prior distribution is placed on each item characteristic curve. The resulting model has full support among models for which there is a one-dimensional trait underlying exam performance. Features of the model will be described. Results from fitting the new model and traditional, parametric models to results from an undergraduate statistics exam will be presented and contrasted. We find a clear need to fit the nonparametric model.

email: snm@stat.ohio-state.edu

DIFFERENTIAL ITEM FUNCTIONING IN A GRADED RESPONSE IRT MODEL: A BAYESIAN APPROACH TO ITEM DISCRIMINATION

Mark E. Glickman*, Boston University
Susan Eisen, Boston University
Pradipta Seal, Boston University

Mental health survey instruments are often useful at diagnosing and summarizing well-being of respondents. A typical survey involves respondents evaluating themselves on a number of items via a set of ordinal choices. If, however, certain subgroups of patients with the same mental health status give systematically different responses to certain items, the instrument is said to exhibit differential item functioning (DIF). In this talk, we develop a Bayesian approach to identifying differences in item response patterns with particular application to Samejima's (1969) graded response model. Our approach extends the usual item-response theory (IRT) models by including item-specific parameters to measure the extent of DIF, and allows for the incorporation of baseline covariate information. We apply our approach to the analysis of responses by mental health patients from the BASIS-24, a widely-used self-report mental health assessment instrument, to study differences among different cultural and language groups in item response patterns.

email: mg@bu.edu

A MARKOV CHAIN MONTE CARLO APPROACH TO CONFIRMATORY ITEM FACTOR ANALYSIS

Michael C. Edwards*, The Ohio State University

Item factor analysis has a rich tradition in both the structural equation modeling and item response theory frameworks. While great strides have been made in the past three decades in parameter estimation for these types of models, significant limitations remain. The goal of this paper is to examine the feasibility of using Markov chain Monte Carlo (MCMC) estimation methods to estimate parameters of a wide variety of confirmatory item factor analysis models. After providing a brief overview of item factor analysis and MCMC, results from a small feasibility study will be discussed. Where possible comparisons are made to estimates from "gold-standard" estimators such as WLS and MML/EM, but the bulk of the examples focus on models that are problematic for these estimators. In all cases MCMC provided reasonable parameter estimates. Future directions, including software development, model fit, and model extensions, will also be addressed.

email: edwards.134@osu.edu

5. MULTIPLICITY AND REPRODUCIBILITY IN SCIENTIFIC STUDIES: RESULTS FROM A SAMSI WORKSHOP

IDENTIFYING MEANINGFUL PATIENT SUBGROUPS

Robert L. Obenchain*, Eli Lilly and Company

The Local Control (LC) approach to estimation of treatment effects on humans is based upon a treatment-within-cluster nested ANOVA model that is frequently less restrictive and more robust than traditional Covariate Adjustment (CA) models. Because LC systematically forms subgroups, compares subgroups and (following overshooting) recombines subgroups, LC proceeds via built-in sensitivity analyses. Specifically, the analyst views graphical displays that identify the most effective treatment in each patient subgroup within a family generated by varying the number of clusters, the clustering metric and/or the clustering (unsupervised learning) algorithm. The LC approach uses generalized definitions of Treatment Effects, TEs, and main effects. TEs are distributions of local main effects, and the overall Main Effect (ME) of treatment is the unknown true mean of all such local TE distributions. It is then natural to ask whether [a] an observed TE distribution could be a mixture of two or more well defined sub-distributions and whether [b] it is possible to predict the numerical size or sign of local main effects directly from the baseline patient X-characteristics used to define clusters.

email: ochain@lilly.com

SOME ASPECTS OF MULTIPLE TESTING

Susie Bayarri*, University of Valencia and SAMSI
James Berger, SAMSI and Duke University

The main purpose of this talk is to review, from a Bayesian perspective, recent and popular methods to address the problem of multiple testing. We consider both frequentist and classical methods, as well as eclectic ones that seem to combine both approaches; in particular, we look at a variety of False Discover Rate approaches. The main goal is to gain understanding of these methods from a Bayesian perspective. We also consider these methods from a decision-theoretic perspective. For instance, a key question is whether FDR can arise naturally in a decision setting. This talk summarizes a variety of issues about multiple testing discussed during the SAMSI Summer Program on Multiplicity and Replicability in Scientific Studies.

email: susie.bayarri@uv.es

SUBGROUP ANALYSIS - A STYLIZED BAYES APPROACH

Siva Sivaganesan*, University of Cincinnati
Prakash Laud, Medical College of Wisconsin
Peter Müller, The University of Texas-M.D. Anderson Cancer Center

We introduce a new approach to inference for subgroups in clinical trials. The main elements of the proposed approach are the use of a priority ordering on covariates to define potential subgroups and the use of posterior probabilities to identify subgroup effects for reporting. As usual in Bayesian clinical trial design we compute frequentist operating characteristics (OC). We achieve desired OCs by obtaining a suitable threshold for the posterior probabilities.

email: sivagas@ucmail.uc.edu

BAYESIAN DECISION THEORY FOR MULTIPLICITIES

Kenneth M. Rice*, University of Washington

Many Bayesian analyses conclude with a summary of the posterior distribution, thus summarizing uncertainty about parameters of interest. But this is only half the story; it neglects to state what it is about the parameters that we want actually want to know. Formally, deciding our criteria for a 'good' answer defines a loss function, or utility, and is usually only considered for point estimation problems. For interval estimation, we provide sensible, interpretable loss functions which formally justify some 'standard' but essentially ad-hoc Bayesian intervals. Developing these measures of utility for problems with multiple parameters is straightforward, and the Bayes rules remain attractive and simple. Direct connections can be made with frequentist methods of multiplicity-adjustment.

email: kenrice@u.washington.edu

6. RECENT DEVELOPMENTS IN BAYESIAN SURVIVAL ANALYSIS

PRIOR ELICITATION AND VARIABLE SELECTION IN REGRESSION MODELS WITH HIGH DIMENSIONAL DATA

Joseph G. Ibrahim*, University of North Carolina at Chapel Hill
Mayetri Gupta, University of North Carolina at Chapel Hill

One of the most important modern day challenges in analyzing high dimensional data (such as microarray data) jointly with continuous, dsicrete, or time-to-event data is that one immediately encounters the $p > n$ problem, in which the number of covariates in the regression model greatly exceeds the number of subjects. In this talk, we develop a methodology for the specification of a class of prior distributions for generalized linear models and survival models that accommodate the $p > n$ paradigm. The class of proper prior distributions are based on the generalization of the g-prior as well as having a 'ridge parameter' that facilitates propriety for $p > n$. The resulting prior has the flavor and operating characteristics of ridge regression. Various properties of the prior are discussed and a real dataset is analyzed to illustrate the proposed methodology.

email: ibrahim@bios.unc.edu

JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA USING MIXTURES OF POLYA TREES

Timothy Hanson, University of Minnesota
Adam Branscum, University of Kentucky
Wesley Johnson*, University of California-Irvine

We discuss a new approach to jointly modeling survival data with time dependent covariates. Historically, time dependent covariates (TDC} were not modeled but rather analysis proceeded with them regarded as fixed. However, this type of analysis was based on the premise that the TDC's did not vary between observational time points. If TDC's are monitored frequently enough, the necessity to model them is diminished. We develop and discuss the problem assuming the need for such modeling and allow baseline distributions to be drawn from broad families in the context of the traditional Cox model, the proportional odds model and the Cox and Oakes model, all with time dependent covariates. We compare results based on these and parametric counterparts using a real data set.

email: wjohnson@uci.edu

LONGITUDINAL STUDIES WITH OUTCOME-DEPENDENT FOLLOW-UP: MODELS AND BAYESIAN REGRESSION

Bani Mallick*, Texas A&M University
Duchwan Ryu, Chase
Debajyoti Sinha, MUSC
Stuart Lipsitz, MUSC

We propose Bayesian parametric and semiparametric partially linear regression methods to analyze the `outcome-dependent follow-up'data when the random time of a follow-up measurement of an individual depends on the history of outcomes at the previous times. We begin with the investigation of the simplifying assumptions and extend the correlation structure of the errors by allowing subject-specific correlations. To accommodate both the longitudinal measurements and the follow-up times as responses of interest, we present a new model for analyzing such data by introducing a subject-specific latent variable. An extensive simulation study shows that a Bayesian partially linear regression method facilitates accurate estimation of the true regression line and the regression parameters. We illustrate our new methodology using data from a longitudinal observational study.

email: bmallick@stat.tamu.edu

7. RECONCILING DIFFERENCES IN ESTIMATION METHODS: A CASE STUDY OF MANATEE POPULATION DYNAMICS

POPULATION GROWTH RATE OF FLORIDA MANATEES: CONTRADICTORY ESTIMATES FROM INDEPENDENT METHODS

John E. Reynolds, III*, Mote Marine Laboratory
Michael C. Runge, USGS Patuxent Wildlife Research Center

Although Florida manatees have been statutorily protected for well over 100 years, management strategies for species recovery remain contentious. The fact that manatee habitat coincides with the preferred coastal real estate and boating corridors of a rapidly growing human population has led to calls from the public and managers for better science, especially regarding population size and trends, to justify regulatory actions. In 2004, two studies were published that used entirely independent methods (aerial surveys and a Bayesian hierarchical model; mark-recapture data and an age structured matrix model) to estimate the growth rate of the manatee population along the Atlantic coast. Whereas the estimates generated by the two approaches appear tantalizingly close, the distributions do not overlap much and the management implications of the two estimates are quite different. The existence of entirely independent methods to estimate a quantity relevant to management is rare. The tension between the results from these methods provides something even rarer: a test of assumptions and the opportunity to examine and improve both methods. Rather than compete in a game of dueling models, the authors of these studies have chosen to work together to reconcile their estimates.

email: reynolds@mote.org

BAYESIAN TREND ANALYSIS OF THE FLORIDA MANATEE ALONG THE ATLANTIC COAST VIA AERIAL SURVEYS

Bruce A. Craig*, Purdue University

In many animal population studies, the construction of a Bayesian hierarchical model provides an effective way to capture underlying biological and sampling characteristics which contribute to the overall variation in the data. In this talk, I will discuss such a model developed to assess the population trend of the Florida manatee from aerial survey data collected at winter aggregation sites. This model attempts to account for the method by which the manatees were counted, their possible movement between surveys, and the potential increase/decrease of the total population over time. Because a recent 2004 study that analyzed survey data between 1982 and 2001 resulted in more optimistic growth estimates compared to estimates based on mark-recapture studies, I will focus on the key mechanisms and potential biases of the model that could drive these differences. This includes the specification of priors and the omission of a key covariate in the observation or movement process.

email: bacraig@stat.purdue.edu

MARK-RECAPTURE ESTIMATES OF ADULT MANATEE SURVIVAL RATE: MODELING POSSIBLE SOURCES OF CONFLICT WITH AERIAL SURVEY ESTIMATES OF GROWTH

Catherine A. Langtimm*, USGS-Florida Integrated Science Center

Estimates of adult survival rates from mark-recapture analysis of photo-identification data of Florida manatees are key parameters in matrix models used to estimate population growth rate and to assess population status. Because estimates of growth rate from a Bayesian analysis of aerial count data do not coincide with estimates based on our mark-recapture studies, I focus on three issues that could drive the potential differences: 1) censoring the photo-sighting data to better represent the population sampled with aerial surveys, 2) modeling variation in temporary emigration that can bias survival estimates, and 3) addressing potential bias at the end of the time series due to limitations in model methodology.

email: clangtimm@usgs.gov

POPULATION GROWTH RATE OF FLORIDA MANATEES: RECONCILING ESTIMATES FROM INDEPENDENT METHODS

Michael C. Runge*, USGS Patuxent Wildlife Research Center
Bruce A. Craig, Purdue University
Catherine A. Langtimm, USGS Florida Integrated Science Center
John E. Reynolds, Mote Marine Laboratory

To understand the discrepant estimates of manatee population growth rate arising from two independent methods, we began by articulating the assumptions of each method and positing six hypotheses for the discrepancy: (1) a different sample frame for each method, coupled with immigration into the population sampled by aerial survey; (2) negative bias in estimates of adult survival rate from mark-recapture; (3) negative bias in reproductive or calf survival rates used in the matrix method; (4) transient effects in the age-structured population; (5) a positive temporal trend in sighting probability during the aerial survey; and (6) bias induced by the particular specification of the hierarchical model used to evaluate the aerial survey data. Recent analyses of adult survival have suggested that some of the discrepancy can be explained by negative bias in previous estimates, but we have not eliminated the possibility of other sources of bias. More detailed comparisons across a variety of time frames will allow us to investigate the circumstances under which the two methods produce comparable results. In this talk, we describe the process by which we sought to reconcile the results from the independent methods, as well as our plans for future integrated analysis of aerial survey and demographic mark-recapture data.

email: mrunge@usgs.gov

8. FINDING THE BEST DOSE IN CLINICAL TRIALS

FINDING AN ACCEPTABLE REGION IN CANCER DOSE-FINDING STUDIES MODELING BOTH TOXICITY AND EFFICACY

Bo Huang*, University of Wisconsin-Madison
Rick Chappell, University of Wisconsin-Madison

A new class of dose-finding designs in phase I/II clinical trials in cancer combining both toxicity and efficacy are presented. Instead of trying to find a single optimal dose, we would like to identify an acceptable region in which doses are both effective and safe, along with the optimal one within the region. For each cohort of three patients, the estimated minimum effective dose (MED), most successful and acceptable dose (MSAD), which maximizes the probability of joint efficacy and nontoxicity in the acceptable region, and maximum tolerated dose (MTD) are randomly assigned to them. Simulation results show that the new designs exhibit good operating characteristics and have competitive performance compared with the existing designs.

email: bohuang_uw@yahoo.com

ASSESSING THE PRECISION OF MED ESTIMATORS IN DOSE-RESPONSE STUDIES

Shanhong Guan*, University of Wisconsin-Madison
Jose C. Pinheiro, Novartis Pharmaceuticals
Frank Bretz, Novartis Pharma AG

Identifying and estimating target doses, such as the minimum effective dose (MED), is a key goal in the drug development. A unified approach, multiple comparison procedure and modeling (MCP-Mod), was proposed by Bretz, Pinheiro and Branson (2005) for designing and analyzing dose finding studies, including the estimation of the MED. To help the clinical team understand the accuracy of the results of a dose finding trial, it is of critical importance to include a measure of precision for the MED estimates. In this paper, we describe several bootstrap procedures for estimating the precision of the MED derived via MCP-Mod methodology, including the standard error and the 90% confidence intervals. These results are further compared with the asymptotic standard errors and associated confidence intervals. Simulation studies are conducted to evaluate the performance of boothstrap methods and asymptotic variance formula. Sample size calculations taking into account the desired precision of the MED estimates are also discussed and illustrated.

email: shanhong@stat.wisc.edu

ON IDENTIFYING MINIMUM EFFICACIOUS DOSES IN COMBINATION DRUG TRIALS

Julia N. Soulakova*, University of Nebraska-Lincoln
Allan R. Sampson, University of Pittsburgh

In the case of a combination drug, in addition to demonstrating safety and efficacy the FDA requires demonstrating that each component makes a contribution to the claimed effects. We consider a case of a combination compound with at least one component known to be effective at the considered doses. In this case any combination superior to each component is then also considered to be effective. We term such a combination as an efficacious combination. When administering higher doses of the drug can cause serious side effects, the goal is to identify the “lowest” (minimum) among efficacious doses (MeD’s). We introduce the notion of the MeD-set and propose a closed testing procedure for identifying the MeD-set. Our procedure requires first, identifying all possible patterns of the population MeD-set for a given design. Next, the closed hypothesis family is constructed and the proper step-down partial testing order is specified. Then, the hypotheses are tested using the closed testing scheme. The pattern of rejected and accepted hypotheses provides the estimated MeD-set. We present the results of a simulation study to describe the goodness of the estimation procedure and to characterize the population configurations when the procedure performs the best.

email: jsoulakova2@unlnotes.unl.edu

BAYESIAN DOSE-FINDING IN CLINICAL TRIALS FOR DRUG COMBINATIONS

Guosheng Yin*, The University of Texas-M. D. Anderson Cancer Center
Ying Yuan, The University of Texas-M. D. Anderson Cancer Center

There has been a growing trend of combining several different drugs to treat patients in clinical trials. We propose a Bayesian adaptive clinical trial design for dose-finding to account for the synergistic or anti-synergistic effect of two combined drugs. We focus on finding the maximum tolerated dose based on the toxicity outcomes from the sequentially accrued patients. We continuously update the posterior estimate for each combined dose toxicity probability based on the cumulated data of previously treated patients in the trial. We conduct extensive simulation studies to examine the operating characteristics of the proposed method under various practical scenarios. Finally, we illustrate our Bayesian dose-finding procedure with a recent clinical trial of drug combinations at M. D. Anderson Cancer Center.

email: gsyin@mdanderson.org

A STATISTICAL LOOK AT RELATIVE POTENCY IN A CROSSOVER DESIGN

Guoyong Jiang*, Cephalon, Inc.

The relative potency of an innovative drug to a standard therapy is often evaluated based on the parallel line assay in crossover dose-response studies. This presentation will address some statistical issues on the point and interval estimates of relative potency in crossover designs based on the parallel line assay. It will also address the statistical issue of determining parallelism or linearity between pairs of dose-response data sets. In addition, it will discuss the sample size and power estimations related to relative potency.

email: jjiang@cephalon.com

OPTIMIZATION IN MULTIVARIATE GENERALIZED LINEAR MODELS

Siuli Mukhopadhyay*, Medical College of Georgia
Andre I. Khuri, University of Florida

Combination drug therapy involves the use of more than one drug in order to provide greatest benefit in the treatment of a certain disease. One of the objectives of this therapeutic approach is the finding of the combined dose levels of the drugs which maximize their intended effects while minimizing any adverse side effects they may cause. In this article, we discuss the determination of the settings of the factors which simultaneously optimize several responses that can be represented by a multivariate generalized linear model (GLM). The generalized distance approach of Khuri and Conlon (1981), for the simultaneous optimization of several linear response surface models, is adapted to this multivariate GLM situation. An application of the proposed methodology is presented in the special case of a bivariate binary distribution resulting from a drug testing experiment concerning two responses, namely, efficacy and toxicity of a particular drug combination. This optimization procedure is used to find the dose levels of two drugs that simultaneously maximize their therapeutic effect and minimize their toxic effect.

email: smukhopadhyay@mcg.edu

A UNIFIED APPROACH TO PROOF OF ACTIVITY AND DOSE ESTIMATION FOR BINARY DATA

Bernhard Klingenberg*, Williams College

We suggest to unify dose-response modeling and target dose estimation into a single framework for the benefit of a more comprehensive and powerful analysis. Bretz, Pinheiro and Branson (Biometrics, 2006) implemented a similar idea for independent normal data by using optimal contrasts as a selection criterion among various candidate dose-response models. From a comprehensive set of candidate models a few are chosen that best pick up the dose-response. To decide which models, if any, significantly pick up the signal we construct the permutation distribution of the maximum penalized deviance statistic over the candidate set. This allows us to find critical values and multiplicity adjusted p-values, controlling the error rate of declaring spurious signals as significant. After settling on significant models, we use them to estimate doses worth of future investigation, such as the minimum effective dose. A thorough evaluation and comparison of our approach to popular contrast tests such as the Cochran-Armitage, Williams or Dunnett tests reveals that its power is as good or better in detecting a dose-response signal under a variety of situations, with many more additional benefits.

email: bklingen@williams.edu

9. MISSING DATA AND MEASUREMENT ERROR IN EPIDEMIOLOGY

ESTIMATION OF INCUBATION PERIODS DISTRIBUTION UNDER DIFFERENT SCENARIOS OF INFECTIOUS DISEASE OUTBREAK

Xiaojun You*, Johns Hopkins University
Ron Brookmeyer, Johns Hopkins University

The incubation period is an important quantity in developing effective epidemic control programs. When a chain of infection happens, a possible problem arises that we only have information about the times of symptom onset but not the times of initial infection. We developed statistical approaches for using epidemic chains of symptomatic cases to estimate the distribution of incubation periods. We used the geometric distribution to approximate the date of infection, and derived estimators of the distribution parameters. We also examined imputation methods. We performed simulation studies to compare the performance of the various estimators.

email: xyou@jhsph.edu

DOUBLE SAMPLING DESIGNS FOR ADDRESSING LOSS TO FOLLOW-UP IN ESTIMATING MORTALITY

Ming-Wen An*, Johns Hopkins University
Constantine E. Frangakis, Johns Hopkins University
Donald B. Rubin, Harvard University
Constantin T. Yiannoutsos, Indiana University School of Medicine

Loss to follow-up is an important challenge that arises when estimating mortality, and is of particular concern in developing countries. In the absence of more active follow-up systems, resulting mortality estimates may be biased. One design approach to address this is 'double sampling', where a subset of patients who are lost to follow-up is chosen to be actively followed, often subject to resource constraints. We discuss design and analysis to guide data collection and sampling procedures under such constraints, with the goal of obtaining valid and efficient estimators. We demonstrate our results using data from Africa, which were collected to estimate HIV mortality as part of an evaluation of the President's Emergency Plan for AIDS Relief (PEPFAR).

email: man@jhsph.edu

A COMBINED ANALYSIS OF MATCHED AND UNMATCHED CASE-CONTROL STUDIES: A LATENT GROUP APPROACH

Mulugeta G. Gebregziabher*, Medical University of South Carolina
Paulo Guimaraes, Medical University of South Carolina

Common in case-control studies is to use two or more different sets of controls to validate or confirm a positive or negative finding. This is usually done using separate analysis comparing cases with each unmatched or individually matched control group. But, when the parameter estimates from separate models for each pairing of cases and controls are not independent (for instance when there is common case group), it could result in inflated type II error for a homogeneity test and a large standard-error for the common pooled measure. The problem is compounded when one control group is matched and the other is unmatched. While there are available methods for combining matched and unmatched binary response case-control studies, there is no work for case-control studies with multiple control groups. Thus, we propose a unified latent group approach for testing homogeneity and for obtaining a pooled estimate that takes into account non-independence. A simulation study is used to evaluate the relative performance of the proposed method and a case-control study of multiple-myeloma risk and a SNP in the Inter-Leukin-6 is used to illustrate our findings. The proposed method gives better control of type II error and more precise estimate of the pooled measure.

email: gebregz@musc.edu

NONIGNORABLE MISSING DATA IN MATCHED CASE-CONTROL STUDY

Samiran Sinha*, Texas A&M University
Tapabrata Maiti, Iowa State University

We will talk about how to deal with informative missing (IM) exposure data in matched case-control studies. When the missingness mechanism depends on the unobserved exposure values, modeling the missing data mechanism is inevitable. Therefore a full likelihood based approach for handling IM data has been proposed by positing a model for selection probability, and a parametric model for the partially missing exposure variable among the control population along with a disease risk model. The model parameters are estimated via the EM algorithm. The method will be illustrated by real data example and simulation study.

email: sinha@stat.tamu.edu

ANALYSIS OF A DISEASE AND PROBABILITY OF EXPOSURE ASSOCIATION USING A REPLICATED ERROR-PRONE EXPOSURE ASSESSMENT

Chengxing Lu*, Emory University
Robert H. Lyles, Emory University

In environmental epidemiologic studies, it is common for a binary exposure to be assessed multiple times in a manner subject to misclassification. In a case-control setting, we focus on exploring the association between a disease and the probability of exposure given such replicates and in the absence of a gold standard for exposure. Assuming a beta distribution for the exposure probability, we obtain the estimated association by maximizing the marginal likelihood of the observed exposure replicates and the disease status. In simulation studies, we compare the performance of the proposed method with that of a logical but biased approach that replaces the unknown true exposure probability by the sample mean of the replicates. The proposed method is shown to be superior in terms of bias and confidence interval coverage. We present a real data example investigating the association between infants birth weight and the probability of parental occupational lead exposure, from the Baltimore Washington Infant Study (BWIS).

email: clu@emory.edu

IMPUTATION FOR MISSING CONTINUOUS OUTCOMES IN COMMUNITY INTERVENTION TRIALS

Monica Taljaard*, Ottawa Health Research Institute and University of Ottawa
Allan Donner, Schulich School of Medicine-University of Western Ontario and Robarts Clinical Trials
Neil Klar, Schulich School of Medicine-University of Western Ontario

We consider the problem of missing continuous outcomes in community intervention trials with a completely randomized design. The validity and power of inferences for the effectiveness of the intervention based on the adjusted two-sample t-test are compared under six different approaches for dealing with randomly missing outcomes, namely Complete Case Analysis and five different imputation strategies, including Cluster Mean imputation, parametric and non-parametric Multiple Imputation (MI) procedures which do not account for the intracluster correlation coefficient, and parametric MI accounting for intracluster correlation. We thus compare validity and power of tests when imputation is based on all the data pooled across communities in a group versus the data from each community only; and evaluate the consequences of failing to account for intracluster correlation during imputation. A simulation study is used to evaluate the role of cluster size, number of clusters, degree of intracluster correlation, and variability among cluster follow-up rates.

email: mtaljaard@ohri.ca

ACCOUNTING FOR ERROR DUE TO MISCLASSIFICATION OF EXPOSURES IN CASE-CONTROL STUDIES OF GENE-ENVIRONMENT INTERACTION

Li Zhang*, The Cleveland Clinic Foundation
Bhramar Mukherjee, University of Michigan
Malay Ghosh, University of Florida

We consider the analysis of an unmatched case-control study data with a binary genetic factor and a binary environmental exposure when both the genetic and the environmental exposures could be potentially misclassified. We devise an estimation strategy which corrects for misclassification errors and also exploits the gene-environment independence assumption. The proposed corrected point estimates and confidence intervals for misclassified data reduce back to standard analytic forms as the misclassification error rates go to zero. We illustrate the methods by simulating unmatched case-control datasets under varying association and misclassification scenarios.

email: zhangl3@ccf.org

10. LONGITUDINAL DATA, INCLUDING MISSING DATA AND MARKOV MODELS

TWO LOCAL INFLUENCE APPROACHES FOR TWO BINARY OUTCOMES WITH NON-MONOTONE MISSINGNESS

Caroline Beunckens*, Hasselt University
Cristina L. Sotto, Hasselt University
Geert Molenberghs, Hasselt University

Many models to analyze incomplete longitudinal data have been developed, which allow the missingness to be not at random (MNAR). Since such models rely on unverifiable modelling assumptions, research nowadays is devoted to assess the sensitivity thereof. A popular sensitivity tool is based on local influence (Cook, 1986), a technique which studies the effect of small perturbations around a given null model to detect subjects which have an influence on the analysis. Jansen et al (2003) developed a local-influence approach for binary data, subject to non-montone missingness, focussing on the model proposed by Baker, Rosenberger, and DerSimonian (1992). In this case perturbations of a given BRD model are considered in the direction of a model with one more parameter in which the original model is nested. Next to the influence of perturbations in the parameters of BRD models, perturbations in the observed cell counts arising from the contingency tables from two binary outcomes subject to non-monotone missingness, can also yield influence, e.g., on the complete data cell counts, and on important inferential quantities. Therefore, we derived influence measures functions of the model parameters in general, following the reasoning of Cook (1986). Both local influence approaches are applied to the Slovenian Public Opinion survey data.

email: caroline.beunckens@uhasselt.be

A SIMULATION STUDY COMPARING WEIGHTED ESTIMATING EQUATIONS WITH MULTIPLE IMPUTATION BASED ESTIMATING EQUATIONS

Caroline Beunckens, Hasselt University
Cristina L. Sotto*, Hasselt University
Geert Molenberghs, Hasselt University

Missingness frequently complicates the analysis of longitudinal data. A popular solution for dealing with incomplete longitudinal data is the use of likelihood-based methods, when, for example, linear, generalized linear, or non-linear mixed models are considered, due to their validity under the assumption of missing at random (MAR). Semi-parametric methods such as generalized estimating equations (GEE) offer another attractive approach but requires the assumption of missing completely at random (MCAR). Weighted GEE (WGEE) have been proposed as an elegant way to ensure validity under MAR. Alternatively, multiple imputation (MI) can be used to pre-process incomplete categorical/binary repeated measures, after which GEE is applied (MI-GEE). In this paper, we compare both using the so-called asymptotic, as well as small-sample, simulations, in a variety of correctly specified as well as incorrectly specified models. While WGEE often exhibits less bias asymptotically, the small-sample simulations show situations where the order is reversed. Arguably, MI-GEE is most useful when a variety of analyses are envisaged on the same set of incomplete data, one of the original motivations for MI, or when there is large uncertainty about the missingness mechanism.

email: cristina.sotto@uhasselt.be

A PATTERN MIXTURE MODEL WITH LEAST SQUARE SPLINES IN THE ANALYSIS OF LONGITUDINAL DATA WITH NON-IGNORABLE DROPOUT: RCT DATA FROM LIDODERM PATCH 5% FOR PAIN FROM OSTEOARTHRITIS OF THE KNEE

Qinfang Xiang*, Endo Pharmaceuticals
Suna Barlas, Endo Pharmaceuticals

Missing data are frequently encountered in the analysis of data based on randomized clinical trials. The problem is exacerbated in longitudinal studies, particularly those conducted in the areas of depression and pain. Missing data in these studies are usually of non-ignorable type and therefore pose a greater statistical challenge. Inappropriate handling of the missing data in the analyses can result in biased estimates and lead to incorrect conclusions. This is especially problematic when the dropout mechanism is a function of the unobserved outcome (non-ignorable missing). This study examines data based on a randomized trial comparing efficacy of Lidocaine Patch 5% with Celecoxib in patients suffering from pain associated with osteoarthritis of the knee. The study suffered from a significant amount of missing data (48%) that could not be assumed to be of ignorable type. A random-effects pattern mixture least squares spline model with fixed knots is applied to compare the two treatment groups in terms of the outcome of interest. This model is compared to those based on the last observation carried forward (LOCF), complete case analysis (CC), and mixed model repeated measurements (MMRM).

email: xiang.qinfang@endo.com

NON-DELETION APPROACH TO DETECT DISCORDANT SUBJECTS IN REPEATED MEASUREMENTS

Jungwon Mun*, University of Wisconsin-Madison

Many papers have proposed methods for detecting discordant subjects and observations for longitudinal or functional data. Most of these methods adapt existing methods for regression data to mixed effect models (mainly for linear models). This article suggests a new method to more effectively detect discordant subjects and observations and provides greater information by utilizing (revised) residuals. The proposed method is valid for linear and non-linear mixed effect models and enables the investigation of both subject-wise and observation-wise outliers.

email: stat.chris@gmail.com

TESTING FOR TRENDS IN A TWO-STATE MARKOV MODEL WITH APPLICATIONS IN SMOKING CESSATION STUDIES

Charles G. Minard*, The University of Texas M.D. Anderson Cancer Center and The University of Texas
Wenyaw Chan, The University of Texas
David Wetter, The University of Texas M.D. Anderson Cancer Center
Carol J. Etzel, The University of Texas M.D. Anderson Cancer Center

Intervention trials, such as studies on smoking cessation, may observe multiple, discrete outcomes over time. When the outcome is dichotomous, participant observations may alternate between two states (e.g. abstinent or relapsed) over the course of a study. The generalized estimating equations approach is commonly used to analyze binary, longitudinal data in the context of independent variables. Including an interaction term in the model between an independent variable and time allows for the evaluation of a trend in the outcome of interest. However, the sequence of observations may be assumed to follow a Markov chain with stationary transition probabilities when observations are made at fixed time points. Participants favoring the transition to one particular state over the other would provide evidence of a trend in the observations. Assuming a lognormal trend parameter, the determinants of a trend in binary, longitudinal data may be evaluated by maximizing the likelihood function. New methodology is presented here to test for the presence and determinants of a trend in binary, longitudinal observations. Empirical studies are evaluated, and comparisons are made with respect to the generalized estimating equations approach. Practical application of the proposed method is made to data available from a smoking cessation study.

email: cgminard@mdanderson.org

HIDDEN MARKOV MODELS FOR ALCOHOL DATA

Kenneth E. Shirley*, The Wharton School, University of Pennsylvania
Dylan Small, The Wharton School, University of Pennsylvania
Kevin Lynch, University of Pennsylvania

In a clinical trial of a treatment for alcoholism, the usual response variable of interest is the number of alcoholic drinks consumed by each subject each day. Subjects in these trials are typically volunteers who are alcoholics, and thus are prone to erratic drinking behaviors, often characterized by alternating periods of heavy drinking and abstinence. For this reason, many statistical models for time series that assume steady behavior over time and white noise errors do not fit alcohol data well. In this paper, we propose to describe subjects' drinking behavior using a Hidden Markov model (HMM) for categorical data, where the counts of drinks per day are summarized as a categorical variable with three levels, as is the convention in alcohol research. We compare the HMM's properties to those of other models, focusing on interpretability as well as out-of-sample prediction error; to do this, we analyze a real data set using each model. The HMM performs at a level comparable to the other models with respect to out-of-sample prediction error, and contains unique features that allow for useful clinical interpretations.

email: kshirley@wharton.upenn.edu

A BAYESIAN SEMIPARAMETRIC HIDDEN MARKOV MODEL FOR INCIDENCE ESTIMATION

Alejandro Jara*, Catholic University of Leuven
María José García-Zattera, Catholic University of Leuven
Emmanuel Lesaffre, Catholic University of Leuven

Motivated by a longitudinal oral health study, we present a Bayesian semiparametric hidden non-homogeneous Markov model for incidence estimation in which the underlying and within-subject multiple sequences of binary variables are subject to an unconstrained misclassification process. The elements of the transition matrices and of the initial distributions are described in terms of mixed-effects models in which the subject-specific effects and the association structure are captured by latent variables. The uncertainty in the mixing distribution is characterized by a mixture of Dirichlet process prior. We developed an appropriate simulation scheme for posterior computations using Markov chain Monte Carlo methods under the presence of absorbent states. The methodology is illustrated with an analysis of caries experience in deciduous molars using data from the Signal Tandmobiel$ ^ \circledR$ study.

email: alejandro.jaravallejos@med.kuleuven.be

11. GENOMICS: PATHWAYS AND NETWORKS

INCORPORATING GENE FUNCTIONAL ANNOTATIONS INTO REGRESSION ANALYSIS OF DNA-PROTEIN BINDING DATA AND GENE EXPRESSION DATA TO CONSTRUCT TRANSCRIPTIONAL NETWORKS

Peng Wei*, University of Minnesota
Wei Pan, University of Minnesota

Useful information on transcriptional networks has been extracted by regression analyses of gene expression data and DNA-protein binding data. However, a potential limitation of these approaches is their assumption on the common and constant activity level of a transcription factor(TF) on all the genes in any given experimental condition, which is not always true. Rather than assuming a common linear regression model for all the genes, we propose using separate regression models for various gene groups; the genes can be grouped based on their functions or some clustering results. Furthermore, to take advantage of the hierarchical structure of many existing gene function annotation systems, such as Gene Ontology(GO), we propose a shrinkage method that borrows information from relevant gene groups. In addition, we propose using cross-validation(CV) to do model selection for each level of the hierarchical structure. Applications to a yeast dataset and simulations lend support for our proposed method. In particular, we find that the shrinkage method consistently works well under various scenarios. We recommend the use of the shrinkage method as a useful alternative to the existing method.

email: weixx035@umn.edu

CLUSTER-NETWORK MODEL

Lurdes Y.T. Inoue*, University of Washington
Mauricio Neira, Vancouver General Hospital
Colleen Nelson, Vancouver General Hospital
Martin Gleave, Vancouver General Hospital
Ruth Etzioni, Fred Hutchinson Cancer Center

Network models are the focus of a growing number of researchers concerned with discovering novel gene interactions and regulatory relationships between genes from expression data. In this talk we will present a model-based approach that unifies the processes of inferring networks and clustering genes. Specifically, we provide a probabilistic framework for inferring clusters from gene expression profiles. Genes within the same cluster are expected to share a similar expression profile. We build a network over clusters using state-space models. We will illustrate the methods with simulation studies and a case study using time course microarray data arising from animal models on prostate cancer progression. Keywords: model-based clustering, Bayesian network, mixture model, dynamic linear model, time course microarray data, prostate cancer

email: linoue@u.washington.edu

BAYESIAN SEMIPARAMETRIC METHOD FOR PATHWAY ANALYSIS

Inyoung Kim*, Yale University
Herbert Pang, Yale University
Hongyu Zhao, Yale University

Pathways are sets of genes that serve a particular cellular or physiological function. The genes within the same pathway are expected to function together and hence may interact with each other. It is, therefore, of scientific interest to study their overall effect rather than each individual effect. Limited work has been done in the regression settings to study the effects of clinical covariates and large numbers of gene expression levels on a clinical outcome. In this paper we propose a Bayesian MCMC method for identifying pathways related to a clinical outcome based on the regression setting. A semiparametric mixed model (Liu, et al., 2006) is used to build dependence among genes using covariance structure with Gaussian, Polynomial, and Neural network kernels. The clinical covariate effects are modeled parametrically but gene expression effects are modeled nonparametrically. All variance components and nonparametric effect of genes are directly estimated using Bayesian MCMC approach. We compare our Bayesian MCMC approach with the method proposed by Liu et al. (2006) which was developed by connecting a least squares kernel machine with a linear mixed model.

email: inyoung.kim@yale.edu

A GENOME-BASED STATISTICAL CLUSTERING METHOD USING DYNAMIC PATTERNS: APPLICATION TO BUDDING YEAST DATA FOLLOWING GLUCOSE-GALACTOSE SHIFT

Wonsuk Yoo*, Wayne State School of Medicine
Sungchul Ji, Rutgers University

This study provides a new statistical clustering method to analyze cDNA microarray gene expression data. Even though cDNA microarray technology has rapidly emerged as a novel method to analyze thousand of expressed genes data coincidently, current microarray methods have focused on analysis through investigation of measuring mRNA levels. Recently a couple of papers have indicated that mRNA levels depend on two components, transcription rates (TR) and transcript degradation rates (TD), and that transcription rates and the rate of mRNA degradation play an important role in determining concentration of transcript (Garcia-Martinez et al. 2004, Fan et al. 2002). We show that an analysis based on data using both dynamic patterns from joint information of transcription rates (TR) and transcript levels (TL) can be more informative rather than that with only mRNA levels data since it describes dynamic process of gene expression on TL-TR joint information space. We provide a clustering statistical method which considers investigating trajectories of dynamic sequences of both TL and TR data, and apply to budding yeast data following Glucose-Galactose Shift.

email: wyoo@med.wayne.edu

JOINT MODELING OF GENE EXPRESSION AND MOTIF PROFILE DATA TO INFER TRANSCRIPTIONAL MODULES AND MODULE SPECIFIC ACTIVATION OF MOTIFS

Roxana A. Alexandridis*, University of Wisconsin-Madison
Sunduz Keles, University of Wisconsin-Madison
Rebecka Jornsten, Rutgers University

With the increase in the amount and quality of the databases containing putative and known transcription factor binding sites, as well as gene expression data, a task of immediate importance is to find transcriptional modules and the sets of motifs that control their regulatory mechanisms. A transcriptional module is defined as a group of co-regulated genes, whose similar mRNA expression profiles under various experimental conditions are governed by a common set of transcription factors. These transcription factors bind to motifs present in the promoter regions of the genes and, to a first approximation, activate or repress the transcription of the genes. We propose a probabilistic model that makes use of both gene expression and regulatory sequence data to assign genes to transcriptional modules. In particular, we consider a mixture model approach which factorizes into an expression mixture model and a regulatory sequence mixture model, conditional on the module membership. We make use of currently available mixture models for the expression part, and develop novel models and inference procedures based on Poisson mixtures for the regulatory sequence part. The developed methods are illustrated with S.cerevisiae case studies.

email: alexandridis@wisc.edu

VARIABLE SELECTION WITH STRONG HIERARCHY CONSTRAINTS AND ITS APPLICATION TO IDENTIFICATION OF GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS

Nam Hee Choi*, University of Michigan
Ji Zhu, University of Michigan

In this paper, we propose a new variable selection method for simultaneously fitting a regression model and identifying important interaction terms. Our method is in the framework of 'loss + penalty'. Unlike most of the existing penalization methods, the method we propose automatically enforces the hierarchy (or heredity) constraint, i.e., an interaction effect can be included in the model, only if the corresponding main effects are also included in the model. Numerical results on both simulation data and real data indicate that our method tends to remove non-relevant variables more effectively and provide better prediction performance than the classical LASSO method.

email: nami@umich.edu

INCORPORATING COVARIATES IN MAPPING HETEROGENEOUS TRAITS - A HIERARCHICAL MODEL USING EMPIRICAL BAYES ESTIMATION

Swati Biswas*, University of North Texas Health Science Center
Shili Lin, The Ohio State University

Complex genetic traits are inherently heterogeneous, that is, they may be caused by different genes, or non-genetic factors, in different individuals. So, for mapping genes responsible for these diseases using linkage analysis, heterogeneity must be accounted for in the model. Heterogeneity across different families can be modeled using a mixture distribution by letting each family have its own heterogeneity parameter. A substantial gain in power is expected if covariates that can discriminate between the families of linked and unlinked types are incorporated in this modeling framework. To this end, we propose a hierarchical Bayesian model, in which the families are grouped according to various (categorized) levels of covariate(s). The heterogeneity parameters of families within each group are assigned a common prior, whose parameters are further assigned hyper-priors. The hyper-parameters are obtained by utilizing the empirical Bayes estimates. We compare the proposed approach with one that does not take covariates into account and show that our approach leads to considerable gains in power to detect linkage and in precision of interval estimates through various simulation scenarios. An application to the asthma datasets of Genetic Analysis Workshop~12 also illustrates this gain in a real data analysis.

email: sbiswas@hsc.unt.edu

12. GENOME-WIDE ASSOCIATION STUDIES

GENETIC SIMILARITY MATCHING FOR GENOME-WIDE ASSOCIATION STUDIES

Weihua Guan*, University of Michigan
Liming Liang, University of Michigan
Michael Boehnke, University of Michigan
Gonçalo R. Abecasis, University of Michigan

Recently, genome-wide association studies have drawn great interest as a promising tool to dissect complex diseases such as hypertension, diabetes, and bipolar disorder. Although case-control association tests are generally more powerful than family-based association tests, population stratification that can lead to spurious disease-marker association or mask a true association remains a major concern. Several methods have been proposed to match cases and controls prior to genotyping or using genotype data for a modest number of genetic markers. Here, we describe a method for efficient matching of cases and controls after genotyping a large number of genetic markers in a genome-wide association study or large-scale candidate gene association study. Our method is comprised of three steps: 1) calculating similarity scores using the genotype data; 2) conducting optimal matching based on the scores so that matched cases and controls have similar genetic background; 3) using conditional logistic regression to perform association tests. Through simulations we show that out strategy can correctly control false positive rates and improve power to detect true disease predisposing variants. We also show that the optimal many-to-one matching has substantial advantages over one-to-one matching.

email: wguan@umich.edu

WINNER’S CURSE IN GENETIC ASSOCIATION STUDIES

Rui Xiao*, University of Michigan
Michael Boehnke, University of Michigan

Studies of gene-disease association are now commonly used to localize the genetic loci that propose to disease susceptibility. It is also of interest to estimate the genetic effect of each identified locus on the disease. It is known that the initial positive findings of the genetic effect estimate tend to be upwardly biased, given it was the first one to reach the statistical significance level. This phenomenon is called the ‘winner’s curse’, which is originated from auctions. However, the expected degree of the overestimation has not previously been investigated systematically. In our study, we model the winner’s curse in the context of case-control genetic association studies. We quantify its impact on the naïve estimators of the allele frequency difference between cases and controls as a function of several factors including sample size, minor allele frequency in controls, and the chosen statistical significance level. We also propose a maximum likelihood method to improve the estimate of the allele frequency difference corrected for the ascertainment. The simulation results indicate that the proposed method reduces the overestimation by different degrees, depending on the sample size and the chosen significance level.

email: xiaor@umich.edu

A GENERAL POPULATION GENETIC MODEL FOR HAPLOTYPING A COMPLEX TRAIT IN GENETIC ASSOCIATION STUDIES

Min Lin*, Duke University
Rongling Wu, Unversity of Florida

The past decade has witnessed the extensive development of statistical approaches to mapping complex traits with high-throughput single nucleotide polymorphisms (SNPs). The current approaches have now allowed for the identification of DNA sequence variants constructed by a series of linear SNPs (i.e., haplotype) that encode the expression of a complex trait. These approaches are based, however, on an important assumption that the population genotyped is at Hardy-Weinberg equilibrium (HWE). When this assumption is violated, a case that is likely to happen, especially for a small isolated population, the use of these approaches will be limited. In this talk, we present a statistical model for sequence mapping by incorporating a general population genetic model into analysis. The general model takes into account the deviation of haplotypes from HWE, allowing for the estimation and test of Hardy-Weinberg disequilibrium (HWD) for individual SNPs and linkage disequilibria of different orders. The results from simulation studies suggest that the new model covers current approaches and can be used in any population no matter whether it is at HWE or HWD. The new model has been validated by an example in which significant haplotype effects have been detected on a pharmacogenetic trait.

email: annie.lin@duke.edu

APPLICATION OF BAYESIAN VARIABLE SELECTION INCORPORATING LINKAGE DISEQUILIBRIUM FOR GENETIC ASSOCIATION STUDIES

Brooke L. Fridley*, Mayo Clinic
Mariza de Andrade, Mayo Clinic

Variable and model selection is becoming increasing important with the advent of high throughput genotyping methods resulting in many thousands of markers/SNPs. Currently, one is often either analyzing each SNP one at a time to determine genetic associations with the phenotype, either continuous or binary, or through haplotypes. To fit a model with multiple SNPs not in the same haplotype or region of chromosome, one is left with determining which SNPs to include in the model. A disadvantage of this single SNP approach is the multiple testing issues. Another disadvantage of the single SNP analysis approach for complex diseases and phenotypes is that, more than likely, multiple SNPs in various locations on the chromosome have individually small effects but collectively they may have a large effect on the phenotype of interest. Variable selection can be completed within the Bayesian framework in which variables and or models and be selected. We will illustrate the use Bayesian variable selection, incorporating linkage disequilibrium, for genetic association studies. In doing so, we will present results from a simulated dataset and a pharmacogenomic study.

email: fridley.brooke@mayo.edu

A WALD'S SPRT-BASED GROUP-SEQUENTIAL TESTING PROCEDURE FOR GENOME-WIDE ASSOCIATION SCANS

Andres Azuero*, University of Alabama at Birmingham
David T.Redden, University of Alabama at Birmingham

A common issue in genetic association studies and a major obstacle in genome-wide association scans is that a large number of markers are tested in the same group of subjects, increasing the probability of detecting false positives due to multiple comparisons. Wald's Sequential Probability Ratio Test (SPRT) provides an efficient way to sequentially test a single hypothesis as sample size accumulates subject by subject, but requires continuous monitoring which often makes it impractical. We propose an SPRT-based group-sequential procedure that combines the efficiency of the SPRT with a flexible correction for multiplicity, in the context of Genetic Association Studies and Genome-Wide Association Scans. We use simulated datasets to study the behavior of the procedure in regards to false positive detection rate, power, and cost-efficiency.

email: andreo@uab.edu

A BAYESIAN MODEL TO DESCRIBE THE ASSOCIATION BETWEEN GENOTYPE DATA AND A PHENOTYPE DEFINED IN A LIMITED RANGE

Ling Wang*, Boston University School of Public Health
Vikki Nolan, Boston University School of Public Health
Clinton T. Baldwin, Boston University School of Public Health
Martin H. Steinberg, Boston University School of Public Health
Paola Sebastiani, Boston University School of Public Health

In many genetic studies, the phenotype is a continuous variable such as a rate, a probability, or some percentage, and takes values between 0 and 1. The statistical analysis that uses linear regression with transformed phenotype as dependent variable may lead to false positive associations and, more importantly, may produce poor fit because of the data transformation. Here we propose a Bayesian approach that uses the Beta distribution to model the phenotype in the proper range of definition. We show how to estimate the phenotype-genotype association with or without adjustment for confounding in a Bayesian framework, using Markov Chain Monte Carlo methods. We evaluate this procedure by assessing the false positive rates and true positive rates in simulated data and we show that, compared to the model that assumes a lognormal distribution, our method has a comparable false positive rate but a much higher true positive rate. We apply this technique to a real genetic study targeting genes associated with fetal hemoglobin concentration in patients with sickle cell anemia and we discover a few novel genes that modulate fetal hemoglobin concentration.

email: wangling@bu.edu

SIMULTANEOUS CONFIDENCE INTERVALS FOR ODDS RATIOS IN CANDIDATE GENE STUDIES

Melinda H. McCann*, Oklahoma State University
Stephanie A. Monks, Oklahoma State University

Candidate gene studies often involve construction of multiple confidence intervals for odds ratios for the various haplotypes. We present methods to adjust these confidence intervals for multiplicity.

email: mccann@okstate.edu

13. NONPARAMETRIC SURVIVAL ANALYSIS

NONPARAMETRIC ESTIMATION OF MEAN RESIDUAL LIFE FUNCTION USING SCALE MIXTURES

Sujit K. Ghosh*, North Carolina State University
Shufang Liu, North Carolina State University

The mean residual life function (MRLF) of a subject is defined as the average residual life of the subject given that the subject has survived up to a given time point. It is well known that under mild regularity conditions an MRLF completely determines the probability distribution of the subjects' lifetime. In practice, the advantage of the MRLF over the more popularly used hazard function lies in its interpretation in many applications where the primary goal is often to characterize the remaining life expectancy of a subject instead of the instantaneous survival rate. For example, patients in a clinical trial might be more interested to know how many more years they are expected to survive given that they began a treatment at a certain time ago. A smooth nonparametric estimator of the MRLF is proposed using on a scale mixture of the empirical estimate of the MRLF based right-censored data. Asymptotic properties are established and compared with other nonparametric estimators currently available in the literature. Empirical performances of the proposed estimator are studied based on simulated data sets and finally, a real data set is used to illustrate the practical relevance of the proposed estimator.

email: ghosh@stat.ncsu.edu

SEMIPARAMETRIC REGRESSION WITH TIME-DEPENDENT COEFFICIENT FOR FAILURE TIME DATA ANALYSIS

Zhangsheng Yu*, The Ohio State University
Xihong Lin, Harvard School of Public Health

We propose a working independent profile likelihood method for the semiparametric time-dependent coefficient model with correlation. Kernel likelihood is used to estimate time-dependent coefficient. Profile likelihood for the parametric coefficient is formed by plugging in the nonparametric estimator. For independent data, the estimator is shown to be asymptotically normal and achieve the asymptotic semiparametric efficiency bound. We evaluate the performance of proposed nonparametric kernel estimator and the profile estimator, and apply the method to the western Kenya parasitemia data.

email: zyu@sph.osu.edu

EFFICIENT ESTIMATION IN ACCELERATED FAILURE TIME MODEL

Donglin Zeng*, University of North Carolina
Danyu Lin, University of North Carolina

The accelerated failure time model provides a natural formulation of the effects of covariates on potentially censored response variable. The existing semiparametric estimators are computationally intractable and statistically inefficient. In this article, we propose an approximate nonparametric maximum likelihood method for the accelerated failure time model with possibly time-dependent covariates. We estimate the regression parameters by maximizing a kernel-smoothed profile likelihood function. The maximization can be achieved through conventional gradient-based search algorithms. The resulting estimators are consistent and asymptotically normal. The limiting covariance matrix attains the semiparametric efficiency bound and can be consistently estimated. We also provide a consistent estimator for the error distribution. Extensive simulation studies demonstrate that the asymptotic approximations are accurate in practical situations and the new estimators are considerably more efficient than the existing ones. Illustrations with clinical and epidemiological studies are provided.

email: dzeng@email.unc.edu

EXPONENTIAL TILT MODELS IN THE PRESENCE OF CENSORING

Chi Wang*, Johns Hopkins University
Zhiqiang Tan, Johns Hopkins University

Various semi-parametric models can be used to describe survival distributions. Assuming a constant change in hazards leads to proportional hazard models. Alternatively, assuming a scale change in survival functions leads to accelerated failure time models. In this talk, we consider exponential tilt models that assume a parametric form for the density ratio of the two survival distributions, and develop a nonparametric likelihood method for estimation in the presence of censoring. Although exponential tilt models in the absence of censoring have been widely studied in connection with case-control studies and biased sampling, our work presents a first step in extending applications of such models for survival analysis.

email: chwang@jhsph.edu

MODELING REGRESSION MEAN RESIDUAL LIFE FUNCTION USING SCALE MIXTURES

Sujit K. Ghosh, North Carolina State University
Shufang Liu*, North Carolina State University

The mean residual life function (mrlf) of a subject is defined as the average remaining (residual) life of the subject given that the subject has survived up to a given time. It is well known that under mild regularity conditions, the mrlf determines the distribution uniquely. Therefore, the mrlf can be used to formulate a statistical model just as the survival and hazard functions. One of the main advantages of the mrlf is that it provides a direct measure of the average remaining as opposed to the hazard function. The commonly used proportional mean residual life model (PMRL) and linear mean residual life model (LMRL) have limited parameter scope. The regression model we propose using scale mixtures does not have any constraint. In the presence of censoring, we use full likelihood to develop the statistical inference for the regression coefficients. A simulation study is carried out to assess the properties of the estimators of the regression coefficients. We illustrate our regression model by applying it to the well-known Veterans Administration cancer survival data.

email: sliu@ncsu.edu

ESTIMATION OF HAZARD FUNCTION UNDER SHAPE RESTRICTIONS

Desale H. Habtzghi*, Georgia College and State University
Mary C. Meyer, University of Georgia
Somnath Datta, University of Louisville

The problem of estimation of hazard function has received considerable attention in the statistical literature. In particular, assumptions of increasing, decreasing, bathtub-shaped and convex hazard function are common in literature, but practical solutions are not well developed. In this talk we introduce a new nonparametric method for estimation of hazard functions under shape restrictions to handle the above problem. We adopt a nonparametric approach in assuming that the density and hazard rate have no specific parametric form with the assumption that the shape of the underlying hazard rate is known (either decreasing,increasing, concave, convex or bathtub-shaped). We also show how the estimation procedures can be used when dealing with right censored data. We evaluate the performance of the estimator via simulation studies and illustrate it on some real data set.

email: desale.habtzghi@gcsu.edu

NONPARAMETRIC INFERENCE ON MEDIAN RESIDUAL LIFE FUNCTION

Jong-Hyeon Jeong*, University of Pittsburgh
Sin-Ho Jung, Duke University
Joseph P. Costantino, University of Pittsburgh

A simple approach to estimation of the median residual lifetime is proposed for a single group by inverting a function of the Kaplan-Meier estimators. A test statistic is proposed to compare two median residual lifetimes at any fixed time point. The test statistic does not involve estimation of the underlying probability density function of failure times under censoring. Extensive simulation studies are performed to validate the proposed test statistic in terms of type I error probabilities and powers at various time points. One of the oldest datasets from the National Surgical Adjuvant Breast and Bowel Project (NSABP), which has more than a quarter century of follow-up, is used to illustrate the method. The analysis results indicate that, without systematic post-operative therapy, a significant difference in median residual lifetimes between node-negative and node-positive breast cancer patients persists for about 10 years after surgery. The new estimates of the median residual lifetime could serve as a baseline for physicians to explain any incremental effects of post-operative treatments in terms of delaying breast cancer recurrence or prolonging remaining lifetimes of breast cancer patients.

email: jeong@nsabp.pitt.edu

14. INNOVATIONS IN CLINICAL TRIALS DESIGN

ADVANCES IN CLINICAL TRIAL DESIGN AND EMERGING PROBLEMS

H.M. James Hung*, U.S. Food and Drug Administration

Statistical methodology for clinical trial designs is rapidly advanced in the recent literature. The advances are primarily to provide a number of frameworks under which some elements of a clinical trial design can be properly modified during the course of the trial. In addition, a sequence of classical clinical trials can be combined in a flexible way, under the so-called flexible design framework. Under the new frameworks, statistical testing can be validly performed in the sense of a proper control of overall false positive rate. However, the valid statistical testing strategies come with a lot more emerging problems when they are used in practice. In this presentation, the key methods will be surveyed. The utility and pitfalls of these methods will be discussed in the context of the clinical program development for pharmaceutical products.

email: hsienming.hung@fda.hhs.gov

DATA-DRIVEN INTERIM ANALYSES WITH INTENT TO CHEAT

KuangKuo Gg Lan*, Johnson & Johnson
Peter Hu, Johnson & Johnson

When an alpha spending function is employed to design a sequential study, the frequencies of interim looks are not required to be pre-specified. However, data-driven interim looks may inflate the alpha level. The extent of inflation has been investigated by Proschan-Follmann-Waclawiw (1992), where they concluded for the Pocock and the O'Brien-Fleming spending functions, the alpha inflation caused by data-driven looks is minimal. However, the alpha inflation will be a serious problem for data-driven interim looks for spending functions with big jumps in small time intervals. A similar problem occurs for the maximum information designed study. We will look into this problem and suggest some reasonable solutions in practice.

email: glan@prdus.jnj.com

ADAPTIVE SAMPLE SIZE BASED ON A NUISANCE PARAMETER: SHOULD WE CONDITION ON THE SAMPLE SIZE ACTUALLY USED?

Michael A. Proschan*, National Institute of Allergy and Infectious Diseases
Martha C. Nason, National Institute of Allergy and Infectious Diseases

Sample size calculations in clinical trials typically require specification of values of one or more nuisance parameters. For a continuous outcome, the nuisance parameter is the variance or standard deviation (assumed equal across arms). For a dichotomous outcome, the nuisance parameter is either the control probability or the overall probability among all patients combined across arms. Internal pilot studies use within-trial data to estimate the nuisance parameter and modify the sample size, if needed. If an internal pilot study is used, should inference be based on the sample size actually used, or averaged over all possible sample sizes? These two approaches are shown to yield very different answers in some cases, such as with a dichotomous outcome and sample size recalculation based on the control event proportion.

email: ProschaM@mail.nih.gov

15. METABOLOMICS

LINKING GENETIC PROFILES TO BIOLOGICAL OUTCOME

S. Stanley Young*, Metabolon, Inc., NISS
Paul Fogel, Consultant, France
Doug Hawkins, University of Minnesota

Micro array, proteomics, metabolomics, etc., all produce data sets where there are many more predictor variables than observations. There are correlations among these variables; indeed, as n<<p and biological systems have to work together, the many variables/predictors can not all be independent of one another. The correlations can be utilized to improve the linking predictors to outcomes (disease, drug effects, etc.). New inference methods will be presented which combine statistical testing with matrix factorization. The methods will be demonstrated using a real data set.

email: young@niss.org

EXPLORING A COMPLEX METABOLOMICS DATA SET

Susan J. Simmons*, University of North Carolina-Wilmington
Emilea Norris, AAI Pharma
Matthew Mitchell, Metabolon

Data from metabolomics experiments pose a number of interesting statistical challenges. Some issues that arise from this type of data are correlated metabolites, non-normal distributions, number of samples less than the number of metabolites ($n < p$), and non-random missing values. In this presentation, we focus on the missing value problem and how missing values affect feature selection. Based on a number of simulations, we recommend appropriate methods to use when dealing with this type of data.

email: simmonssj@uncw.edu

PATHWAY-BASED ANALYSIS OF METABOLIC PROFILES

Jacqueline M. Hughes-Oliver*, North Carolina State University

Metabolomics is emerging as an attractive component of the extensive body of platforms for systems biology. The relatively small number of metabolites and their mostly known network of interconnectivity and relation to disease make metabolomics a prime candidate for improving both disease diagnosis and treatment. Unfortunately, analysis of metabolomic data typically ignores knowledge of disease pathways, and hence the opportunity for taking full advantage of domain knowledge is lost. This work develops a simple systemized view of metabolites and their existence in pathways, and then incorporates this knowledge in analyzing the impact of pathway measures on disease status. Our focus is more interpretable diagnosis of disease occurrence.

email: hughesol@stat.ncsu.edu

16. STATISTICAL METHODS IN HIV GENOMICS

GRAPHICAL MODELS FOR THE ACCUMULATION OF HIV DRUG RESISTANCE MUTATIONS

Niko Beerenwinkel*, Harvard University

Drug resistance is a major limitation of antiretroviral therapy. I present graphical models for the evolution of drug resistance, which is characterized by the accumulation of resistance-associated mutations in the viral genome. The statistical models impose certain contraints on the order in which mutations can occur. I will discuss methods for inferring these order constraints and the model parameters from cross-sectional as well as from longitudinal HIV sequence data. The usefulness of this approach is demonstrated in predicting phenotypic drug resistance and treatment outcome from viral genotypes.

email: beerenw@fas.harvard.edu

A RESAMPLING-BASED APPROACH TO MULTIPLE TESTING WITH UNCERTAINTY IN PHASE

Andrea S. Foulkes*, University of Massachusetts
Victor G. DeGruttola, Harvard School of Public Health

Characterizing the genetic correlates to complex diseases requires consideration of a large number of potentially informative biological markers. Attention to alignment of alleles within or across chromosomal pairs, commonly referred to as phase, may be essential for uncovering true biological associations. In the context of population based association studies, phase is generally unobservable. Preservation of type-1 error in a setting with multiple testing presents a further analytical challenge. Our research combines a likelihood-based approach to handling missingness in phase with a resampling method to adjust for multiple testing. This method can be extended to allow for parametric models that adjust for covariates. For settings in which no parametric methods are available, semiparametric methods can also be employed with a small cost in power and large gain in robustness. Through simulations we demonstrate preservation of the family-wise error rate and reasonable power for detecting associations. The method is applied to a cohort of 626 HIV-1 infected individuals receiving highly active anti-retroviral therapies, to ascertain potential genetic contributions to abnormalities in lipid profiles. The haplotypic effects of 2 genes, hepatic lipase and endothelial lipase, on high-density lipoprotein cholesterol are tested.

email: foulkes@schoolph.umass.edu

NON- AND SEMI-PARAMETRIC ANALYSIS OF MULTIPLE CATEGORICAL PREDICTORS AND SEVERAL OUTCOMES

A. Gregory DiRienzo*, Harvard University

Investigating the relationship between an outcome variable and the joint combinations of multiple categorical predictors is challenging in several ways. First, an apparent association between a joint predictor combination and outcome may actually be redundant in the sense that the association is completely driven by one or more lower-order combinations of predictors. Second, the total number of joint predictor combinations to consider is often unrealistically large. This paper proposes a nonparametric method for approximating a full multivariate comparison between the predictors and outcome, along with an efficient computational algorithm. The method estimates the set of predictor combinations that have a non-redundant association with outcome. The probability that the estimate contains one or more predictor combinations that are either not associated with the outcome, or possess a redundant association, is asymptotically bounded above by any pre-specified level for arbitrary data-generating distributions. A statistically formal procedure for selecting a final model is proposed. Simultaneous analysis of several outcome variables is easily accomplished, as is adjustment for low- dimensional factors. The method is shown to perform well for finite samples in simulation study, and is applied to model the relationship between week 24 HIV-1 RNA and genotype in ACTG398.

email: dirienzo@hsph.harvard.edu

FINDING LOW DIMENSIONAL STRUCTURE IN HIGH DIMENSIONAL DATA

Hugh Chipman, Acadia University
Andrea Foulkes, University of Massachusetts
Edward George*, University of Pennsylvania
Robert McCulloch, University of Chicago

Consider the canonical regression setup where one wants to model the relationship between y, a variable of interest, and x, a high dimensional vector of potential predictors. For this general problem we propose BART (Bayesian Additive Regression Trees), a new approach to discover the form of f(x)= E(Y | x) and draw inference about it. BART approximates f by a Bayesian 'sum-of-trees' model where each tree is constrained by a prior to be a weak learner as in boosting. Fitting and inference are accomplished via an iterative backfitting MCMC algorithm. By using a large number of trees, essentially an over-complete basis for f, we have found BART to be remarkably effective at finding highly nonlinear relationships hidden within a large number of irrelevant potential predictors. Applications to problems in HIV genomics are illustrated.

email: edgeorge@wharton.upenn.edu

17. NEW APPROACHES FOR ANALYZING FUNCTIONAL DATA

BAYESIAN METHOD FOR CURVE CLASSIFICATION USING WAVELETS

Xiaohui S. Wang*, University of Texas-Pan American
Shubhankar Ray, Merck Research Laboratories
Bani K. Mallick, Texas A&M University

We propose classification models for binary and multicategory data where the predictor is a random function. We use Bayesian modeling with wavelet basis functions which have nice approximation properties over a large class of functional spaces and can accommodate a variety of functional forms observed in real life applications. We develop an unified hierarchical model to encompass both the adaptive wavelet based function estimation model as well as the logistic classification model. These two models are coupled together to borrow strengths from each other in this unified hierarchical framework. The use of Gibbs sampling with conjugate priors for posterior inference makes the method computationally feasible. We compare the performance of the proposed model with other classification methods such as the existing naive plug-in methods by analyzing simulated and real datasets.

email: xhwang@utpa.edu

FUNCTIONAL DATA ANALYSIS FOR GENE EXPRESSION TIME COURSES

Hans-Georg Müller*, University of California-Davis

Temporal gene expression profiles characterize the time-dynamics of the expression of specific genes. Functional warping methods can be useful for the analysis of such profiles as we illustrate with a simple time-shift warping approach applied to the yeast cell cycle data. To analyze the relation of dynamic changes in various sets of temporal gene expression profiles, functional linear regression models are useful. The interpretation of such models may be simplified by a decomposition into a series of simple linear regressions of functional principal component scores of response on those of predictor trajectories. We discuss issues such as functional coefficient of determination, functional regression diagnostics and inference via bootstrap, and demonstrate the application of these methods for gene expression profiles belonging to specific gene groups, relating expressions of late and early developmental phases in Drosophila. This presentation is based on joint work with Jeng-Min Chiou and Xiaoyan Leng.

email: mueller@wald.ucdavis.edu

A BAYESIAN MODEL FOR SPARSE FUNCTIONAL DATA

Wesley K. Thompson*, University of Pittsburgh
Ori Rosen, University of Texas-El Paso

We propose a method for analyzing data which consist of curves on multiple individuals, i.e., longitudinal or functional data. We use a Bayesian model where curves are expressed as linear combinations of basis functions with random coefficients. The curves are estimated as posterior means obtained via Markov Chain Monte Carlo (MCMC) methods, which automatically select the local-level of smoothing. The method is applicable to situations where curves are sampled sparsely and/or at irregular timepoints. We construct posterior credible intervals for the mean curve and for the individual curves. We also incorporate covariates into the model. This methodology is demonstrated on two applications and monte carlo simulation.

email: wesleyt@pitt.edu

18. NEW METHODS FOR GENETIC ASSOCIATION STUDIES

STATISTICAL APPROACHES TO WHOLE GENOME ASSOCIATION TESTING

Andrew Clark*, Cornell University

Whole-genome association testing is widely cited as having promise for identification of genetic variants that are causal to elevated risk of complex disorders like cardiovascular disease, diabetes, and cancers. The magnitude of the multiple testing problem in whole-genome association testing is, however, daunting. One approach that we are developing is Bayesian classification, a promising approach for inference when the number of predictors (SNPs) is large, but where the prior expectation is that most SNPs will have zero effect. The model has a three-component mixture prior with a high point mass at zero (no effect) as well as positive and negative effects on risk. Fitting is done by Monte Carlo Markov chain and by stochastic variable selection. We apply the model to gene expression data obtained from cell lines from the 270 subjects of the HapMap project (each having more than 4 M SNP genotypes). The Bayesian classification approach will be contrasted with linear model based approaches. Both case-control and random cohort data will be addressed. Performance of the methods in the face of missing and erroneous data will be quantified.

email: ac347@cornell.edu

DESIGN AND ANALYSIS OF GENOME WIDE ASSOCIATION STUDIES: APPLICATION TO TYPE 2 DIABETES

Michael Boehnke*, University of Michigan
Andrew Skol, University of Chicago School of Medicine
Goncalo Abecasis, University of Michigan
Laura Scott, University of Michigan

The catalog of human genetic variants and the drop in genotyping costs have made genome wide association studies (GWAs) a practical approach to study the genetic basis of human diseases. GWAs often require genotyping 100Ks of genetic markers on 100s-1000s of subjects. In this talk, we discuss optimal experimental design and analysis for two-stage GWAs in which a subset of a sample is genotyped on all markers in stage 1, and the remaining samples are genotyped on the most interesting markers in stage 2. Consistent with prior work of Satagopan, Elston, Thomas, and colleagues, we find that two-stage designs can maintain nearly the same power to detect association as the one-stage design in which all samples are genotyped for all markers. We argue that joint analysis of stage 1 and 2 samples is more powerful than replication-based analysis, despite the much larger number of tests implied. We address the impact of design parameters (proportion of sample in stage 1 and proportion of markers followed up in stage 2) and study setting (per genotype cost differences and etiologic heterogeneity between in stages 1 and 2) on optimal design and analysis approach. We illustrate these ideas with results from stage 1 of a two-stage GWA of type 2 diabetes carried out as part of the Finland-United States Investigation of NIDDM Genetics (FUSION) study.

email: boehnke@umich.edu

LIKELIHOOD-BASED INFERENCE ON HAPLOTYPE-DISEASE ASSOCIATIONS

Danyu Lin*, University of North Carolina
Donglin Zeng, University of North Carolina

A haplotype is a specific sequence of nucleotides on a single chromosome. The population associations between haplotypes and disease phenotypes provide critical information about the genetic basis of complex human diseases. It is highly challenging to make statistical inference about these associations because of unknown gametic phase in genotype data. The common practice of probabilistically inferring individual haplotypes leads to biased and inefficient analysis of association. We present proper statistical methods for inferring haplotype-phenotype associations. We consider all commonly used study designs, including cross-sectional, case-control, cohort, case-cohort and nested case-control studies, as well as family-based studies. The phenotype can be a disease indicator, a quantitative trait or a potentially censored time to disease variable. The effects of haplotypes on the phenotype are formulated through flexible regression models, which can accommodate a variety of genetic mechanisms and gene-environment interactions. We construct appropriate likelihood functions for various study designs and disease phenotypes, and establish the consistency, asymptotic normality and efficiency of the maximum likelihood estimators. We develop simple and stable numerical algorithms to implement the corresponding likelihood-based inference procedures. Simulation studies demonstrate that the proposed methods perform well in practical settings. We provide applications to several real studies.

email: lin@bios.unc.edu

19. MISCLASSIFIED OR MISMEASURED DATA IN EPIDEMIOLOGY

INFERRING EXPOSURE-DISEASE RELATIONSHIPS WHEN EXPOSURE IS MISMEASURED: NO ADJUSTMENT VERSUS SENSITIVITY ANALYSIS VERSUS BAYESIAN ADJUSTMENT

Paul Gustafson*, University of British Columbia

There is wide acknowledgement that statistical inferences concerning exposure-disease associations should adjust for exposure mismeasurement (of either a categorical or continuous exposure). However, such adjustment is not common in practice. One problem is that full identification of exposure-disease parameters requires considerable knowledge of the mismeasurement process, through some combination of a priori assumptions and richness of the data (the availability of gold-standard exposure measurements for a sub-sample, for instance). This gap between the ideal situation and the practical situation seems to encourage either not attempting adjustment, or carrying out sensitivity analysis, whereby adjusted inferences arising from a variety of assumptions about the mismeasurement process are reported. This talk examines whether meaningful adjustment can be carried out without sufficient assurances of identifiability, and whether sometimes the data contain useable information about the parameters being varied `by hand’ in a sensitivity analysis. These issues are addressed in a Bayesian framework, which we argue is well-suited for this class of problems.

email: gustaf@stat.ubc.ca

MULTIVARIATE META-ANALYSIS OF DIAGNOSTIC TESTS WITHOUT A GOLD STANDARD

Haitao Chu*, The Johns Hopkins Bloomberg School of Public Health
Sining Chen, The Johns Hopkins Bloomberg School of Public Health

In diagnosis practice, it is common to apply two imperfect tests jointly or sequentially to a study population. Quite often, neither test can be regarded as a gold standard. To estimate the accuracy of microsatellite instability testing (MSI) and traditional mutation analysis in predicting germline mutations of mismatch repair (MMR) genes, Bayesian methods (Chen, Watson, and Parmigiani 2005) have been proposed to handle missing data resulting from partial testing and the lack of a gold standard for both tests in a recent meta-analysis. In this paper, we demonstrate improved estimation of the sensitivity and specificity of MSI and traditional mutation analysis by using a nonlinear mixed model and a Bayesian hierarchical model, which account for the heterogeneity across studies through study-specific random effects. The methods can be used to estimate the accuracy of two imperfect diagnostic tests in other meta-analyses when the prevalence of disease, the sensitivity and/or the specificity of diagnostic tests are heterogeneous among studies.

email: hchu@jhsph.edu

IDENTIFIABILITY IN THE PRESENCE OF MISCLASSIFICATION ERROR

Daniel O. Scharfstein*, The Johns Hopkins Bloomberg School of Public Health
Yong Chen, The Johns Hopkins Bloomberg School of Public Health
Shuai Wang, The Johns Hopkins Bloomberg School of Public Health

In this talk, we consider issues of identifiability in a setting where one is interested in understanding the relationship between a binary response Y and a binary covariate X that cannot be directly observed due to misclassification error. We assume that we observe two binary surrogates for X, namely X1 and X2. We discuss identification of the distribution of Y given X, under (1) non-differential misclassification error: X1 and X2 are jointly independent of Y given X and (2) conditional independence: X1 and X2 are independent given X. While local identifiability is relatively easy to establish, global identifiability is much more difficult. We also discuss how to relax assumption (2), through a set of sensitivity analysis assumptions that are indexed by a parameter that expresses varying levels of dependence between X1 and X2 after conditioning on X.

email: dscharf@jhsph.edu

CLASSIFYING HELTHCARE ASSOCIATED INFECTIONS USING DATE OF ONSET

Justin Lessler*, The Johns Hopkins Bloomberg School of Public Health
Ronald Brookmeyer, The Johns Hopkins Bloomberg School of Public Health
Trish M. Perl, Johns Hopkins Hospital

Date of symptom onset is often used to distinguish healthcare associated infections from community acquired infections. Those patients developing symptoms early in an inpatient stay are considered to have community acquired infection, while those developing symptoms later are considered nosocomially infected. We have preformed a probabilistic analysis of this technique, showing how misclassification rates depend on disease incubation period and the incidence rate ratio of infections among inpatients versus the community. We provide quantitative results that aid in selecting the decision rule for classifying patients that will best meet desired performance criteria. This analysis shows that identification of nosocomial infection using only date of symptom onset can perform well for important illnesses. For example, using date of onset: influenza can be classified with NPV >= PPV = 87%, strep throat can be classified with NPV >= PPV = 84%, measles can be classified with NPV >= PPV = 90%, and legionnaires disease can be classified with NPV >= PPV = 77%. These results increase the utility of classifying infections using date of onset by providing theoretically sound measures of performance; and are applicable beyond the hospital setting.

email: jlessler@jhsph.edu

20. CAUSAL INFERENCE

RECURSIVE PATH ANALYSIS FOR CATEGORICAL VARIABLES

Haihong Li*, University of Florida

Recursive path analysis is a useful tool for causal inference about a chain of three or more response variables in which the causal effects are in one direction. The primary objective in such analysis is to decompose the total effect of each variable into its direct and indirect components. Methods for recursive analysis of a chain of continuous variables are well developed but there is a lack of uniform methodology when the variables are categorical. We propose an approach for categorical response variables that is based on generalized linear models. An illustration using real data is provided.

email: hli@biostat.ufl.edu

MULTIVARIATE PATH MODELS AND THE CALCULUS OF COEFFICIENTS

Youngju Pak*, The SUNY at Buffalo
Randy L. Carter The SUNY at Buffalo

Path analysis is useful to explain the interrelationships among sets of observed variables in a causal chain. However, path analysis has been underutilized in health science and epidemiology research, because of its restrictive requirement of a complete causal ordering of variables. In order to solve this problem, we suggest the use of multivariate path models and derive the ¡°Calculus of Coefficients (COC)¡±, which results in a partitioning of the matrix of total effects into a sum of the matrix of direct effects and all matrices of indirect effects through intermediate outcome vectors. The multivariate COC derived in this study extends the classical univariate path model and associated results to the multivariate case, where vectors of outcome variables replace single variables in the causal chain. Estimated indirect effects are tested by intersection-union tests. Both theoretical considerations and an application are presented. A data set from the Western New York Health Study is used to illustrate the methods. We partition the total effects of health behavior variables into direct and indirect effects on cardiometabolic disease risk factors through anthropometric variables and through composite blood measures that reflect chronic inflammation, endogenous steroid levels, anemia, and blood viscosity.

email: ypak@buffalo.edu

SENSITIVITY ANALYSIS TO ACCOUNT FOR A NON-IGNORABLE MISSING COVARIATE IN THE ESTIMATION OF SACE

Brian L. Egleston*, Fox Chase Cancer Center
Daniel O. Scharfstein, Johns Hopkins University

A common study design for investigating outcomes among those who have had a catastrophic health event is to enroll a sample of all individuals with the event and then interview the survivors. Although information from patients who die might be abstracted from medical records, pre-event information such as usual functional status is only available among those who are interviewed. This can result in covariate data that is missing almost exclusively among those who die. Because the reason for missingness is so linked to death, researchers might believe that data are not missing at random. In order to make meaningful inferences concerning non-mortality outcomes in the presence of death, many have suggested examining the effect of a treatment only on the subset of individuals who would survive either with or without an exposure, sometimes referred to as the Survivors Average Causal Effect (SACE). We demonstrate how the method of sensitivity analysis can be used to identify this effect when covariates are missing exclusively on those who die.

email: brian.egleston@fccc.edu

ESTIMATING A CAUSAL TREATMENT EFFECT ON A MARK VARIABLE WITH COMPLICATIONS OF FAILURE TIME CENSORING

Jing Ning*, Johns Hopkins University
Mei-Cheng Wang, Johns Hopkins University

In a randomization study, treatment effects on a mark variable measured at the time of failure events are important index for evaluating treatment efficacy. To analyze the data, a difficulty is that the values of the mark variable are not observable when failure times are censored and the mark variable is typically observed subject to selection bias. Thus, comparisons based on mark variables measured at post-treatment events do not have a causal interpretation. Further more, when a failure time censoring is present, the marginal distribution of the mark variable may not be identifiable. In this talk we consider models and required assumptions to address the identifiability and estimation of causal treatment effects on mark variables. Formulating the problem by the principal stratification framework of causal inference, we propose a class of causal treatment effects under a randomized study setting. We develop analytical procedures by borrowing information from failure time data to correct selection bias from observed mark variable data. Asymptotic properties of the proposed estimators are established. Numerical studies demonstrate that the estimators perform well with practical sample sizes.

email: jning@jhsph.edu

THE SIGN OF THE BIAS OF UNMEASURED CONFOUNDING

Tyler J. VanderWeele*, University of Chicago
Miguel A. Hernan, Harvard School of Public Health
James M. Robins, Harvard School of Public Health

A result is given which allows the researcher in certain cases to determine the sign of the bias which arises when control for confounding is inadequate. The result is given within the context of the directed acyclic graph causal framework and is rigorously stated in terms of signed edges. Rules governing these signed edges are provided. Cases in which intuition concerning signed edges fails are clearly articulated. The central result provides an alternative to sensitivity analysis and can be used to provide concrete bounds on the effect of a particular exposure on an outcome in the presence of unmeasured confounding variables.

email: vanderweele@uchicago.edu

COMBINING INFORMATION FROM RANDOMIZED AND OBSERVATIONAL DATA: A SIMULATION STUDY

Eloise E. Kaizar*, The Ohio State University
Howard Seltman, Carnegie Mellon University
Joel Greenhouse, Carnegie Mellon University

Randomized controlled trials have become the gold standard of evidence in medicine. They earned this status because they offer strong internal validity. However, subject recruitment may introduce selection bias that limits trials' external validity. To mitigate the selection bias some turn to meta-analysis to widen the recruitment pool; in practice, this method is not likely to eliminate all selection bias. Observational studies are also commonly used in medical and epidemiological research. Complementary to randomized trials, these studies tend to have strong external validity or broad generalizability, but because of treatment self-selection often have severely limited internal validity. We propose a response surface framework for combining both randomized and observational data in a single overarching probability model that models the selection bias of the randomized studies and the self-selection bias of the observational studies. Simulations show that our framework may produce a single estimate with less bias than estimates derived using current methods.

email: ekaizar@stat.ohio-state.edu

21. LONGITUDINAL DATA APPLICATIONS

CLASSIFICATION RULES FOR TRIPLY MULTIVARIATE DATA WITH AN AR(1) CORRELATION STRUCTURE ON THE REPEATED MEASURES OVER TIME

Anuradha Roy*, The University of Texas at San Antonio
Ricardo Leiva, F.C.E., Universidad Nacional de Cuyo-Argentina

Under the assumption of multivariate normality we study the classification rules for triply multivariate data (multivariate in three directions), where more than one response variable is measured on each experimental unit on more than one site at several time points. It is very common in clinical trial study to collect measurements on more than one variable at different body positions (sites) repeatedly over time. The new classification rules, with and without time effect, and with certain structured and unstructured mean vectors and covariance structures, are very efficient in small sample scenario, when the number of observations is not adequate to estimate the unknown variance-covariance matrix. We introduce a parametrically parsimonious model for the classification rules by introducing an "equicorrelated (partitioned) matrix" on the measurement vector over sites in addition to AR(1) correlation structure on repeated observations over time. Computation algorithms for maximum likelihood estimates of the unknown population parameters are presented. Simulation results show that the introduction of the sites in the classification rules improves its performance.

email: aroy@utsa.edu

FLEXIBLE ESTIMATION OF SERIAL CORRELATION IN LINEAR MIXED MODELS

Jan Serroyen*, Hasselt University, Belgium
Geert Molenberghs, Hasselt University, Belgium
Marc Aerts, Hasselt University, Belgium
Geert Verbeke, Katholieke Universiteit Leuven, Belgium

The linear mixed effects model has, arguably, become the most commonly used tool for analyzing continuous, normally distributed longitudinal data. In its general model formulation four structures can be distinguished: fixed effects, random effects, measurement error and serial correlation. Broadly speaking, serial correlation captures the phenomenon that the correlation structure within a subject depends on the time lag between two measurements. While the general linear mixed model is rather flexible, the need has arisen to further increase flexibility. In response, quite some work has been done to relax the model assumptions and/or to extend the model. For example, the normality assumption for the random effects has been generalized in several ways. Comparatively less work has been devoted to more flexible serial correlation structures. Therefore, we propose the use of spline-based modeling of the serial correlation function. The approach is applied to data from a pre-clinical experiment on the eating and drinking behavior in rats.

email: jan.serroyen@uhasselt.be

IMPLEMENTATION OF A NEW CORRELATION STRUCTURE IN FRAMEWORK OF GEE WITH R SOFTWARE

Jichun Xie*, University of Pennsylvania
Justine Shults, University of Pennsylvania

The generalized estimating equation (GEE) approach is extremely popular because it extends generalized linear models for correlated data. However, one limitation is that GEE has only been implemented for a few correlation structures that are used to describe the pattern of association. For a study in Ophthalmology we use quasi-least squares to implement a correlation structure that has not previously been implemented in the framework of GEE. We discuss the structure and the analysis results that we obtained using the QLSinR procedure that is under development.

email: jichun@mail.med.upenn.edu

IMPLEMENTING SEMIPARAMETRIC VARYING-COEFFICIENT PARTIALLY LINEAR MODELS FOR LONGITUDINAL DATA

Jialiang Li*, National University of Singapore
Yingcun Xia, National University of Singapore
Mari Palta, University of Wisconsin,-Madison

A flexible semiparametric model is studied for its application for longitudinal data. We estimate the model via the profile least square method and employ the cross-validation method to select the bandwidth. The model is implemented for a population based analysis to study the temporal relationship between systolic blood pressure and urinary albumin excretion. The regression coefficients of some important risk factors for systolic blood pressure are shown to display a nonparametric shape. Our computation strategy makes a calss of complicated yet useful models more accessible to applied researchers.

email: stalj@nus.edu.sg

MODELING MULTIVARIATE LATENT TRAJECTORIES AS PREDICTORS OF A UNIVARIATE OUTCOME

Sujata M. Patil*, Memorial Sloan-Kettering Cancer Center
Trivellore E. Raghunathan, University of Michigan
Jean T. Shope, University of Michigan

We describe a model where two or more longitudinal markers are predictors of a univariate outcome. Each longitudinal marker is summarized by a few latent trajectory variables in stage one of the model. In stage two of the model, these latent trajectory variables are used as predictors of the outcome. A fully Bayesian approach is used to obtain estimates for model parameters. We apply the proposed methods to study the effect of several adolescent substance use trajectories on motor vehicle offenses incurred during young adulthood. Results from a corresponding simulation study and methodological challenges in fitting these models will be discussed.

email: patils@mskcc.org

JOINT LONGITUDINAL ANALYSIS OF ALCOHOL AND DRUG USES FOR YOUNG ADULTS

Liang Zhu*, University of Missouri-Columbia
Jianguo Sun, University of Missouri-Columbia
Phillip Wood, University of Missouri-Columbia

Alcohol and drug uses are common in today's society and especially among college students and it is well-known that they can lead to serious consequences. Corresponding to this, many studies have been conducted in order, for example, to learn or understand short- or long-term temporal processes of alcohol and drug uses, to identify fixed or time-dependent risk factors that affect the alcohol and drug uses, and/or to assess their relationships with some serious outcomes such as dropping out of school and causing traffic accidents. This paper considers a prospective study of alcohol and drug uses on college freshmen from a large, midwestern university. %A number of authors have analyzed the study from many different points of views, but %most of them focused on the univariate analysis of a particular response variable. We focus on joint analysis of both alcohol and drug uses and for the purpose, several statistical models and approaches are presented and applied to the resulting data set. In particular, a marginal mean model is proposed that leaves the correlation between response outcomes arbitrary. The analysis focuses on the outcomes that represent negative consequences of alcohol and drug uses and the results indicate that the consequences of alcohol and drug uses decrease along with ages.

email: lzkn7@mizzou.edu

FLEXIBLE MODELING OF EXPOSURE DATA WITH INFORMATIVE NUMBER OF REPEATED MEASUREMENTS

Huichao Chen*, Emory University
Amita K. Manatunga, Emory University
Robert H. Lyles, Emory University
Limin Peng, Emory University
Michele Marcus, Emory University

The distribution of exposure measurements in epidemiologic and environmental studies often tends to be highly skewed, containing heaps and coarse values. To model the decay over time based on repeated measurements of such an exposure, we propose a generallatent model suitable for highly skewed and grouped data, assuming that observed exposures are determined by an unobservable Weibull-distributed variable. To accommodate correlations among repeated responses over time, we introduce a general random effect from the power variance function (PVF) family of distributions (Hougaard, 2000). Since the number of measurements per subject is correlated with his/her initial exposure, we further extend this model by incorportaing a geometric distribution for the number of observations. With or without modeling the number of measurements per subject, the resulting marginal likelihoods have closed forms. We study the inference under these models, including estimation and hypothesis testing, with different choices of random effect distributions. Simulation studies are conducted to evaluate their performance. Finally, we apply the proposed method to the polybrominated biphenyl (PBB) exposure data collected from the Michigan Female Health Study (MFHS).

email: hchen4@sph.emory.edu

22. MISSING DATA

DETECTING MULTIPLE SOURCES OF INFORMATIVE DROPOUT IN CLUSTERED LONGITUDINAL DATA

Sara B. Crawford*, Emory University
John J. Hanfelt, Emory University

Longitudinal studies tracking the rate of change are subject to patient dropout. Not only might this dropout be informative, but also heterogeneous, in the sense that different causes might contribute to multiple patterns of informative dropout. We propose a random effects approach to testing for multiple sources of informative dropout when reasons for dropout are known or unknown. The proposed score test is robust in that it does not depend on the underlying distribution of the random effects. It also does not require complete knowledge of the dropout process, as the proposed score test can be applied whether or not the reasons for dropout are known. The test allows for an additional level of clustering among participating subjects, as might be found in a family study.

email: svyrost@sph.emory.edu

HANDLING MISSING RESPONSES IN GENERALIZED LINEAR MIXED MODEL WITHOUT SPECIFYING MISSING MECHANISM

Hui Zhang*, Columbia University
Myunghee Cho Paik, Columbia University

In longitudinal studies, missingness of data is often unavoidable. Valid estimators from the generalized linear mixed model usually rely on the correct specification of the missing data mechanism. When the missing mechanism is incorrectly specified, the estimates may be biased. In this paper, we propose a class of unbiased estimating equations using pairwise conditional method to deal with generalized linear mixed model without specifying the missing mechanism. We show that the proposed estimator is consistent and asymptotically normal. We illustrate the method using longitudinal course of neuropsychcological data.

email: hz2107@columbia.edu

ALTERNATIVES TO TOP-CODING FOR STATISTICAL DISCLOSURE CONTROL

Di An*, University of Michigan
Roderick J.A. Little, University of Michigan

Top-coding of extreme values of variables like income is a common method of statistical disclosure control, but it creates problems for the data analyst. This article proposes two alternative methods to top-coding for SDC based on multiple imputation (MI). We show in simulation studies that the MI methods facilitate better inferences of the publicly-released data than top-coding, using straightforward MI methods of analysis, while maintaining good SDC properties. We illustrate the methods on data from the 1995 Chinese household income project.

email: dianch@umich.edu

BAYESIAN ANALYSIS OF INCOMPLETE DATA IN CROSSOVER TRIALS

Sanjib Basu, Northern Illinois University
Sourav Santra*, Northern Illinois University

Crossover designs are common in clinical trials, especially in bioavailability and bioequivalence studies. The observed data from a crossover trial are often incomplete due to dropout or other reasons. The case of missing data in crossover trials and its analysis are discussed in Jones and Kenward (2003). Chow and Shao (1997) propose transformation of the observed data to eliminate the random subject effects that may influence the missing process. We propose joint models for the observed data and the missing process in a Bayesian set up. We use Bayesian model comparison methods to compare and evaluate missing completely at random (MCAR), missing at random (MAR), informative drop out (ID) and random coefficients models. We illustrate the proposed method in a real example.

email: santra@math.niu.edu

CORRELATION ANALYSIS FOR LONGITUDINAL DATA: APPLICATIONS TO HIV AND PSYCHOSOCIAL RESEARCH

Yan Ma*, University of Rochester
Xin Tu, University of Rochester

Correlation analysis is widely used in biomedical and psychosocial research for characterizing relationships among variables.As longitudinal study designs become increasingly popular, the need for modeling correlations over time with repeated measures data is growing strong.In this paper, we consider non-parametric inference for the product-moment correlation within a longitudinal data setting based on the Pearson correlation estimators. Although methods exist for longitudinal data analysis, none addresses missing data, a very common problem with longitudinal studies. We consider inference of product-moment correlation under both the missing completely at random (MCAR) and missing at random (MAR) assumptions and derive consistent estimators and their asymptotic distributions. The approach is illustrated with real data from several psychiatric and HIV prevention research studies.

email: yan_ma@urmc.rochester.edu

ESTIMATING MEAN COST UNDER DEPENDENT CENSORING

Wenqin Pan*, Duke University
Donglin Zeng, University of North Carolina

We study the estimation of mean medical cost when censoring is dependent and a large amount of auxiliary information is present. Under the missing at random assumption, we propose two working models, which are semiparametric, to obtain condensed covariate information. The estimate of the total cost can be derived nonparametrically using this condensed information. We show that when either working model is correct, the estimator is consistent and asymptotically normal. The asymptotic variance can be consistently estimated using the bootstrapping method. The small-sample performance of the proposed estimator is evaluated via simulation studies. Finally, our approach is applied to a real data set.

email: wendy.pan@duke.edu

23. POWER ANALYSIS AND SAMPLE SIZE

APPROXIMATE CONFIDENCE INTERVALS FOR POWER IN UNIREP ANALYSES

Matthew J. Gribbin*, University of North Carolina
Jacqueline L. Johnson, University of North Carolina
Keith E. Muller, University of Florida

Unlike standard mixed model tests which often have inflated test size, particularly in small samples, and have few methods developed for computing power, tests using the univariate approach to repeated measures (UNIREP) accurately control test size even in small samples and have available accurate power approximations. Muller, Edwards, et al (in review) presents methods that allow for accurate power approximations for the Geisser-Greenhouse, Huynh-Feldt, Box conservative and Uncorrected UNIREP tests. We present new theoretical methods for computing accurate confidence interval approximations for power in UNIREP tests based on the methods developed by Muller, Edwards, et al. We demonstrate our method with simulations of these confidence intervals for a variety of examples, and illustrate the application of the methods using Diffusion Tensor Imaging data.

email: mgribbin@gmail.com

SAMPLE SIZE FOR TUMOR XENOGRAFT STUDIES

Carin J. Kim*, University of Pennsylvania School of Medicine
Daniel F. Heitjan, University of Pennsylvania School of Medicine

The tumor xenograft experiment is a common tool of preclinical cancer research. In a typical experiment, interest is in comparing the rates of growth as a result of treating a tumor with different agents. In many instances, the analysis of a tumor xenograft experiment simply involves the comparison of the shapes of the log tumor volume curves; such a model is at least satisfactory for determining power and sample size. In this article, we develop simple formulas for the sample size needed in a typical two-group experiment with follow-up at equally spaced times, assuming that either a compound symmetry or autocorrelation error structure holds. The formulas involve only the difference in slopes (growth rates), the CV of the raw scale data, the number of follow-up times, Type I and II errors, and the correlation parameter. The simple formulas are applicable even when the CV has to be estimated, with an addition of two animals per group. We also studied that the misspecification of autocorrelation as compound symmetry error model had little effect.

email: carinkim@mail.med.upenn.edu

SAMPLE SIZE CALCULATION FOR THE WILCOXON-MANN-WHITNEY TEST ADJUSTING FOR TIES

Yan Zhao*, Eli Lilly and Company
Dewi Rahardja, University of Indianapolis
Yongming Qu, Eli Lilly and Company

In this paper we study sample size calculation methods for the asymptotic Wilcoxon-Mann-Whitney (WMW) test for data with or without ties. The existing methods are applicable either to data with ties or to data without ties but not to both cases. While the existing methods developed for data without ties perform well, the methods developed for data with ties have limitations in that they are either applicable to proportional odds alternatives or have computational difficulties. We propose a new method which has a closed form formula and therefore is very easy to calculate. In addition, the new method can be applied to both data with or without ties. Simulations have demonstrated that the new sample size formula performs very well because the corresponding actual powers are close to the nominal powers.

email: zhaoyan1@gmail.com

A LIKELIHOOD APPROACH IN SAMPLE SIZE CALCULATION

Yong Chen*, Johns Hopkins University
Charles Rohde, Johns Hopkins University

For given type I and type II error rates, the sample size formula can be induced using Neyman-Pearson theory. From the view of the statistical evidence, the analogues to the type I and type II errors are the probabilities of observing misleading evidence and weak evidence. In general, these two probabilities are functions of sample size. Ideally, we can calculate these two probabilities and control them under some given values. Then we can solve for minimal required sample size. However, these probabilities are hard to calculate. Therefore we study the asymptotic properties of the likelihood ratio and use them to approximate the probabilities of interest. With the idea described above, for the fixed dimensional parametric models with simple hypothesis, we give the sample size formula. In many cases, testing just one component of the vector parameter is of more interest. The use of profile likelihood ratio is justified (i.e. the probabilities of observing misleading evidence and weak evidence can be controlled as small as we want). Consequently, a sample size formula is given for this composite test.

email: yonchen@jhsph.edu

DESIGNING LONGITUDINAL STUDIES TO OPTIMIZE THE NUMBER OF SUBJECTS AND NUMBER OF REPEATED MEASUREMENTS

Xavier Basagana*, Harvard School of Public Health
Donna Spiegelman, Harvard School of Public Health

At the design stage of a longitudinal study, there might be interest in computing the required number of participants (N) to achieve a specific power when the number of repeated measures (r) is fixed a priori, or in computing the required r when N is fixed (e.g. when one wants to perform a longitudinal analysis using subjects who participated in a previous cross-sectional study). When neither N nor r are fixed by design, there is a level curve in (N, r) which gives a fixed amount of power. If C monetary units are available, given the ratio of costs of the first visit vs. the rest, the optimal combination (N, r) that maximizes the power to detect the hypothesized effect, subject to the cost constraint C, can be obtained. We examine this optimal combination and the factors that influence it for longitudinal studies with a continuous response and a binary exposure, both for a exposure difference that is constant over time and for a time by exposure interaction. Three popular correlation structures (compound symmetry, random intercepts and slopes and damped exponential) are studied. Attention is given to specific issues that arise uniquely in observational research.

email: xbasagan@hsph.harvard.edu

A GENERAL APPROACH FOR SAMPLE SIZE AND STATISTICAL POWER CALCULATIONS ASSESSING THE EFFECTS OF INTERVENTIONS USING A MIXTURE MODEL IN THE PRESENCE OF DETECTION LIMITS

Lei Nie*, Georgetown University
Haitao Chu, Johns Hopkins University
Stephen Cole, Johns Hopkins University

A zero-inflated log-normal mixture model with left censoring due to assay measurements falling below detection limits has been applied to compare treatment groups. The sample size calculation has not been studied for this type of data under the assumption of equal proportions of true zeros in the treatment and control groups. In this article, we derive the sample sizes based on the expected differences between the non-zero values of individuals in treatment and control groups. Methods for calculation of statistical power are also presented. When computing the sample sizes, caution is needed as some irregularities occur, namely that the location parameter is sometimes underestimated due to the mixture distribution and left censoring. In such cases, the aforementioned methods fail. We calculated the required sample size for a recent randomized chemoprevention trial estimating the effect of oltipraz on reducing aflatoxin. A Monte Carlo simulation study was also conducted to investigate the performance of the proposed methods. The simulation results illustrate that the proposed methods provide adequate sample size estimates.

email: ln54@georgetown.edu

PREEMPTIVE POWER AND THE CONSULTING STATISTICIAN

David F. Sawrie*, University of Alabama at Birmingham

A statistical experiment often begins in a clinical setting long before consultation with any statistician. Suppose an experimenter identifies a question of interest to the scientific community and collects data to answer it without further design consideration. Once the data reaches statistician in this condition, testing the hypothesis (i.e. the question) is risky. Without a prospective power analysis, the potential null result is difficult to interpret. It may imply the absence of an effect but may also suggest an underpowered hypothesis. This paper introduces a preemptive power method which preempts immediate analysis of the data with an interim decision point. Before any hypothesis test, a preemptive method recasts the circumstance as a special case of an internal pilot design. Using known internal pilot methods to mine the data for an interim variance estimate, statistician projects additional observations, if needed, to achieve pre-specified target power for the experimenter's hypothesis. The experimenter then faces a decision point: (i) collect the additional data and then test the final hypothesis or (ii) immediately conduct the test. This paper outlines and applies the preemptive power method with respect to a fixed sample design under the GLUM assuming fixed effects.

email: dsawrie@uab.edu

24. MULTIVARIATE SURVIVAL, INCLUDING ADJUSTMENT FOR QUALITY OF LIFE

ESTIMATION METHODS FOR MULTIPLE HEALTH-RELATED QUALITY-OF-LIFE ADJUSTED TIMES-TO-EVENT

Adin-Cristian Andrei*, University of Wisconsin-Madison
Fernando J. Martinez, University of Michigan

In clinical trials, particularly those involving aggressive treatment strategies, close attention is paid to the patient health-related quality-of-life (HRQOL). Oftentimes, investigators are interested in both patient survival and HRQOL in connection to specific landmark events, such as disease episodes. The HRQOL-adjusted time-to-event (HRQOL-TTE) is a commonly used measure that provides a comprehensive approach for gaining additional insight into the synergistic aspects of the patient quantity/quality-of-life interplay. Existing methods for HRQOL-TTE only deal with situations where a single HRQOL measure is collected for each patient. However, there are numerous instances when multiple facets of patient's health status are simultaneously of high relevance for the medical investigator, in conjunction with overall survivorship. Using inverse probability of censoring weighting techniques, we propose and fully characterize an estimator for the joint distribution of several HRQOL-TTE. We then evaluate its performance in simulated settings and we illustrate its practical usefullness by presenting a National Emphysema Treatment Trial-related example.

email: andrei@biostat.wisc.edu

NONPARAMETRIC INFERENCE FOR PAIRED QUALITY-OF-LIFE ADJUSTED TIME-TO-EVENT DATA

Kristine L. Cooper, University of Michigan
Susan Murray*, University of Michigan
Hongwei Zhao, University of Rochester

This research makes available a nonparametric quality-of-life adjusted survival method for testing differences in paired censored time to event data. This statistic is based on integrated quality-adjusted survival curves with the standardized test appropriately adjusted for correlation in the estimated curves. Quality-of-life adjusted survival analysis is desirable when there is interest not only in the overall time to event but also in the quality of that time. The asymptotic distribution of the test statistic is given along with variance estimation formulae. We conduct simulations to study finite sample properties of the proposed statistic under both null and alternative hypotheses. We apply this method to a diabetic retinopathy study comparing alternate treatments in paired eyes.

email: skmurray@umich.edu

ON CONSISTENCY OF KENDALL'S TAU UNDER CENSORING

David Oakes*, University of Rochester Medical Center

Necessary and sufficient conditions for consistency of a simple estimator of Kendall's tau under bivariate censoring are presented. The results are extended to data subject to bivariate left truncation as well as right censoring.

email: oakes@bst.rochester.edu

EFFICIENT ESTIMATION FOR THE PROPORTIONAL HAZARDS MODEL

Lianming Wang*, Biostatistics Branch-National Institute of Environmental Health Sciences
Jianguo Sun, University of Missouri
Xingwei Tong, Beijing Normal University

We consider efficient estimation of regression and association parameters for bivariate current status data with the marginal proportional hazards model. Current status data occur in many fields including demographical studies and tumorigenicity experiments and several approaches have been proposed for regression analysis of univariate current status data. We discuss bivariate current status data and propose an efficient score estimation approach for the problem. In the approach, the copula model is used for joint survival function with the survival times assumed to follow the proportional hazards model marginally. Simulation studies are performed to evaluate the proposed estimates and suggest that the approach works well in practical situations. Key words: Bivariate current status data; Efficient estimation; Counting processes; Martingale; Proportional hazards model; Sieve method.

email: apolo2001cn@yahoo.com.cn

ON THE ASSOCIATION MEASURE IN COPULA MODELS WITH APPLICATION TO A HIV STUDY

Suhong Zhang*, University of Iowa
Ying J. Zhang, University of Iowa

Some prior studies suggest that GBV-C, a harmless virus, delays the progression of HIV disease, while some others fail to find the beneficial effect. Those studies compared the Kaplan-Meier estimators of HIV survival function between GBV-C positive and negative patients. Since GBV-C is subject to clearance, the absence of GBV-C at one observation time does not truly reflect whether or not these HIV patients were co-infected with GBV-C, which in turn invalidates the two-sample comparison. We consider the clearance time of GBV-C as an event time in addition to the HIV survival time. Since many studies only had one observation on GBV-C status, the clearance time of GBV-C is observed as current status data. To study the association between HIV survival time and GBV-C persistence time, we proposed a two-stage semiparametric estimator of dependence measure between two correlated event times, in the case that one of the paired event times is subject to right censoring and the other is observed as current status data. We established consistency and normality of the proposed estimator and our simulation studies showed its good performance given a reasonable sample size.

email: suhong-zhang@uiowa.edu

ANALYZING TIME-TO-EVENT DATA FOR A COMPOSITE ENDPOINT HAVING A SILENT COMPONENT

Peng Zhang*, Harvard University School of Public Health
Stephen W. Lagakos, Harvard University School of Public Health

We consider inference for a composite time-to-event endpoint where one component event is only observed periodically and with error. The motivating example is HIV-free survival, that is, time to the earlier of HIV infection or death, in infants born to HIV-infected mothers. Such infants are assessed periodically for HIV infection using an imperfect diagnostic test. As a result, time of occurrence of the composite event is never observed with certainty. Interpretation of such data is further complicated by informative drop-outs that occur prior to all scheduled diagnostic tests. We first determine the identifiable aspects of data arising from such a setting, and develop sharp upper and lower bounds for the distribution of time to the composite endpoint. We then propose likelihood-based methods for estimating the identifiability bounds, and for estimating the distribution of HIV-free survival when drop-outs occur non-informatively. We illustrate the results from a recently-completed clinical trial for the prevention of mother-to-child HIV transmission in Botswana.

email: pzhang@hsph.harvard.edu

THE PROPORTIONAL ODDS MODEL FOR MULTIVARIATE INTERVAL-CENSORED FAILURE TIME

Man-Hua Chen*, University of Missouri-Columbia
Xingwei Tong, Beijing Normal University
Jianguo Sun, University of Missouri-Columbia

The proportional odds model is one of the most commonly used regression models in failure time data analysis and has been discussed by many authors (Bennett,1983; Chen, 2001; Huang and Rossini, 1997; Rabinowita et al., 2000; Yang and Prentice, 1999). It specifies that the covariate effect is a multiplicative factor on the baseline odds function and is often used when, for example, the covariate effect diminishes over time. Most of the existing methods for the model are for univariate failure time data. In this paper, we discuss how to fit the proportional odds model to multivariate interval-censored failure time data. For inference, a maximum likelihood approach is developed and evaluated by simulation studies, which suggest that the method works well for practical situations. The method is applied to a set of bivariate interval-censored data arising from an AIDS clinical trial. KEY WORDS: interval-censoring; marginal inference approach; multivariate failure time data; the proportional odds model

email: manhua50@hotmail.com

25. GENERAL METHODS I

FIDUCIAL INTERVALS FOR VARIANCE COMPONENTS IN AN UNBALANCED TWO-COMPONENT NORMAL MIXED LINEAR MODEL

Lidong E*, Colorado State University
Hannig Jan, Colorado State University
Hari K. Iyer, Colorado State University

Mixed effects linear models are useful in applications that require an understanding of the components of variability arising from multiple sources. For example, in animal breeding studies mixed linear models with two variance components are often used. One variance component accounts for genetic variability and the other accounts for variability due to environmental factors. Our focus in this paper is on unbalanced normal mixed linear models with two variance components. We develop confidence interval procedures for the two variance components and also for the intra-class correlation coefficient based on an extension of R. A. Fisher's fiducial argument. A simulation study is conducted to compare the resulting interval estimates with other competing confidence interval procedures from the literature. Our results demonstrate that the proposed fiducial intervals have satisfactory performance in terms of coverage probability. In addition these intervals have shorter average confidence interval lengths overall. We also prove that these fiducial intervals have asymptotically exact frequentist coverage probability. The computations for the proposed procedures are illustrated using examples from animal breeding applications.

email: e@stat.colostate.edu

MULTIDIMENSIONAL ARRAY-BASED GROUP TESTING IN THE PRESENCE OF TEST ERROR

Hae-Young Kim*, University of North Carolina at Chapel Hill
Michael G. Hudgens, University of North Carolina at Chapel Hill

Berger et al. (Biometrics, 2000) derived the efficiency of multidimensional array-based group testing algorithms without test errors, i.e., no false negative or false positive tests. We extend their results to allow for imperfect testing. In particular, we derive efficiency and pooling measurement error rates such as specificity, sensitivity, positive and negative predictive values, per-family error rate, and per-comparison error rate for three dimensional array-based pooling algorithms when there exist test errors. Algorithms with and without master pool testing are considered. Our results are compared with previously derived operating characteristics for hierarchical and square array-based group testing algorithms in the presence of test errors.

email: hkim@bios.unc.edu

RELATIVITY OF TESTS' OPTIMALITY, WITH APPLICATIONS TO CHANGE POINT DETECTION AND MIXTURE TYPE TESTING

Albert Vexler*, National Institute of Child Health and Human Development
Chengqing Wu, National Institute of Child Health and Human Development
Kai F. Yu, National Institute of Child Health and Human Development

When Statistics was initiated as science, creating new tests was very important issue. Now commonly we have many tests that can correspond to one practical application. Suppose we have K different test statistics based on our data. Which test statistic is appropriate for our study? Investigators used to solve this problem by basing on Monte Carlo simulations. Here we propose that any reasonable test is optimal. The problem is to interpret the optimality of the context of common operating characteristics (e.g. the power of testing etc.). However, if we will be able to obtain areas of K-tests' optimality then by choosing the area, which is close to our objective of a study, we could suggest the most appropriate test. We apply our method to demonstrate a valuable optimality of retrospective Shiryayev-Roberts change point detection. We evaluate and propose optimal mixture type tests that can be very reasonable in the cases when data have limited number of observations.

email: vexlera@mail.nih.gov

FINITE MIXTURE INFERENCE USING THE QUADRATIC INFERENCE FUNCTION

Daeyoung Kim*, The Penn State University
Bruce G. Lindsay, The Penn State University

We investigate the use of quadratic inference function(QIF) methods in finite mixture models. In these methods maximum likelihood methods have optimal efficiency but also multiple failings. For example, the maximum likelihood estimator(MLE) is non-robust in the sense of sensitivity to outliers or errors in the components density specification. When used in a normal mixture model with unequal variances, the likelihood has many singularities. These performance problems are all exacerbated when parameters are near the boundaries. We will show by simulation that the estimators based on a properly constructed QIF behave better than the MLE in these circumstances, all without losing any asymptotic efficiency relative to the MLE.

email: dzk123@psu.edu

STATISTICAL MODELING OF ADVERSE EVENT COUNTS IN CLINICAL TRIALS

Michael A. O'Connell*, Insightful Corporation
Tim Hesterberg, Insightful Corporation
David Henderson, Insightful Corporation

Assessment of drug safety is a primary goal in drug development. However, in contrast to the design and analysis of efficacy endpoints, there is typically little planning or innovative statistical analysis of safety data. In particular, adverse events are usually reported as tables of counts for treatment and control, sometimes accompanied by results from a statistical test that is not well-suited to the data structure. In this paper we present some statistical and graphical analyses of adverse event data that address issues such as the sparse nature of the count data and the many adverse event terms under consideration. Statistical models considered include hierarchical generalized linear models and least angle regression; along with exploratory methods such as random forests. Results from these analyses are presented as S-PLUS graphical summaries that highlight key aspects of drug risks such as risk difference and effect probability for preferred terms and body systems.

email: moconnell@insightful.com

SOME PRACTICAL CONCERNS ON BAYESIAN SAMPLE SIZE CRITERIA ACC AND ALC

Jing Cao*, Southern Methodist University
Jack J. Lee, M D Anderson Cancer Center

One argument against Bayesian sample size determination (SSD) methods is the subjective choice on which criterion to use and how to assign the priors. In this talk, we compare two popular criteria, the average coverage criterion (ACC) and average length criterion (ALC). With the estimation of a single normal mean, we demonstrate that there exists a threshold on the significance level which explains the difference in the ACC and ALC. Furthermore, we investigate the effect of the prior parameters on the Bayesian SSD. Closed form expressions of the true coverage of the posterior credible interval are developed. Based on that, we sugguest some guidelines for choosing the prior parameters.

email: jcao@smu.edu

PROBLEMS WITH EXACT TWO-SIDED TESTS AND THE ASSOCIATED CONFIDENCE INTERVALS FOR DISCRETE DISTRIBUTIONS

Paul W. Vos, East Carolina University
Suzanne S. Hudson*, East Carolina University

Exact confidence intervals for parameters of discrete distributions will be less conservative when defined by inverting an exact test that does not require equal probability in each tail. However, the p-value obtained from such tests can exhibit undesirable properties which in turn results in undesirable properties in the associated confidence intervals. We illustrate these difficulties using p-values for binomial proportions, the difference between binomial proportions, and the parameters of other discrete distributions.

email: hudsons@ecu.edu

26. FUNCTIONAL AND STRUCTURAL NEURO-IMAGING DATA: MODELING AND INFERENCE

BAYESIAN HIERARCHICAL MODELING OF FUNCTIONAL NEUROIMAGING DATA

F. DuBois Bowman*, Emory University
Brian Caffo, Johns Hopkins University

Functional magnetic resonance imaging (fMRI) is a powerful noninvasive tool used to characterize behavior-related neural activity and to investigate regional associations in brain activity. Functional neuroimaging is also useful for addressing important scientific questions regarding differences in distributed neural processing between subgroups, e.g. cocaine addicts and healthy controls. Many challenges are involved in analyzing functional neuroimaging data including the massive amount of data collected, the large number of spatial locations (voxels), and the complex patterns of spatial and temporal correlations, to name a few. In this talk, we develop a modeling approach in which the second stage specifies a Bayesian hierarchical model for fMRI data that captures spatial correlations between voxels within a given region as well as inter-regional correlations. We demonstrate the applicability of our model using a study of response inhibition among cocaine-dependent subjects. We also explore functional connectivity between different brain regions associated with the response inhibition task.

email: dbowma3@sph.emory.edu

LEAD EXPOSURE, BEHAVIOR AND NEURONAL VOLUME

Brian S. Caffo*, Johns Hopkins University
Sining Chen, Johns Hopkins University
Brian Schwartz, Johns Hopkins University

Neuronal volumes in adults are determined by the convolution of several complex genetic and environmental processes, including the cross-sectional differences in intrinsic brain size, longitudinal loss in volume due to aging, the presence or absence of neuronal diseases and longitudinal volume loss due to environmental exposure to toxicants. In this talk we explore the complex relationships between (i) neuronal volume, as measured by magnetic resonance images (MRIs), (ii) behavior, as measured by a battery of neuro-behavioral tests and (iii) lead exposure, as measured by peak tibia lead measurements. The population under study is comprised of former organo-lead workers and controls, resulting in over 600 MRI images. Techniques of voxel based morphometry (VBM) will be used to address the causal relationships between these processes. Particular attention will be paid to the potential mediating effect of structure on the relationship between lead exposure and function. Due to the complexity of the questions and the high-throughput nature of the MRI data, novel statistical methods will be employed to address the research goals.

email: bcaffo@jhsph.edu

IMPROVING CLUSTER-WISE INFERENCE WITH DIFFERENT TYPES COMBINED STATISTICS

Hui Zhang*, University of Michigan
Thomas E. Nichols, GlaxoSmithKline
Timothy D. Johnson, University of Michigan

In fMRI data analysis inference is based on either cluster size or voxel intensities. Cluster-size methods are sensitive to spatially extended signals, while voxel-wise methods are sensitive to signal magnitude. Others have ave proposed combining inferences from cluster size and voxel intensitie. Alternatively, cluster mass is another way to combine the cluster size and intra-cluster average height, and evidence suggests it to be more powerful than other combining methods. We examine the properties of different combining statistics using permutation methods when the asymptotic distribution of these statistics are unknown. We also develope random field theory (RFT) for the asymptotic distribution of cluster mass under certain circumstances. We also derive the joint distribution of cluster mass and intra-cluster peak intensity.

email: huizhang@umich.edu

27. METHODS FOR VACCINE TRIALS WITH RARE EVENTS, SMALL SAMPLE SIZES, AND MISSING DATA

ON COMPARING INFECTEDS IN RANDOMIZED VACCINE TRIALS

Dean A. Follmann*, NIAID
Michael Fay, NIAID
Michael Proschan, NIAID

This paper proposes new tests to compare the vaccine and placebo groups in randomized vaccine trials when a small minority of volunteers become infected. The Burden of Illness test of Chang, Guess, & Heyse test assigns 0s to the uninfecteds, a severity score X>0 to the infecteds and compares the mean of this lumpy random variable between the two groups. But this approach can have poor power if infections are rare. Hudgens, Hoering, & Self introduced tests of the equality of X in the principle stratum of those who are doomed to be infected. While this can be more powerful than the BOI test, it requires the assumption that the vaccine cannot cause infections in anyone. We suggest new tests that can be more powerful than other approaches, and do not require the assumption that the vaccine harms no one. The basic idea is to toss out an equal number of zeros from both groups and then perform a test on the remaining data which is mostly Xs >0. A permutation approach is used to derive a null distribution. The tests are compared to other procedures via simulation.

email: dfollmann@niaid.nih.gov

A PROTOTYPE PROOF-OF-CONCEPT TRIAL DESIGN FOR CELL MEDIATED IMMUNITY-BASED VACCINES

Devan V. Mehrotra*, Merck Research Laboratories

Almost all vaccine efficacy trials to date have evaluated vaccines designed to protect against clinical infection through antibody-based immunity. A new class of vaccines designed to elicit cell mediated immunity (CMI) are being developed. It is hypothesized that vaccine-induced CMI responses will either prevent infection, or will help contain the post-infection pathogen load at a low level in those who become infected despite vaccination. The large uncertainty about the efficacy of such vaccines motivates conduct of relatively small proof-of-concept (POC) trials to establish biological efficacy, before qualifying them for potential large-scale licensure trials. A prototype POC efficacy trial design will be presented, with emphasis on the statistical strategy for establishing POC in a timely and resource-efficient manner. The novel design was recently implemented for evaluating a CMI-based HIV vaccine in a landmark clinical trial. (This is joint work with Xiaoming Li, Peter Gilbert and other collaborators from Merck and the HIV Vaccine Trials Network.)

email: devan_mehrotra@merck.com

EVALUATION OF MULTIPLE IMPUTATION IN AN VACCINE IMMUNOGENICITY TRIAL

Michela Baccini, University of Florence
Constantine E. Frangakis, Johns Hopkins University
Fan Li, Harvard Medical School
Fabrizia Mealli, University of Florence
Brian D. Plikaytis, Centers for Disease Control and Prevention
Charles E. Rose, Jr., Centers for Disease Control and Prevention
Donald B. Rubin, Harvard University
Elizabeth R. Zell*, Centers for Disease Control and Prevention

In randomized trials involving the evaluation of the effectiveness of a vaccine with repeated vaccinations over time, it is almost inevitable that some people will skip scheduled visits or drop out before the end of the trial. Ideally, data would be collected even though violations of the protocol have occurred, but it is essentially impossible to obtain immunogenicity data, such as antibody levels, or reactogenicity data, such as negative side effects, for missed visits or after dropout. Thus, some sort of imputation, either implicit or explicit, is required if true intention-to-treat (ITT) analyses are to be conducted using the full randomized population. Because ITT analyses are the standard ones that are validated by the randomization, it is highly desirable to do some sort of imputation for the missing data. Multiple imputation has been shown to be valid, or nearly so, under a broad range of circumstances. However, the current Anthrax Vaccine Trial being conducted presents particular challenges because of the very large number of measurements on each person (e.g., nearly 2000) and the limited number of subjects in each treatment arm (e.g., about 200). This talk will present preliminary results from a systematic large-scale evaluation of the validity of the multiple imputations in this trial, which were created using a complex multiple imputation scheme that is viewed as state of the art.

email: ezr1@cdc.gov

MODELING VACCINE ADVERSE EVENT COUNT DATA USING ZERO-INFLATED AND HURDLE MODELS

Charles E. Rose, Jr.*, Centers for Disease Control and Prevention
Stacey W. Martin, Centers for Disease Control and Prevention
Kathleen A. Wannemuehler, Centers for Disease Control and Prevention
Brian D. Plikaytis, Centers for Disease Control and Prevention

We compared several modeling strategies for vaccine adverse event count data in which the data are characterized by excess zeroes and heteroskedasticity. Count data are routinely modeled using Poisson and Negative Binomial (NB) regression but zero-inflated and hurdle models may be advantageous in this setting. Here we compared the fit of the Poisson, Negative Binomial (NB), zero-inflated Poisson (ZIP), zero-inflated Negative Binomial (ZINB), Poisson Hurdle (PH), and Negative Binomial Hurdle (NBH) models. In general, for public health studies, we may conceptualize zero-inflated models as allowing zeroes to arise from at-risk and not-at-risk populations. In contrast, hurdle models may be conceptualized as having zeroes only from an at-risk population. Choosing between the zero-inflated and hurdle modeling framework, assuming Poisson and NB models are inadequate because of excess zeroes, should generally be based on the study design and purpose. For example, if the study design leads to count endpoints with both structural and sample zeroes then generally the zero-inflated modeling framework is more appropriate, while in contrast, if the endpoint of interest, by design, only exhibits sample zeroes (e.g., at-risk participants) then the hurdle model framework is generally preferred.

email: cvr7@cdc.gov

28. RECENT ADVANCES IN REGRESSION MODELING WITH SURVIVAL DATA

ON GENERAL TRANSFORMATION MODEL FOR CENSORED DATA

Zhezhen Jin*, Columbia University

This talk will focus on the estimation and inference for right censored data based on general transformation models which are distribution free and with unknown monotone transformation. The rank-type estimation method will be presented along with a simple variance estimation.

email: zj7@columbia.edu

SURVIVAL ANALYSIS WITH TEMPORAL COVARIATE EFFECTS

Limin Peng*, Rollins School of Public Health-Emory University
Yijian Huang, Rollins School of Public Health-Emory University

In chronic disease studies, covariate effects can vary over time. For example, the effect of a treatment may diminish due to the development of drug resistance. Unfortunately, standard proportional hazards models do not accommodate such temporal effects. In this paper, we propose a natural generalization of the Cox regression model, where the regression coefficients have direct interpretations as temporal covariate effects on the survival function. Under the conditional independent censoring mechanism, we develop an estimation procedure based on a set of martingale-based equations without involving smoothing. Our estimator is shown to be uniformly consistent and converge weakly to a Gaussian process. A simple resampling method is proposed to approximate the limiting distribution of the estimated coefficients. Second-stage inferences with time-varying coefficients are developed accordingly. Simulations and two real examples illustrate the practical utility of the proposed method. Finally, we extend this proposal of temporal covariate effects for the general class of linear transformation models and also establish a connection with the additive hazards model.

email: lpeng@sph.emory.edu

REGRESSION ANALYSIS OF RECURRENT EPISODES DATA: THE LENGTH-FREQUENCY TRADEOFF

Jason P. Fine*, University of Wisconsin-Madison
Jun Yan, University of Iowa

I consider a special type of recurrent event data, ‘recurrent episode data’ in which when a event occurs it last for a random length of time. Recurrent episode data arise frequently in studies of episodic illness. A naive recurrent event analysis disregards the length of each episode, which may contain important information about the severity of the disease, as well as the associated medical cost and quality of life. Analysis of recurrent episode data can be further complicated if the effects of treatment and other progrnostic factors are not constant over the observation period, as occurs when the covariate effects vary across episodes. I will review existing methods applied to recurrent episode data and approach the length-frequency tradeoff using recently developed temporal process regression. Novel endpoints are constructed which summarize both episode length and frequency. Time varying coefficient models are proposed, which capture time varying covariate effects. New and existing methods are compared on data from a clinical trial to assess the efficacy of a treatment for cystic fibrosis patients experiencing multiple pulmonary exacerbations.

email: fine@biostat.wisc.edu

ANALYSIS OF RECURRENT-MARKER PROCESS DATA USING MODEL-BASED NONPARAMETRIC AND SEMIPARAMETRIC MODELS

Mei-Cheng Wang*, Bloomberg School of Public Health-Johns Hopkins University

This paper considers analysis of recurrent-marker process data in the case that marker measurements are collected at the times when recurrent events occur. A recurrent-marker process is defined using both recurrent events and markers where these two kinds of measurements are possibly correlated. A mean function (MF) at t is defined as the unit-time average of the sum of marker measurements at t. To model the recurrent-marker process, we consider a multiplicative mean model of the mean function. A model-based nonparametric estimator is introduced and its properties studied. A two-step regression model is introduced for modelling recurrent events in the first step and marker measurements in the second step.

email: mcwang@jhsph.edu

29. STATISTICAL SAFETY ANALYSIS OF TIME-DEPENDENT ENDPOINTS IN CLINICAL TRIALS

CHANGE-POINT ANALYSIS AND SURVIVAL DATA

Ramin B. Arani*, Bristol Myers Squibb Company
Seng-Jaw Soong, Comprehensive Cancer Center-University of Alabama at Birmingham

The change-point problem has been studied in a variety of contexts. Most methods utilize a variation of likelihood ratio test, which can be computationally intensive . In most case, estimation of standard errors of the change-point can be tedious.. A general review of analysis methods will be provided, the context generally will focus on survival data. We consider a simple case of single change-point, t0, i.e., a piece-wise constant hazard function. Clearly, the likelihood function is not differentiable at the change-point, therefore MLEs can be obtained through intensive computational methods. We propose an alternative method by approximating the likelihood function by a continuous transformation of Sigmoidal function, /s/(t) = 1/[1+exp(t-t0)]. The difference between the likelihood function and approximate function converges to zero almost every-where. Therefore, consistency of the estimators from the approximate likelihood function is followed. The application of the proposed method in analysis and design of clinical trials will be illustrated.

email: ramin.arani@bms.com

ESTIMATING THE LOCATION AND CONFIDENCE INTERVAL OF UNKNOWN CHANGE POINTS IN HAZARD RATE MODELS WITH AND WITHOUT A BINARY GROUPING COVARIATE

Sandra L. Gardner*, Sunnybrook Health Sciences Centre
Rafia Bhore, Food and Drug Administration

In the statistical literature related to a hazard rate model with a single change point, the parametric bootstrap has been proposed as a method of estimating the confidence interval for the unknown change point. This talk will explore different types of bootstrap confidence intervals through simulations and application to the clinical trial setting. The single change point model and bootstrapping methods will be extended to test if two distinct groups of patients, for example patients from two different sites, have different change points.

email: sandra.gardner@sunnybrook.ca

TESTING FOR CHANGE-POINTS IN ANALYZING TIME-TO-EVENT ENDPOINTS

Thomas Hammerstrom, Food and Drug Administration
Rafia Bhore*, Food and Drug Administration
Mohammad Huque, Food and Drug Administration

This talk will present the results of some simulations on fitting and estimating parameters of Piecewise Exponential (PWE) survival models with change points for time-to-event data. The focus will be on two problems: 1) the accuracy of the likelihood ratio test in choosing the correct model for the times-to-event among exponential, PWE with one change-point, PWE with two change-points, and Weibull; and 2) comparison of true and nominal coverage levels of confidence intervals for parameters computed by bootstrap and parametric methods. Finally we illustrate application of change-point methods with an example of how to analyze time-dependent adverse events in clinical trials (such as nephrotoxicity, liver function tests, or cardiovascular events).

email: rafia.bhore@fda.hhs.gov

30. STATISTICS EDUCATION IN K-12 AND ITS POTENTIAL IMPACTS

A FRAMEWORK FOR TEACHING AND LEARNING STATISTICS IN ELEMENTARY GRADES

Christine A. Franklin*, University of Georgia

A major goal of the report 'A Curriculum Framework for Pre K-12 Statistics Education' is to describe a statistically literate high school graduate and, through a connected curriculum, provide guidelines to achieve this goal. In this session the recommendations from the Framework related to statistics for the elementary grades and their connections to statistical concepts at the middle and secondary levels will be presented. Activities appropriate for integrating statistics in the elementary curriculum and transitioning to the middle school curriculum will be included in this session.

email: chris@stat.uga.edu

A FRAMEWORK FOR TEACHING AND LEARNING STATISTICS IN MIDDLE GRADES

Gary D. Kader*, Appalachian State Univeristy

A major goal of the report ‘A Curriculum Framework for Pre K-12 Statistics Education’ is to describe a statistically literate high school graduate and, through a connected curriculum, provide guidelines to achieve this goal. The recommendations from the Framework related to statistics for the middle grades and their connections to statistical concepts at the elementary and secondary levels will be presented. Activities appropriate for integrating statistics in the middle school curriculum and transitioning to the secondary school curriculum will be included in this session.

email: gdk@math.appstate.edu

AN UPDATE ON AP STATISTICS

Linda J. Young*, University of Florida

In 1997, the AP statistics exam was offered for the first time. The 7,500 students who took that exam was a record number for any first-year offering of an AP exam. The program has grown rapidly, with almost 90,000 students taking the exam in 2006. The curriculum of AP statistics will be reviewed. The potential impact of AP statistics on the undergraduate curriculum will be discussed.

email: LJYoung@ufl.edu

31. VARIABLE SELECTION FOR HIGH DIMENSIONAL DATA

OBJECTIVE BAYES VARIABLE SELECTION: SOME METHODS AND SOME THEORY

George Casella*, University of Florida

A fully automatic Bayesian procedure for variable selection in normal regression model has been developed, which uses the posterior probabilities of the models to drive a stochastic search. The posterior probabilities are computed using intrinsic priors, which are default priors as they are derived from the model structure and are free from tuning parameters. The stochastic search is based on a Metropolis-Hastings algorithm with a stationary distribution proportional to the model posterior probabilities. The performance of the search procedure is illustrated on both simulated and real examples, where it is seem to perform admirably. However, until recently such data-based evaluations were the only performance evaluations available. It has long been known that for the comparison of pairwise nested models, a decision based on the Bayes factor produces a consistent model selector (in the frequentist sense), and we are now able to extend this result and show that for a wide class of prior distributions, including intrinsic priors, the corresponding Bayesian procedure for variable selection in normal regression is consistent in the entire class of normal linear models. The asymptotics of the Bayes factors for intrinsic priors are equivalent to those of the Schwarz (BIC) criterion, and allow us to examine what limiting forms of proper prior distributions are suitable for testing problems. Intrinsic priors are limits of proper prior distributions, and a consequence of our results is that they avoid Lindley's paradox. The asymptotics further allow us to examine some selection properties of the intrinsic Bayes rules, where it is seen that simpler models are clearly preferred. This is joint work with Javier Girón, Lina Martínez, and Elías Moreno.

email: casella@ufl.edu

SPIKE AND SLAB VARIABLE SELECTION

Hemant Ishwaran*, Cleveland Clinic

Rescaled spike and slab models are a new Bayesian approach to selecting variables in high dimensional models. I discuss the importance of prior hierarchical specifications in such models and draw connections to frequentist methods in order to motivate the approach as well as develop theory. Applications of the method include high-throughput genomic data. While it is customary to approach such data through the use of hypothesis testing, I show a variable selection paradigm using rescaled spike and slab models offers unique advantages. Several examples will be presented.

email: hemant.ishwaran@gmail.com

HIERARCHICAL SPARSITY PRIORS IN HIGH-DIMENSIONAL VARIABLE SELECTION

Joe Lucas, Duke University
Carlos Carvalho, Duke University
Mike West*, Duke University

We discuss the development and application of a novel class of sparsity models for problems of variable selection and uncertainty. Motivated by problems of highly multivariate analysis - including multivariate regression, anova and latent factor models - we have been led to extend traditional ‘pont mass mixture’ prior models to a richer class of hierarchical shrinkage models that deal much more adequately with the signal extraction and implicit multiple comparison problems in high-dimensional models. Our presentation here will discuss and exemplify the approach, covering key aspects of model specification and stochastic computation for a range of contexts involving ‘big’ and ‘sparse’ models. The examples will be drawn from large-scale multivariate anova, regression and related sparse latent factor models in number of current applications in gene expression genomics.

email: amy_formike@hotmail.com

---

32. SPECIAL CONTRIBUTED SESSION:

C. FREDERICK MOSTELLER: BIOSTATISTICAL SCIENTIST -- EDUCATOR – MENTOR

PANEL DISCUSSION

Moderator: Marvin Zelen, Harvard University

33. CLINICAL TRIALS

ESTIMATING THE CURRENT TREATMENT EFFECT WITH HISTORICAL CONTROL DATA

Zhiwei Zhang*, Food and Drug Administration

When a randomized, concurrently controlled study is unethical or impractical, researchers often turn to a single-armed, historically controlled study (HCS) as a practical alternative. Causal inference in an HCS is usually carried out using methods designed for a typical observational study (TOS). This paper points out the differences between a TOS and an HCS and attempts to conceptualize the latter in clinically meaningful terms. In particular, it is noted that the current treatment effect, the average of individual treatment effects over the patient population for the current study, may be a more relevant estimand than the quantity estimated by standard TOS methods, which is a weighted average of individual effects over current and historical patients. The current treatment effect can be estimated under an outcome regression model or a covariate density model, the latter corresponding to a propensity score model from a TOS perspective. Augmenting both estimators leads to a doubly robust estimator that is consistent and asymptotically normal if either model is correctly specified. Simulation experiments are conducted to evaluate the finite sample performance of the proposed methods. Practical recommendations are given in regard to the design and analysis of an HCS.

email: zhiwei.zhang@fda.hhs.gov

THE EFFECTS OF AN INACCURATE MARKER ON CLINICAL TRIALS

Nancy Wang, University of California-Berkeley
Nusrat Rabbee*, Genentech, Inc.

This talk explores the impact of an inaccurate diagnostic marker in clinical trials. Previous papers have focused on the effect of incorporating a diagnostic marker with a specified prevalence into Phase II or III trials. Here we explore the scenario where the marker is an imperfect surrogate for a true clinical biomarker, by introducing inaccuracy metrics like sensitivity and specificity. We simulate settings that are close to Phase II trials with a diagnostic marker. We investigate how the confidence intervals, and point estimates of the hazard ratio between treatment and control in the diagnostic positive group is impacted by marker inaccuracy. Our preliminary results show the marker's impact on clinical decision making. We introduce a method for correcting errors in hazard ratio estimates.

email: nrabbee@gene.com

THE EFFECT OF INTERIM SAFETY MONITORING ON END-OF-TRIAL ESTIMATES OF RISK

Michael J. Dallas*, Merck Research Laboratories

Upon termination of a clinical trial that uses interim evaluations to determine whether the trial can be stopped, the proper statistical analysis must account for the interim evaluations. While it is standard practice to adjust terminal statistical analyses due to opportunities to stop for "positive" findings, e.g., group-sequential efficacy trials, adjusting due to opportunities to stop for "negative" findings should also be considered. Here I present a method to account for such designs in the context of a safety trial. The method is demonstrated to have advantages over naive (i.e., unadjusted) methodology that is commonly employed with respect to estimates of risk.

email: michael_dallas@merck.com

ON TREATMENT SELECTION IN ACCELERATED DRUG DEVELOPMENT

Ying Wan*, Temple University

This paper describes a method for the design of a clinical trial to combine phase 2 and 3 of clinical development. The method provides a two-stage study design allowing a fixed or data dependent number of treatment arms to be selected for stage 2. An example is given to illustrate the method and to examine advantages of the two-stage flexible design. We study some key questions to find a good strategy for detecting superior and statistically significant treatment(s) in consideration of limited resources for drug development. Given a prespecified fixed sample size, practical recommendations are made on: whether to select one arm or multiple arms, whether to expand the number of experimental arms in a trial, and what is the appropriate timing of interim analysis.

email: ying_wan@merck.com

ANALYSIS OF BINARY OUTCOMES IN NEONATAL CLINICAL TRIALS WITH TWIN BIRTHS

Michele L. Shaffer*, Penn State College of Medicine
Allen R. Kunselman, Penn State College of Medicine

In neonatal trials of pre-term or low-birth-weight infants, twin births can represent an appreciable percentage (10-20%) of the study sample. Several methods exist for analyzing binary outcomes where related data are present. However, it is unclear which existing methods, if any, work well for mixes of correlated and independent data over a range of correlations. Simulation studies are conducted to compare mixed-effects models and generalized estimating equations for binary outcomes in two-armed clinical trials. Data from the Randomized Clinical Trial of Intravenous Immune Globulin to Prevent Neonatal Infection in Very-Low-Birth-Weight Infants conducted by the National Institute of Child Health & Human Development Neonatal Research Network are used for illustration.

email: mshaffer@hes.hmc.psu.edu

APPLICATION OF NEW LEVENE TYPE TESTS FOR HYPOTHESES ABOUT DISPERSION DIFFERENCES IN CLINICAL DATA ANALYSIS

Xiaoni Liu*, Bristol Myers Squibb Company
Dennis Boos, North Carolina State University
Cavell Brownie, North Carolina State University

Dispersion is sometimes more relevant than means in clinical data analysis. An example is to compare variability in response for different formulations of a drug, especially for drugs with small therapeutic windows. Levene type tests are well known to be robust tests for equality of scale in one-way designs. Current Levene type tests for the two-way design use ANOVA F tests on the absolute values of least squares-based residuals. In this paper, we introduce briefly some new Levene type tests for two-way RCB designs that can assess equality of variances across blocks and/or treatments. We use ANOVA F tests on the absolute values of residuals obtained from robust Huber type fits. We also apply bootstrap methods to these Levene type tests and compare the tests in terms of robustness and power by simulation. In addition, we apply these new methods to one-sequence crossover designs and to randomized crossover designs that are popular in Phase I clinical trials.

email: xiaoni.liu@bms.com

THE EFFECT OF HERD IMMUNITY ON VACCINE TRIAL DESIGN

Blake F. Charvat*, Johns Hopkins Bloomberg School of Public Health
Ronald Brookmeyer, Johns Hopkins Bloomberg School of Public Health
Jay Herson, Johns Hopkins Bloomberg School of Public Health

Clinical trials for infectious disease vaccines involve randomizing subjects within communities to either receive experimental vaccine or placebo. Design of such trials frequently ignore herd immunity, that is, the tendency of disease incidence to decrease both among vaccinated and placebo subjects because subjects are less likely to come into contact with infected individuals. This paper uses computer simulation of an epidemic disease transmission model to address the effect of herd immunity on size and power of a vaccine trial for k:1 assignment. The model allows for heterogeneity in disease transmission rates within and between clusters of persons. We find that naive power calculations that do not account for herd immunity when designing large scale community based vaccine trials can seriously underestimate the power of the study. In fact, we find the surprising result that in some circumstances increasing intra-community study enrollment can actually lead to decreased power. The effect of disease incidence on trial feasibility will also be presented.

email: bcharvat@jhsph.edu

34. MICROARRAY ANALYSIS I

A CLASSIFICATION METHOD USING THE FIEDLER VECTOR, WITH APPLICATION TO MICORARRAY DATA

Choongrak Kim*, Pusan National University-Pusan, South Korea
Minkyung Oh, Pusan National University-Pusan, South Korea
Soonphil Hong, Pusan National University-Pusan, South Korea
Eunyeong Yoon, Pusan National University-Pusan, South Korea
Jinmee Kim, Pusan National University-Pusan, South Korea

The eigenvector corresponding to the nonzero smallest eigenvalue of the Laplacian matrix is called the Fiedler vector in graph theory. In this paper, we suggest a classification method using the Fiedler vector. To be more specific, we first choose informative genes using the F-test, and construct the Laplacian matrix based on the selected genes. A direct application of the Fiedler vector to raw data results in poor classification due to the generic errors in data, so that we use the hard-thresholding method to refine data. To estimate the thresholding parameter, we minimize the misclassification rates in training. We applied this method to two real microarray data sets: the leukemia data (Golub et al. 1999) and the lymphoma data (Alizadeh et al. 2000), and evaluate misclassification rates using the leave-one-out cross-validation. The proposed method gave very good classification results.

email: crkim@pusan.ac.kr

STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENE COMBINATIONS

Yen-Yi Ho*, Johns Hopkins University
Leslie Cope, Johns Hopkins University
Marcel Dettling, Zurcher Hochschule-Winterthur, Switzerland
Giovanni Parmigiani, Johns Hopkins University

Identification of coordinate gene expression changes across phenotypes or biological conditions is the basis of our ability to decode the role of gene expression regulatory networks. Statistically, the identification of these changes can be viewed as a search for groups (most typically pairs) of genes whose expression provides better phenotype discrimination when considered jointly than when considered individually. We define such groups as being jointly differentially expressed. In this paper, we proposed a novel entropy-based method for identifying jointly differentially expressed groups of genes using microarray data. We also compare the power of the entropy-based measure with other available approaches using computer simulations.

email: yho@jhsph.edu

PROXIMITY MODEL FOR EXPRESSION QUANTITATIVE TRAIT LOCI (EQTL) DETECTION

Jonathan A. Gelfond*, University of North Carolina
Joseph G. Ibrahim, University of North Carolina
Fei Zou, University of North Carolina

Expression Quantitative Trait Loci (eQTL) are loci or markers on the genome that are associated with gene expression. It is well known to biologists that some (cis) genetic influences on expression occur over short distances on the genome while some (trans) influences can operate remotely. We use a log-linear model to place structure on the prior probability for genetic control of a transcript by a marker locus so that the loci that are closest to a transcript are given a higher prior probability of controlling that transcript to reflect the important role that genomic proximity can play in the regulation of expression. This Proximity Model is an extension of the mixture over marker (MOM) model for the simultaneous detection of cis and trans eQTL of Kendziorski et al. (Biometrics 2006, Volume 62). The genomic locations of transcripts are used to improve the accuracy of the posterior distribution for the location of the eQTL. We compare the MOM method to our extension with both simulated data and datasets of recombinant inbred mouse lines. We also discuss an extension of the MOM method to model multiple eQTLs, and find many transcripts are likely associated with more than one eQTL.

email: jgelfond@bios.unc.edu

SEQUENTIAL SAMPLE SIZE FOR MICROARRAYS

Rossell David*, Rice University - MD Anderson Cancer Center
Peter Müller MD Anderson Cancer Center

Calculating the sample size for an experiment typically requires defining one or a few primary endpoints and also having some a priori knowledge about what values are expected for those endpoints. This can be specially challenging in the microarray context, in which observations are obtained for thousands of genes simultaneously and the experimenter does not usually know what to expect. In this context it seems natural to work sequentially. That is, instead of fixing the sample size in advance we collect data repeatedly over time until at some point we decide to stop. This stopping decision is taken so that it maximizes some user-defined utility function. We solve the maximization via forward simulation, which in principle allows the approach to work for any utility function and probability model. In practice it is convenient to use models of mild complexity to avoid an excessive computational burden. We illustrate the performance of our approach on real data.

email: rusi@rice.edu

ESTIMATION OF SURVIVAL TIME USING CDNA MICROARRAY DATA

Jimin Choi*, Pusan National University
Minho Kang, Pusan National University
Choongrak Kim, Pusan National University

A typical setting of microarray data set is that the number of genes, p, far exceeds the number of samples, n, which is known as ‘small n, large p problem.’ Standard statistical methodologies can be implemented in the reduced subspace. For this reason, dimension reduction technique, for example principal components analysis or partial least squares, is needed prior to analysis of interest. However, such methods still have the difficulty in understanding between each gene and response variable. In an effort to make interpretation easier, a new method to predict survival time of an individual using weighted means of gene intensities is suggested. In this paper, we estimate median survival time, which is used as a predicted survival time, using the Cox regression model. We apply the proposed algorithm to two real data sets: the diffuse large B-cell lymphoma (Alizadeh et al. 2000) and the breast carcinoma (Sorlie et al. 2001). The prediction results using the leave-one-out cross validation are quite good if the censoring percentage is not too high.

email: jmchoi97@yahoo.co.kr

EXPLORATION, NORMALIZATION, AND GENOTYPE CALLS OF HIGH DENSITY OLIGONUCLEOTIDE SNP ARRAY DATA

Benilton S. Carvalho*, Johns Hopkins University
Henrik Bengtsson, University of California-Berkeley
Terence P. Speed, WEHI, Melbourne, Australia and University of California-Berkeley
Rafael A. Irizarry, Johns Hopkins University

In most microarray technologies, a number of critical steps are required to convert raw intensity measurements into the data relied upon by data analysts, biologists and clinicians. These data manipulations, referred to as preprocessing, can influence the quality of the ultimate measurements. In the last few years, the high-throughput measurement of gene expression is the most popular application of microarray technology. For this application, various groups have demonstrated that the use of modern statistical methodology can substantially improve accuracy and precision of gene expression measurements, relative to ad-hoc procedures introduced by designers and manufacturers of the technology. Currently, other applications of microarrays are becoming more and more popular. In this paper we describe a preprocessing methodology for a technology designed for the identification of DNA sequence variants in specific genes or regions of the human genome that are associated with phenotypes of interest such as disease. In particular we describe methodology useful for preprocessing Affymetrix SNP chips and obtaining genotype calls with the preprocessed data. We demonstrate how our procedure improves existing approaches using data from three relatively large studies including one in which large numbers of independent calls are available.

email: bcarvalh@jhsph.edu

MULTI-CLASS CANCER OUTLIER DIFFERENTIAL GENE EXPRESSION DETECTION

Fang Liu*, University of Minnesota
Baolin Wu, University of Minnesota

It has recently been shown that cancer genes (oncogenes) tend to have heterogeneous expressions across disease samples. So it is reasonable to assume that in a microarray data only a subset of disease samples will be activated (these activated samples are often referred to as outliers), which presents some new challenges for statistical analysis. In this paper, we study the multi-class cancer outlier differential gene expression detection. Statistical methods will be proposed to take into account the expression heterogeneity. Through simulation studies and application to public microarray data, we will show that the proposed methods could provide more comprehensive analysis results and improve upon the traditional differential gene expression detection methods, which often ignore the expression heterogeneity and may loss power.

email: liuxx395@umn.edu

35. GENERAL METHODS: CATEGORICAL AND SURVIVAL DATA

IMPACT OF FAILURE TO PROPERLY ADJUST FOR ONSET OF A TIME-DEPENDENT EXPOSURE

Ayumi K. Shintani*, Vanderbilt University
Patrick G. Arbogast, Vanderbilt University
E. Wesley Ely, Vanderbilt University

When analyzing the effect of a time-varying exposure, it is well-known that bias can occur when ignoring the onset of the exposure. However, the degree of bias is not well understood by the clinical community. We examined the potential magnitude of bias when the onset of a time-varying exposure was not accounted and illustrated this in a study of ICU-associated delirium. The cumulative probability of ICU discharge was computed using modified Kaplan-Meier estimates adjusting for delirium onset (Simon and Makuch, Stat in Med 1984). The corresponding log-rank test indicated no association between delirium and time to ICU discharge (p=0.97), and results were similar with a time-dependent Cox regression model (HR=1.1, 95% CI 0.7-1.7, p=0.65, as reported by Ely et al., JAMA 2004). However, when delirium was categorized as ever vs. never (i.e., time-invariant), delirium was associated with prolonged time to ICU discharge (HR= 1.8 (1.2, 2.6), p=0.002). In an independent simulation study with 1000 replicates, the median (95% CI) of simulated hazard ratios was 12.7 (7.7, 24.9) from time-invarying Cox regression whereas 1.0 (0.6, 1.6) from time-varying Cox regression. Ignoring the onset of the exposure in the analysis of time-varying exposure may lead to severe bias with falsely positive findings.

email: ayumi.shintani@vanderbilt.edu

MAXIMUM LIKELIHOOD ESTIMATION FOR TIED SURVIVAL DATA UNDER COX'S REGRESSION MODEL VIA THE EM ALGORITHM

Thomas H. Scheike, University of Copenhagen
Yanqing Sun*, University of North Caroline-Charlotte

We consider tied survival data based on Cox's proportional regression model. The standard approaches are the Breslow and Efron approximations and various so called exact methods. All these methods lead to biased estimates when the true underlying model is in fact a Cox model. In this paper we review the methods and suggest a new method based on the missing-data principle using the EM-algorithm that is rather easy to implement, and leads to a score equation that can be solved directly. This score has mean zero. We also show that all the considered methods have the same asymptotic properties and that there is no loss of efficiency when the tie sizes are bounded or even converge to infinity at a given rate. A simulation study compares the finite sample properties of the methods.

email: yasun@uncc.edu

FRACTIONAL IMPUTATION METHODS FOR MISSING CATEGORICAL DATA

Michael D. Larsen*, Iowa State University
Shinsoo Kang, KwanDong University
Kenneth J. Koehler, Iowa State University

Fractional imputation (FI) fills-in missing values with two or more donor values, assigns replicate weights to the multiple donor values, and uses replication methods to estimate variances. FI is therefore similar to both multiple imputation in that it represents uncertainty due to the value being missing through multiple possible values and hot deck imputation in that it selects donors from the observed cases. Donors can be selected in a number of ways, including nearest neighbor matching, propensity score matching, and nearest prediction matching. The number of donor values can be set at a specified number or selected based on the quality of the matches. Variance estimation can be more efficient if many donors are used for each missing value. This talk presents studies of methods for using fractional imputation for missing categorical data when predictors are either categorical or continuous. Performance is studied through simulation and examples.

email: larsen@iastate.edu

'SMOOTH' SEMIPARAMETRIC REGRESSION ANALYSIS FOR ARBITRARILY CENSORED TIME-TO-EVENT DATA

Min Zhang*, North Carolina State University
Marie Davidian, North Carolina State University

A general framework for regression analysis of time-to-event data subject to arbitrary patterns of censoring is proposed. The approach is relevant when the analyst is willing to assume that distributions governing model components that are ordinarily left unspecified in popular semiparametric regression models, such as the baseline hazard function in the proportional hazards regression model, have densities satisfying mild 'smoothness' conditions. Densities are approximated by a truncated series expansion that, for fixed degree of truncation,results in a 'parametric' representation, which makes likelihood-based inference coupled with adaptive choice of the degree of truncation, and hence flexibility of the model, computationally and conceptually straightforward under any censoring or tuncation pattern. This formulation allows popular models, such as the proportional hazards, proportional odds, and accelerated failure time models, to be placed in a common framework; provides a principled basis for choosing among them; and renders useful extensions of the models straightforward. The utility and performance of the methods is demonstrated via simulations and by application to data from time-to-event studies.

email: mzhang4@stat.ncsu.edu

ALGORITHMIC PREDICTION OF SUICIDE ATTEMPTS

Steven P. Ellis*, Columbia University
Hanga Galfalvy, Columbia University
Maria Oquendo, Columbia University
John Mann, Columbia University

We use decision theory to analyze the problem of choosing treatments over time in order to prevent suicide attempts by psychiatric patients. Choosing a treatment means choosing the level one is prepared to accept of a numeric treatment cost. Here, 'cost' measures not just monetary costs but all burdens of treatment. Similarly, suicide attempts have costs. On the other hand, a treatment has a benefit which is quantified as the multiplicative factor by which the treatment reduces the probability of an attempt. An optimal treatment decision rule ('dynamic treatment regime') is one that minimizes expected cost. We propose a kernel-based machine learning method for estimating the optimal rule from data. The data consist of predictor data and times to suicide attempt, the latter possibly censored. No special assumptions, such as proportional hazards are made. The algorithm can also be used for estimating the hazard function by maximum likelihood. Our current implementation of the algorithm only makes use of baseline, not time varying, data and only applies to the first attempt after baseline. We demonstrate the algorithm on real suicide attempt data. The method may also find application to learning decision rules for preparing for any kind of unpreventable failure.

email: spe4@columbia.edu

EXACT LIKELIHOOD RATIO TREND TEST USING CORRELATED BINARY DATA

Jonghyeon Kim*, The EMMES Corporation
Neal Oden, The EMMES Corporation
Sungyoung Auh, National Institute of Neurological Disorders and Stroke

Despite its unfortunate sensitivity to the choice of score, the one-sided Cochran-Armitage (CA) test statistic is commonly used to test for a trend in proportions. Both exact and approximate CA test procedures, assuming independent binary data, are already available in most statistical packages. Recently, using quadratic exponential family and sufficiency theory, Corcoran et al. (2001, Biometrics) extended the CA test to handle correlated binary data. Their procedure is available in StatXact®. An alternative method, the likelihood ratio (LR) test statistic for the order-restricted alternative hypothesis, has received a great deal of favorable attention. The LR test statistic does not require uncertain score assignment and has higher power than the one-sided CA test statistic. Both exact and approximate versions of LR test, assuming uncorrelated binary data, are already available. In this talk, parallel to the work of Corcoran et al., we propose an exact conditional LR test procedure for correlated binary data. A small monte carlo study for assessing small sample behavior shows that the LR test statistic outperforms the CA test statistic. We have developed a SAS® macro for these computationally intensive exact (CA and LR) tests for a trend in proportions that is applicable to ophthalmology or otolaryngology data.

email: jhkim@emmes.com

RANGES OF MEASURES OF ASSOCIATIONS FOR FAMILIAL BINARY VARIABLES

Yihao Deng*, Indiana University - Purdue University Fort Wayne

Familial binary data occur in a wide range of scientific investigations. Numerous measures of association have been proposed in the literature for the study of intra-family dependence of the binary variables. These measures include correlations, odd ratios, kappa statistics, and relative risks. In this talk, I will discuss permissible ranges of these measures of association such that a joint distribution exists for the familial binary variables. Joint work with N. Rao Chaganty.

email: dengy@ipfw.edu

36. RECENT ADVANCES IN ASSESSING AGREEMENT

A UNIFIED APPROACH FOR ASSESSING AGREEMENT

Lawrence Lin, Baxter Healthcare
Sam Hedayat, University of Illinois at Chicago
Wenting Wu*, Mayo Clinic

This paper proposes a series of Concordance Correlation Coefficient (CCC) indices to measure the agreement among k raters, with each rater has multiple readings (m) from each of the n subjects for continuous and categorical data. In addition, for normal data, this paper also proposes the coverage probability (CP) and total deviation index (TDI). Those indices are used to measure intra rater, inter rater and total agreement, precision and accuracy. Through a two-way mixed model, all CCC, precision, accuracy, TDI, and CP indices are expressed as functions of variance components, and GEE method is used to obtain the estimates and perform inferences. Most of the previous proposed approaches for assessing agreement become one of the special cases of the proposed approach. For continuous data, the proposed estimates degenerate to the overall CCC (OCCC) independently proposed by several authors. When m=1 and k=2, the proposed estimate degenerates to the original CCC. For categorical data, when k=2 and m=1, the proposed estimate and its inference degenerate to the Kappa for binary data and Weighted Kappa for ordinal data.

email: wu.wenting@mayo.edu

A CLASS OF REPEATED MEASURES CONCORDANCE CORRELATION COEFFICIENTS

Tonya S. King*, Pennsylvania State University College of Medicine
Vernon M. Chinchilli, Pennsylvania State University College of Medicine
Josep L. Carrasco, University of Barcelona

The repeated measures concordance correlation coefficient was proposed for measuring agreement between two raters or two methods of measuring a response in the presence of repeated measurements (King, Chinchilli and Carrasco, 2006, Statistics in Medicine, accepted for publication). This paper proposes a class of repeated measures concordance correlation coefficients that are appropriate for both continuous and categorical data. We illustrate the methodology with examples comparing (1) 1-hour vs. 2-hour blood draws for measuring cortisol in an asthma clinical trial, (2) two measurements of percentage body fat, from skinfold calipers and dual energy x-ray absorptiometry, and (3) two binary measures of quality of health from an asthma clinical trial.

email: tking@psu.edu

A NEW APPROACH TO ASSESSING AGREEMENT BETWEEN QUANTITATIVE MEASUREMENTS USING REPLICATED OBSERVATIONS

Michael Haber*, Emory University
Huiman X. Barnhart, Duke University

We present a new general approach to deriving coefficients of individual agreement between two measurement methods or observers when the observations follow a quantitative scale and the same subject is evaluated more than once by the same method. Our approach compares the disagreement between measurements made by different methods to the disagreement between replicated measurements made by the same method. We consider two situations: (1) comparing two methods that may be used interchangeably, and (2) comparing a new method to an established standard (or reference) method. Disagreement between methods can be quantified via one of several disagreement functions, e.g., the mean squared difference, the mean absolute difference or the mean relative difference. Data on systolic blood pressure will be used to illustrate the new concepts and methods.

email: mhaber@sph.emory.edu

COMPARISON OF CONCORDANCE CORRELATION COEFFICIENT AND COEFFICIENT OF INDIVIDUAL AGREEMENT IN ASSESSING AGREEMENT

Huiman X. Barnhart*, Duke University
Michael Haber, Emory University
Yuliya Lokhnygina, Duke University
Andrzej S. Kosinski, Duke University

In method comparison and reliability studies, it is often important to assess agreement between multiple measurements made by different methods, devices, laboratories, observers, or instruments. For continuous data, the concordance correlation coefficient (CCC) is a popular index in assessing agreement between multiple methods on the same subject where none of the methods is treated as reference. We propose coefficient of individual agreement (CIA) to assess individual agreement between multiple methods for situations with and without a reference method extending the concept of individual bioequivalence from the FDA 2001 guidelines. In this talk, we propose a new CCC for assessing agreement between multiple methods where one of the methods is treated as reference. We compare the properties of the CCC and CIA and their dependency on the relative magnitude of between-subject variability and within-subject variability. We present the relationship between CCC and CIA as well as the impact of between-subject variability. Several examples are presented to explain the interpretation of the CCC and CIA values.

email: huiman.barnhart@duke.edu

A BAYESIAN APPROACH FOR MODELING CENSORED METHOD COMPARISON DATA AND ASSESSING AGREEMENT USING TOLERANCE INTERVALS

Pankaj K. Choudhary*, University of Texas at Dallas
Swati Biswas, University of North Texas Health Science Center

We discuss a mixed effects model type framework for modeling censored method comparison data from two methods. The censoring may be left or right. Left censoring typically arises when there is a lower limit of detection in the measurement method, and right censoring arises when the measurements are times to some event. These data are paired as an individual is measured from both the methods. Our approach is to model the measurements from the methods using marginal distributions of the same form but with different parameters and capture the dependence through a random individual effect. In particular, we study three models in detail: Weibull distribution with gamma random effect, Weibull distribution with lognormal random effect, and lognormal distribution with lognormal random effect. Although we focus on the tolerance interval approach for assessment of agreement, this modeling framework can also accommodate other measures of agreement, such as the concordance correlation. We describe a Bayesian approach based on Markov chain Monte Carlo methods for inference. We also evaluate the frequentist properties of the proposed methodology. Finally, we illustrate its application to a real dataset involving measurements of HIV RNA in plasma samples from two methods with different limits of detection.

email: pankaj@utdallas.edu

COMPARISON OF ICC AND CCC FOR ASSESSING AGREEMENT FOR DATA WITHOUT AND WITH REPLICATIONS

Chia-Cheng Chen*, North Carolina State University
Huiman X. Barnhart, Duke University

The intraclass correlation coefficient (ICC) has been traditionally used for assessing reliability between multiple observers for data with or without replications. Definitions of different versions of ICCs depend on the assumptions of specific ANOVA models. The parameter estimator for the ICC is usually based on the method of moment with the underlying assumed ANOVA model. This estimator is unbiased only if the ANOVA model assumptions hold. Often times these ANOVA assumptions are not met in practice and researchers may compute these estimates without verifying the assumptions. We investigate the impact of the ANOVA assumptions by computing the expected value of the ICC estimator under a very general model to get a sense of the population parameter that the ICC estimator provides. We compare this expected value to the popular agreement index, concordance correlation coefficient (CCC), which is defined without ANOVA assumptions. The main findings are reported for data without replications and with replications and for three ICCs defined by one way ANOVA model, two way ANOVA model without interaction and two way ANOVA model with interaction.

email: cchen4@ncsu.edu

37. GENERAL METHODS II

MODELLING AND INFERENCE FOR AN ORDINAL MEASURE OF STOCHASTIC SUPERIORITY

Euijung Ryu*, University of Florida

An ordinal measure of stochastic superiority can be used to describe the difference between two ordered categorical distributions. This measure summarizes the probability that an outcome from one distribution falls above an outcome from the other, adjusted for ties. Here we develop and compare confidence interval methods for the measure including Wald confidence intervals, a likelihood ratio test-based confidence interval, a score confidence interval, and a pseudo score-type confidence interval. Simulation studies show that with independent multinomial samples the score and the pseudo score-type methods perform well. The score method also works well for fully-ranked data and for matched-pairs data. Finally, we consider a logit model for the measure with explanatory variables.

email: eryu@stat.ufl.edu

RANDOM CLUSTER SIZE, WITHIN-CLUSTER RESAMPLING AND GENERALIZED ESTIMATING EQUATIONS

Eugenio Andraca-Carrera*, University of North Carolina at Chapel Hill
Bahjat Qaqish, University of North Carolina at Chapel Hill
John Preisser, University of North Carolina at Chapel Hill

In many studies of clustered binary data, it is reasonable to consider models in which both response probability and cluster size are related to unobserved random effects. Estimation procedures based on resampling units and pairs within clusers have been developed for such models. We investigate such procedures with an emphasis on computing and interpreting the induced parameters they strive to estimate. We also investigate \gee\ in a model with random cluster size and contrast its induced parameter to that of unit-resampling.

email: eandraca@bios.unc.edu

A METHOD FOR THE SIGNIFICANCE ANALYSIS OF A TREATMENT EFFECT AMONG A GROUP OF GENES AND ITS APPLICATION

Taewon Lee*, National Center for Toxicological Research
Robert R. Delongchamp, National Center for Toxicological Research
Varsha G. Desai, National Center for Toxicological Research
Cruz Velasco, Louisian State University-Health Science Center

In studies that use DNA arrays to assess changes in gene expression, the goal is to evaluate the statistical significance of treatments on expressions for predefined sets of genes; e.g., sets of genes grouped by gene ontology terms. The presentation proposes statistical tests which are based on meta-analysis methods for combining p-values. Most of meta-analysis methods assume independence between p-values. This presentation explores corrections for dealing with problems caused by correlation between genes. Computer simulations are used to demonstrate that the corrections diminish the overstatement of the significance from the usual meta-analysis methods. The methods are applied to the analysis of gene groups related to mitochondrial respiratory chain. The treatment of nucleoside reverse transcriptase inhibitors are expected to affect gene expressions in these gene group and this is shown in the analysis.

e-mail: taewon.lee@fda.hhs.gov

CORRELATION ESTIMATION FROM INCOMPLETE BIVARIATE SAMPLES

Qinying He*, The Ohio State University
H. N. Nagaraja, The Ohio State University

We assume (X,Y) has either a bivariate normal or Downton's bivariate exponential distribution with unknown correlation coefficient ρ. Suppose our data consists of either only the Y values and the ranks of associated X values or a Type II right censored bivariate sample from (X,Y). Our goal is to investigate the estimators of ρ based on these two types of data. For both distributions we use simulation to examine several estimators and obtain their asymptotic relative efficiencies.

e-mail: he@stat.ohio-state.edu

INVERSE PREDICTION: A CLINICAL APPLICATION

Jay Mandrekar*, Mayo Clinic

An important feature of regression methodology is in the area of prediction. Oftentimes investigators are interested in predicting a value of a response variable (Y) based on the known value of the predictor variable (X). However, sometimes there is a need to predict a value of the predictor variable (X) based on the known value of the response variable (Y). In such situations, it is improper to simply switch the roles of the response and predictor variables to get the desired predictions i.e., regress X on Y. This is because the primary assumption that X is measured without error and Y is a dependent, random and normally distributed variable is violated. A method that accounts for the underlying assumptions while estimating or predicting X from known Y is known as inverse prediction. This approach will be illustrated using clinical data, including calculations for the 95% confidence limits for a predicted X from a known Y.

e-mail: mandrekar.jay@mayo.edu

LATENT TRAIT MODELS FOR DEVELOPMENT OF A SUMMARY INDEX IN INFANT HEALTH CARE

Xuefeng Liu*, Wayne State University

Birth defect, abnormal condition, developmental delay or disability and low birth weight are four major pregnancy outcomes which are associated with infant morbidity. Most studies have focused on assessment of the effects of risk factors on each of these outcome variables or of the relationship among these outcomes or both. Little attention has been paid to development of a composite index which is a summary construct of infant morbidity outcomes. In this paper, we develop extended latent trait models and modified Gauss-Newton algorithms for multiple multinomial morbidity outcomes with complete responses. Instead of modelling the marginal distribution of the latent variable, we model conditinal probabilities of each outcome as a function of the latent variable. Estimated generalized nonlinear least square method is used to solve equations for parameters of interest. The models are applied to an infant morbidity data set. A new single variable, called infant morbidity index which is a summary of these four morbidity outcomes and represents propensity for infant morbidity, is developed. The validity of this index is assessed in detail. It is shown that this index is correlated with each of the individual outcome, with infant mortality and with a face-valid index of morbidity outcomes, and can be used in future research as a measure of infant propensity for morbidity.

e-mail: xli@med.wayne.edu

META-ANALYSIS OF TRENDS IN DIETARY STUDIES

Michael P. LaValley*, Boston University School of Public Health

Evaluation of trends in an outcome according to the levels of a risk factor is an area where meta-analysis has been applied. As an example, Gao et al. (JNCI, 2005) evaluated the level of dairy intake as a risk factor for prostate cancer in men using the summary results of eight nutritional epidemiology studies. Such meta-analyses are complicated by the use of correlated within-study summary data, usually in the form of multiple odds ratios comparing groups of subjects with increasing risk factor levels to the subjects in single reference group, and by the need to accommodate trends that may not be linear in form. I evaluate the use of penalized spline regression, fit within standard mixed modeling software, for meta-analytic modeling of such correlated data and nonlinear trends. Simulated data and the data from the Gao et al. study will be used to evaluate the results.

e-mail: mlava@bu.edu

38. CHALLENGES AND SOLUTIONS IN THE ANALYSIS OF OBSERVATIONAL DATA

ASSESSING THE SAFETY OF RECOMBINANT HUMAN ERYTHROPOIETIN USING INSTRUMENTAL VARIABLE METHODS

M. Alan Brookhart*, Brigham and Women's Hospital and Harvard Medical School

Recombinant human erythropoietin (EPO) is the primary treatment for anemia associated with chronic kidney disease (CKD). RCTs have shown that treatment with EPO can improve quality of life in CKD patients; however, a recent RCT found that CKD patients who were treated to higher target hemoglobin levels had an increased risk of the composite end point of death, myocardial infarction, hospitalization for congestive heart failure, and stroke. Observational studies are underway to better characterize the potential risks of EPO, but such studies are challenging due to the medical complexity of CKD patients and the limitations of the available data. We describe a study design for this problem based on instrumental variables that are defined at the level of the dialysis facility. The strengths and limitations of this approach are discussed and some preliminary results are reported.

email: abrookhart@rics.bwh.harvard.edu

AN APPLICATION OF MARGINAL STRUCTURAL MODELS IN THE ANALYSIS OF ANEMIA MANAGEMENT IN PATIENTS ON HEMODIALYSIS

David T. Gilbertson*, Minneapolis Medical Research Foundation
Eric Weinhandl, Minneapolis Medical Research Foundation

In a typical marginal structural model (MSM), the outcome model includes only time-independent predictors, and then weights each observation to account for accumulated treatment assignment bias from the beginning of follow-up until the time point at which the outcome model is measured. In contrast, a history-adjusted (HA) MSM includes time-independent predictors and time-dependent predictors that are measured a fixed amount of time before measurement of the outcome. Each observation is then weighted to account for accumulated treatment assignment bias from the conclusion of time-dependent factor measurement until the measurement of the outcome. Depending on the question at hand, a standard MSM or a HA-MSM may be more appropriate. In the analysis of anemia management in patients on hemodialysis, HA models may be more appropriate because treatment assignment bias likely does not accumulate over an extended period of time. Treatment decisions are made on recent hemoglobin measurements; patient characteristics and hemoglobin measurements 6 months prior have little impact on current month treatment, given patient characteristics and hemoglobin values from more recent months prior. For this study, we compared a standard MSM with a HA-MSM in the analysis of anemia management in patients with end-stage renal disease receiving hemodialysis.

email: dgilbertson@cdrg.org

NOVEL APPROACHES TO DEALING WITH MISSING CONFOUNDER INFORMATION IN HEMODIALYSIS MANAGEMENT

Marshall M. Joffe*, University of Pennsylvania

In analyses of the effect of erythropoetin (EPO) on mortality depend on the availability of information on time-varying confounders. Some sources of data for subjects on hemodialysis are dependent on billing records. In these data, information on an important confounding variable, hemoglobin level, is collected only when a treatment is provided; otherwise, information is not collected. Further, information on some other potential confounders is not collected. Standard methods require confounder information to be available from all subjects at all times. For the problem of occasionally missing confounding information, I propose an approach based on structural nested model and a modification of G-estimation which allows consistent estimation of causal effects when confounder information is unavailable at times when no treatment is provided. The modification involves modeling dose for subjects who have nonzero doses. For the problem of missing confounder information. I consider analysis under the assumption that future the future hemoglobin levels that would have been seen in the absence of EPO capture the unmeasured confounding. I propose an approach for jointly modeling the effect of EPO on subsequent hemoglobin and outcome using structural nested models that allows control for this confounding.

email: mjoffe@cceb.upenn.edu

39. DESIGN AND ANALYSIS OF BEHAVIORAL INTERVENTION STUDIES

INFERENCE ABOUT CAUSAL CONTRASTS BETWEEN TWO ACTIVE TREATMENTS IN RANDOMIZED TRIALS WITH NONCOMPLIANCE: THE EFFECT OF SUPERVISED EXERCISE TO PROMOTE SMOKING CESSATION

Jason Roy*, University of Rochester
Joseph W. Hogan, Brown University

In behavioral medicine trials, such as smoking cessation trials, two or more active treatments are often compared. Noncompliance by some subjects with their assigned treatment poses a challenge to the data analyst. Even if subjects in one arm do not have access to the other treatment(s), the causal effect of each treatment typically can only be identified within certain bounds. We propose the use of compliance-predictive covariates to help identify the causal effects. Our approach is to specify marginal compliance models conditional on covariates within each arm of the study. Parameters from these models can be identified from the data. We then link the two compliance models through an association model that depends on a parameter that is not identifiable, but has a meaningful interpretation; this parameter forms the basis for a sensitivity analysis. We demonstrate the benefit of utilizing covariate information in both a simulation study and in an analysis of data from a smoking cessation trial.

email: jason_roy@urmc.rochester.edu

POST-RANDOMIZATION INTERACTION ANALYSES WITH RANK PRESERVING MODELS

Thomas R. Ten Have*, University of Pennsylvania School of Medicine
Jennifer Faerber, University of Pennsylvania
Marshall M. Joffe, University of Pennsylvania School of Medicine

In the context of two randomized psychiatry trials of behavioral interventions, we present a linear rank preserving model approach for analyzing the interaction between a randomized baseline intervention and post-randomization factor on a univariate follow-up outcome. Unlike standard interaction analyses, our approach does not assume that the mediating factor is randomly assigned to individuals (a form of sequential ignorability). However, there is a tradeoff with other assumptions. Consistent estimation of causal interaction effects without sequential ignorability employs weights under the G-estimation approach that are optimal in terms of semi-parametric efficiency but under sequential ignorability. In this context, we will present analyses of interactions between behavioral interventions on reducing depression outcomes and medication and therapy factors in two randomized psychiatry trials. Comparisons will be made with principal stratification, which offers an alternative way of assessing such interactions.

email: ttenhave@cceb.med.upenn.edu

THE MULTI-PHASE OPTIMIZATION STRATEGY: A NOVEL WAY TO DEVELOP MULTI-COMPONENT BEHAVIORAL INTERVENTIONS

Bibhas Chakraborty*, University of Michigan
Linda M. Collins, Pennsylvania State University
Susan A. Murphy, University of Michigan
Vijayan N. Nair, University of Michigan
Victor J. Strecher, University of Michigan

The Multi-phase Optimization Strategy (MOST) is an experimental procedure to proactively develop, optimize, and evaluate multi-component behavioral interventions. MOST consists of three ordered phases of experimentation: screening, refining, and confirming; and it utilizes a sophisticated class of experimental designs called fractional factorials, when there are a large number of components. An overview of this methodology will be given in the talk. Moreover, using an extensive simulation study, we will illustrate the superiority of the new procedure over the traditional approach of formulating a multi-component intervention and immediately proceeding to a confirmatory two-arm randomized trial.

email: bibhas@umich.edu

HETEROGENEOUS BETWEEN- AND WITHIN-SUBJECTS VARIANCE MODELS FOR LONGITUDINAL BEHAVIORAL DATA

Donald Hedeker*, University of Illinois at Chicago
Robin Mermelstein, University of Illinois at Chicago

Mixed models are commonly used for analysis of longitudinal behavioral data. Typically, interest centers around describing and statistically comparing time-trends across groups of subjects. These comparisons focus on statistical tests of the fixed effects of the model. The error variance and the variance parameters of the random effects are usually considered to be homogeneous across subject groups. These variance terms characterize the within-subjects (error variance) and between-subjects (random-effects variance) variation in the longitudinal data. In this presentation, we allow both the between- and within-subject variances to be modeled in terms of explanatory variables according to a log-linear structure. Specifically, in addition to including variables as fixed effects, explanatory variables can enter the model to influence either variance in a multiplicative manner. We describe an application from a longitudinal behavioral study of adolescent smoking to illustrate the usefulness of this class of models, both for continuous and ordinal outcomes.

email: hedeker@uic.edu

40. INTEGROMICS

GENOMIC INTEGRATION OF COPY NUMBER AND EXPRESSION DATA

Debashis Ghosh*, University of Michigan

Recently, there has been a plethora of cancer studies in which samples have been profiled using both gene expression and copy number microarrays. While statistical methods for the analysis of gene expression microarrays have become quite abundant, those for copy number microarrays, and in particular their integration, are less well-developed. We give a brief introduction to copy number and describe an empirical study of copy number/gene expression correlation in several publicly available cancer cell line studies. We then describe an Empirical Bayes methodology for assessing significance of correlations in these types of studies that represents an extension of false-discovery rate procedures that has recently been popular. We next describe approaches for joint modelling of copy number effects on expression using adaptations of machine learning and data mining procedures. Application of the techniques to real datasets will be presented.

email: ghoshd@umich.edu

A HIDDEN MARKOV MODEL FOR COMBINING ESTIMATES OF COPY NUMBER AND GENOTYPE CALLS IN HIGH THROUGHPUT SNP ARRAYS

Robert B. Scharpf*, Johns Hopkins University
Giovanni Parmigiani, Johns Hopkins University
Jonathan Pevsner, Johns Hopkins University
Ingo Ruczinski, Johns Hopkins University

High density single nucleotide polymorphism microarrays (SNP chips) provide information on a subject's genome, such as copy number and genotype (heterozygosity/homozygosity) at a SNP. Copy number estimates are useful for classifying homozygous deletions (zero copies), hemizygous deletions (one copy), gene duplications (greater than two copies), and mosaicism (non-integer copies). Gene function may also be modulated by mechanisms that do not affect copy number (copy neutral). For instance, uniparental isodisomy (UPD) occurs when a subject inherits 2 copies of a chromosome or chromosomal segment from 1 parent. Regions of LOH, of which UPD is a special case, are detectable in high throughput SNP assays by a reduction in the proportion of SNPs called heterozygotes. Joint estimates of copy number and genotype calls can discriminate hemizygous deletion LOH from copy neutral LOH. Using estimates of copy number and genotype as a starting point, we combine these two sources of information in a hidden Markov model that identifies regions of probable deletion, LOH, and amplification.

email: rscharpf@jhsph.edu

STATISTICAL METHODS FOR RECONSTRUCTING TRANSCRIPTIONAL REGULATORY NETWORKS

Ning Sun, Yale University School of Medicine
Raymond J. Carroll, Texas A&M University
Hongyu Zhao*, Yale University School of Medicine

Transcription regulation is a fundamental biological process, and extensive efforts have been made to dissect its mechanisms through direct biological experiments and regulation modeling based on physical–chemical principles and mathematical formulations. Despite these efforts, transcription regulation is yet not well understood because of its complexity and limitations in biological experiments. Recent advances in high throughput technologies have provided substantial amounts and diverse types of genomic data that reveal valuable information on transcription regulation, including DNA sequence data, protein–DNA binding data, microarray gene expression data, chromatid structure data, and others. In this presentation, we discuss a Bayesian error analysis model to integrate protein–DNA binding data, gene expression data, and other data types to reconstruct transcriptional regulatory networks. The usefulness of this approach is demonstrated through its application to infer transcriptional regulatory networks in the yeast cell cycle.

email: hongyu.zhao@yale.edu

41. INTERVAL CENSORING IN STUDIES OF INFECTIOUS DISEASE

NONPARAMETRIC ESTIMATION OF THE JOINT DISTRIBUTION OF A SURVIVAL TIME SUBJECT TO INTERVAL CENSORING AND A CONTINUOUS MARK VARIABLE

Michael G. Hudgens*, University of North Carolina-Chapel Hill
Marloes H. Maathuis, University of Washington
Peter B. Gilbert, Fred Hutchinson Cancer Research Center and University of Washington

We consider three nonparametric estimators of the joint distribution function for a survival time subject to interval censoring and a continuous mark variable. Finite and large sample properties are described for the nonparametric maximum likelihood estimator (NPMLE) as well as estimators based on midpoint imputation (MIDMLE) and coarsening the mark variable (CRMLE). The estimators are compared using data from a simulation study and a recent Phase III HIV vaccine efficacy trial where the survival time is the time from enrollment to infection and the mark variable is the genetic distance from the infecting HIV sequence to the HIV sequence in the vaccine. Theoretical and empirical evidence are presented indicating the NPMLE and MIDMLE are inconsistent. Conversely, the CRMLE is shown to be consistent in general and thus is preferred.

email: mhudgens@bios.unc.edu

RELATING A HEALTH OUTCOME TO SUBJECT-SPECIFIC CHARACTERISTICS BASED ON LEFT- OR INTERVAL-CENSORED LONGITUDINAL EXPOSURE DATA

Robert H. Lyles*, Emory University
Kathleen A. Wannemeuhler, Emory University
Amita K. Manatunga, Emory University
Renee H. Moore, University of Pennsylvania
Michele Marcus, Emory University

Random effects in models for longitudinal data often provide meaningful subject-specific measures of exposure that might be associated with health-related outcomes. This motivates two-stage or, ideally, unified models to tie together the outcome and exposure information. In the two-stage approach, interest lies in the properties of predictors of random effects and their relative performances as covariates at the second stage. While more challenging, a unified modeling approach arguably provides an inherent and efficient adjustment for covariate measurement error. Either approach can face complications, however, when the exposure data are subject to detection limits, are coarse due to rounding, or are otherwise interval-censored. We consider an application in environmental and reproductive health that motivates likelihood-based approaches to handling censored exposure data and linking them to outcomes. We assess the use of empirical Bayes and empirical constrained Bayes predictions at the second stage, and compare the resulting estimated parameters of interest from the health effects model with those obtained under a joint modeling approach.

email: rlyles@sph.emory.edu

TARGETED MAXIMUM LIKELIHOOD ESTIMATION: APPLICATION TO INTERVAL CENSORED DATA

Mark van der Laan*, University of California-Berkeley
Dan Rubin, University of California-Berkeley

Maximum likelihood (ML) estimation involves estimation of the density of the data according to a model. In high dimensional models ML estimation needs to involve model selection trading off bias due to model mis-specification and variance. This trade-off involves the bias and variance of the density of the data as a whole, and typically represents a wrong trade off for specific parameters representing a scientific question of interest: to illustrate this point, the bandwidth selected by likelihood based criteria (e.g., likelihood based cross-validation) for a kernel density estimator of a univariate density results in an estimator of the cdf which does not even achieve the standard root-n rate of convergence, and this wrong trade-off for the purpose of a particular smooth parameter deteriorates with increasing dimension. The cause of this problem is that ML estimation is not targeted towards the scientific parameter of interest, and often results in non-robust estimators and invalid tests of a null hypothesis about the parameter of interest. We present a new method ''Targeted ML Estimation'' unifying ML estimation and estimating function based methodology and providing important improvements. We illustrate it in the estimation of an effect of a variable on an outcome, adjusting for a user supplied set of variables/confounders, based on interval and right censored data.

email: laan@stat.berkeley.edu

42. NONPARAMETRIC BAYES CLUSTERING FOR COMPLEX BIOLOGICAL DATA

ENRICHED STICK BREAKING PROCESSES FOR FUNCTIONAL DATA

David B. Dunson*, NIEHS/NIH
Bruno Scarpa, University of Padua,-Italy

In many applications involving functional data, prior information is available about the proportion of curves having different attributes. It is not straightforward to include such information in existing procedures for functional data analysis. Generalizing the functional Dirichlet process (FDP), we propose a class of stick-breaking priors for distributions of functions. These priors incorporate functional atoms drawn from a Gaussian process. The stick-breaking weights are specified to allow user-specified prior probabilities for curve attributes, with hyperpriors accommodating uncertainty. Compared with the FDP, the random distribution is enriched for curves having attributes known to be common. Theoretical properties are considered, methods are developed for posterior computation, and the approach is illustrated using data on temperature curves in menstrual cycles.

email: dunson1@niehs.nih.gov

NONPARAMETRIC BAYESIAN METHODS FOR GENOMIC DATA

Marina Vannucci*, Texas A&M University

Variable selection has been the focus of much research in recent years. This talk will focus on the the development of Bayesian methods for variable selection in problems that aim at clustering the samples. A novel methodology will be described that uses infinite mixture models via Dirichlet process mixtures to define the cluster structure. A latent binary vector identifies the discriminating variables and is updated via a Metropolis algorithm. Inference on the cluster structure is btained via a split-merge MCMC technique. Performances of the methodology are illustrated on simulated data and on DNA microarray data.

email: mvannucci@stat.tamu.edu

A HIDDEN MARKOV DIRICHLET PROCESS MODEL FOR JOINT INFERENCE OF POPULATION STRUCTURE, LINKAGE DISEQUILIBRIUM, AND RECOMBINATION HOTSPOTS

Eric Xing*, Carnegie Mellon University
Kyung-Ah Sohn, Carnegie Mellon University

The problem of inferring the population structure, linkage disequilibrium pattern, and chromosomal recombination hotspots from genetic polymorphism data is essential for understanding the origin and characteristics of genome variations, with important applications to the genetic analysis of disease propensities and other complex traits. Statistical genetic methodologies developed so far mostly address these problems separately using specialized models ranging from coalescence and admixture models for population structures, to hidden Markov models and renewal processes for recombination; but most of these approaches ignore the inherent uncertainty in the genetic complexity (e,g., the number of genetic founders of a population) of the data and the close statistical and biological relationships among objects studied in these problems. We present a new statistical framework called hidden Markov Dirichlet process (HMDP) to jointly model the genetic recombinations among possibly infinite number of founders and the coalescence-with-mutation events in the resulting genealogies. The HMDP posits that a haplotype of genetic markers is generated by a sequence of recombination events that select an ancestor for each locus from an unbounded set of founders according to a 1st-order Markov transition process. Conjoining this process with a mutation model, our method accommodates both between-lineage recombination and within-lineage sequence variations, and leads to a compact and natural interpretation of the population structure and inheritance process underlying haplotype data. We have developed an efficient sampling algorithm for HMDP based on a two-level nested P\'{o}lya urn scheme, and we present experimental results on joint inference of population structure, linkage disequilibrium, and recombination hotspots based on HMDP. On both simulated and real SNP haplotype data, our method performs competitively or significantly better than extant methods in uncovering the recombination hotspots along chromosomal loci; and in addition it also infers the ancestral genetic patterns and offers a highly accurate map of ancestral compositions of modern populations.

email: epxing@cs.cmu.edu

43. STATISTICAL DATA MINING FOR ADVERSE DRUG EVENT SURVEILLANCE

DATA MINING THE WHO DATABASE OF SUSPECTED ADVERSE DRUG REACTIONS

Andrew Bate*, WHO Collaborating Centre for International Drug Monitoring

Spontaneous reporting of suspected adverse drug reactions (ADRs) is well-established as the best overall method for detection of previously unidentified ADRs of drugs after market approval. The nearly four million suspected ADR reports in the WHO database is the largest of its kind, containing spontaneous reports received from the national centres of 80 countries. The number of reports makes initial quantitative initial filtering of the data for clinical review essential. Data mining has been used routinely since 1998. Many associations have been first found by prospective quantitative screening of WHO data, including SSRIs and neonatal withdrawal syndrome. The method core is the search for disproportional reporting of drug- ADR pairs relative to an expected count estimated from other reporting in the data. Quantitative methods enhance rather than replace traditional ADR surveillance practices. Duplicate copes of the same ADR incident severely hamper data analysis of spontaneous reports. Injudicious stratification can lead to false negatives; while stratum specific disproportionality often highlights data quality issues. More sophisticated methods including regression approaches seem appealing, but will fail to provide useful results unless their implementation allows for the non-random nature of submission of reports, missing data, duplicate records, and the classification of data on reports.

email: andrew.bate@who-umc.org

VACCINE ADVERSE EVENT SURVEILLANCE

Robert L. Davis*, CDC

The Vaccine Safety Datalink Project (VSD) is a collaborative network between 8 health maintenance organizations (HMO) and the Centers for > Disease Control to perform vaccine safety surveillance and research safety. The VSD has begun to utilize a rapid cycle analysis (RCA)to minimize the lag time between the occurrence and identification on of vaccine adverse events (VAE). Using routinely collected HMO administrative data, cohorts of vaccinated persons are created weekly and followed for up to 6 weeks to identify potential VAEs. Rates of events among vaccinated persons are then compared to rates among concurrent non-vaccinated persons or historical controls. Sequential probability ratio testing (SPRT) is carried out on these large groups of weekly cohorts to identify and/or test both pre-specified and unexpected signals of vaccine adverse events. These analyses take advantage of the unique informatics capabilities and data infrastructure of the VSD, and the RCA methodology allows identification of vaccine adverse events in near 'real-time.' We will present our experience in monitoring of new vaccines, including conjugate meningococcal vaccine, rotavirus vaccine, and the yearly influenza vaccine.

email: rad2@cdc.gov

DATA MINING FOR TREATMENT-COMPARATOR DIFFERENCES IN ADVERSE EVENT RATES WITHIN COLLECTIONS OF CLINICAL TRIAL RESULTS

William DuMouchel*, Lincoln Technologies Division of Phase Forward, Inc.
Robert Zambarano, Lincoln Technologies Division of Phase Forward, Inc.

Analysis methods for adverse event frequencies in clinical trials are less developed than are analyses for efficacy. Accounting for multiple comparisons due to the many types of adverse events under observation is problematical, as is the question of how to group event types that seem medically related. Empirical Bayesian approaches to two problems will be developed and illustrated with examples. First, a clustering method for finding potential syndromes related to treatment is based on finding sets of events for which all pairs within the set have enhanced presence within the treatment group compared to the control group. Second, a hierarchical model for parallel logistic regressions allows analyses of medically related events to 'borrow strength' from each other and also permits subgroup analyses that are designed to be resistant to the multiple comparisons fallacy.

email: william.dumouchel@lincolntechnologies.com

BAYESIAN LOGISTIC REGRESSION FOR DRUG SAFETY SURVEILLANCE

David Madigan*, Rutgers University
Aimin Feng, Novartis
Ivan Zorych, New Jersey Institute of Technology

Spontaneous report databases represent the primary data source for monitoring the safety of licensed drugs. This talk will discuss current widely used data mining algorithms for monitoring such databases. I will also discuss some limitations of these algorithms and present some potential alternatives. In particular I will describe the results of a project applying large-scale Bayesian logistic regression in this context.

email: dmadigan@rutgers.edu

44. ADAPTIVE DESIGN

ESTIMATION FOLLOWING AN ADAPTIVE GROUP SEQUENTIAL TEST

Cyrus R. Mehta*, Cytel Inc.

This paper proposes two methods for computing confidence intervals with exact or conservative coverage following a group sequential test in which an adaptive design change is made one or more times over the course of the trial. The key idea, due to Muller and Schafer (2001), is that by preserving the null conditional rejection probability of the remainder of the trial at the time of each adaptive change, the overall type 1 error, taken unconditionally over all possible design modifications, is also preserved. This idea is further extended by considering the dual tests of repeated confidence intervals (Jennison and Turnbull, 1989) and of stage-wise adjusted confidence intervals (Tsiatis, Rosner and Mehta, 1984). The method extends to the computation of median unbiased point estimates.

email: mehta@cytel.com

USING WEIGHTED ESTIMATES TO INCREASE THE POWER OF A SEAMLESS PHASE II/III TRIAL

Kenneth Klesczewski*, Bristol-Myers Squib

The FDA's critical path initiative, started over two years ago, is aimed at catalyzing the creation of tools to advance the science of drug development. One tool, the FDA has stated, is the use of adaptive designs in clinical trials and the FDA is taking steps to help facilitate the continued development of more adaptive clinical trials. One type of adaptive trial is the seamless Phase II/III clinical trial. This type of trial has two parts, one part performs the function of the learning Phase II trial and the second part performs the function of the confirmatory Phase III trial. This design can create efficiencies over performing separate Phase II and Phase III trials. These efficiencies include reducing the "white space" between trials, the cost of restarting sites, and the overhead in initiating a new trial. In addition, the power of the confirmatory part of the trial can be increased by including the data from the first part. In the seamless trial design discussed, part one has multiple active arms and a comparator while the second part has one active arm along with the comparator. Several approaches to analyze the part one data will be reviewed. A weighted method, with pre-specified weights for the treatment arms, will be applied that would provide extra power when it is assumed that certain treatment arms will perform differently. A method for weight selection will also be discussed.

email: kenneth.klesczewski@bms.com

A COMPUTATIONALLY INTENSIVE APPROACH TO OPTIMIZING BAYESIAN GROUP SEQUENTIAL CLINICAL TRIALS

J. Kyle Wathen*, University of Texas: M.D. Anderson Cancer Center
Peter F. Thall, University of Texas: M.D. Anderson Cancer Center

In this talk we present a computationally intensive approach to the problem of deriving an optimal Bayesian design for a randomized group sequential clinical based on right-censored event times. We are motivated by the fact that, if the proportional hazards assumption is not met, then the actual power of a standard group sequential design can differ substantially from its nominal power figure. Using parallel processing, we combine the techniques of Bayesian decision theory, adaptive interim model selection and forward simulation to obtain a group sequential procedure that maintains targeted size and power under a wide range of true event time distributions. A simulation study comparing this design to three commonly used group sequential designs shows that, over many different event time distributions, our proposed method performs at least as well as each of the frequentist designs, and in many cases it provides a much smaller trial.

email: jkwathen@mdanderson.org

SELF-DESIGNING MULTIPLE DOSE STUDY VIA ADAPTIVE RANDOMIZATION

Lin Wang*, Sanofi-Aventis
Lu Cui, Sanofi-Aventis

In clinical trials, in addition to the control, multiple doses of a testing new drug are often used for various purposes, including the determination of a dose response curve or of a potentially more effective dose for further investigation. In this study, a new adaptive randomization method is proposed to improve the efficiency of a multiple dose trial. The new method allows unequal patient allocations, as driven by the data, to different treatment arms according to the trial objectives. The descriptions of the new method will be given together with the assessment of its performance in terms of the power, the type I error rate, and the rate of successful patient allocations. Its implementation, requiring no Bayesian calculations, as well as its applications to Phase II dose response studies and Phase III studies with dropping ineffective dose arms will also be discussed.

email: lin.wang2@sanofi-aventis.com

DOUBLY ADAPTIVE BIASED COIN DESIGN WITH HETEROGENEOUS RESPONSES

Liangliang Duan*, University of Virginia
Feifang Hu, University of Virginia

Heterogeneous responses are present in many sequential experiments. Double adaptive biased coin design is an important family of response adaptive designs, which use accruing data and sequentially updated estimation to skew the allocation probability to favor the treatment performing better thus far. Heterogeneous responses in doubly adaptive biased coin design can bias results of the study. Here we propose a general weighted likelihood method to deal with heterogeneous outcomes and provide a modification of such designs. Strong consistency and asymptotic normality of the new design are obtained under some widely satisfied conditions. Some advantages of the proposed method are discussed. Numerical studies are also presented.

email: ld6g@virginia.edu

A LIKELIHOOD APPROACH FOR TREATMENT SCHEDULE FINDING USING A MIXTURE CURE MODEL WITH A SECTIONAL WEIBULL HAZARD

Changying A. Liu*, University of Michigan
Tom Braun, University of Michigan

Conventional Phase I clinical trials are designed to determine a maximum tolerated dose (MTD) of a therapeutic agent. However, conventional dose-finding Phase I studies are inadequate for trials in which the agent is administered repeatedly over time and evaluation of long-term cumulative effects is important. Braun, Yuan and Thall (2005) constructed a new paradigm for Phase I trials that allows for the evaluation and comparison of several treatment schedules, each consisting of a sequence of administration times. The goal of this design is to determine the maximum tolerated schedule (MTS) instead of a traditional MTD. As an alterative to the triangular hazard model and the Bayesian approach used by Braun, Yuan, and Thall (2005), we propose a cure model with a sectional Weibull distribution to evaluate a fixed number of nested treatment schedules to determine the MTS, in which the event rate modeled by a logistic regression and the conditional hazard function for the susceptible modeled by a mixture of two Weibull distributions to account for the non-monotonic nature of the hazard of toxicity. We use a likelihood approach to estimate parameters of interest. Subject accrual, and outcome adaptive decision-making are done in a sequential fashion as in classical Phase I trials.

email: changying8618@yahoo.com

EXTRACTING INFORMATION FROM AN ON-GOING BLINDED TRIAL

Jitendra Ganju*, Biostatistical Consulting, BiostatWorks
Biao Xing, Genentech

A method is described for re-estimating the variance - and hence re-estimating the sample size - of a continuous endpoint from a blinded trial. The method makes minimal assumption and is shown to work well. The same method can be used to assess the impact of missing data and non-compliance on the variability of the endpoint. It is concluded that useful information can be gleaned from a blinded database.

email: jganju@biostatworks.com

45. PANEL ON THE NEW FDA GUIDANCE FOR THE USE OF BAYESIAN STATISTICS IN MEDICAL DEVICE CLINICAL TRIALS

PANEL DISCUSSION

Gregory Campbell, Food and Drug Administration
Gene Pennello, Food and Drug Administration
Telba Irony, Food and Drug Administration
Gerry Gray, Food and Drug Administration

46. CLASSIFICATION WITH HIGH DIMENSIONAL DATA: GENOMICS, PROTEOMICS, AND METABOLOMICS

TOWARDS BETTER VISUALIZATION, IDENTIFICATION, AND CLASSIFICATION IN METABOLOMICS

Seoung Bum Kim*, University of Texas-Arlington
Dean P. Jones, Emory University

Metabolomics with high-resolution nuclear magnetic resonance (NMR) spectroscopy is emerging as an efficient approach for investigation of metabolic changes within biological systems. Principal component analysis (PCA) has been widely used to facilitate visualization and identify the important features that account for most of the variances in high-resolution NMR spectra. However, various sources of variation in NMR spectra lead to unsatisfactory PCA results for visualization. Moreover, extracting meaningful metabolite features from the reduced dimensions obtained through PCA is complicated because these reduced dimensions are linear combinations of a large number of the original features. This study presents an orthogonal signal correction (OSC) and, a false discovery rate (FDR)-based multiple testing procedure to enhance visualization, provide better feature selection, and improve construction of prediction models from high-resolution NMR spectra. The results showed that OSC-processed spectra achieved a better visualization capability than without OSC. For feature selection, the FDR-based multiple testing procedure identified a set of metabolite features that play an important role in discriminating between different conditions among the samples without losing any information in the original metabolite features.

email: sbkim@uta.edu

LOCALLY ADAPTIVE DISCRIMINANT ANALYSIS OF PROTEOMICS DATA

Yuping Wu*, Cleveland State University
R. Webster West, Texas A&M University

The use of mass spectrometry (MS) as a means of analyzing the proteome for biomarker discovery and cancer diagnostics has been evaluated extensively in recent years due to the new advances in MS technology. Most recent statistical methods for analyzing MS data of proteins summarize the data into a rectangular matrix whose values are intensities associated with peaks. Summarizing data in this manner is somewhat artificial as the number of peaks and the locations of these peaks vary significantly across spectra. In its purest form, processed MS data typically consist of an array of location/intensity pairs for each spectrum with the number of elements in each array varying greatly across spectra. We propose a novel method, locally adaptive discriminant analysis (LADA), for sample classification from MS data in this more primitive form. LADA predicts class labels for each individual peaks in a new spectrum and combines these class labels to obtain the class label for the new spectrum. In comparison to the available rectangular techniques, LADA is more robust to spurious and non-informative peaks. Moreover, LADA takes into account both location and intensity differences in peaks across spectra. LADA can also identify biomarkers more effectively by selecting the set of peaks that best discriminates between different experimental groups.

email: y.wu88@csuohio.edu

QUANTIFICATION AND CLASSIFICATION OF MASS SPECTROMETRY DATA BY KERNEL METHODS

Shuo Chen*, Vanderbilt Ingram Cancer Center
Yu Shyr, Vanderbilt Ingram Cancer Center

Mass spectrometry (MS) technology has been widely applied for biomedical research. The quantification and normalization of MS data is a challenging and crucial task, based on which analysis is performed such as: quality control, biomarker selection, and classification. When measuring the peaks of interest detected by preprocessing, we may introduce variation or lose useful peak shape information by using the local maximum or defining an area under the curve. Also, there have been no "the" methods for baseline correction and normalization of the raw MS data, which affect the peak quantification greatly. In this talk we propose a biomarker selection and classification process based kernel methods. It is more robust and applies more information from the spectra. The results of biomarker selection and classification for real datasets based on SVM with the developed kernels are satisfactory.

email: shuo.chen@vanderbilt.edu

IDENTIFICATION OF MAJOR METABOLITE FEATURES IN HIGH-RESOLUTION NMR

Yajun Mei*, Georgia Institute of Technology
Seoung Bum Kim, University of Texas-Arlington
Kwok Leung Tsui, Georgia Institute of Technology

In this article, we capitalize on the longitudinal nature of metabolic data from high-resolution nuclear magnetic resonance (NMR) spectroscopy, and combine longitudinal data analysis with multiple testing procedures. Our proposed approach uses linear mixed-effects (LME) models to generate p-values for metabolite features and then identifies major metabolite features using multiple testing procedures, especially the false discovery rate(FDR) procedure by Benjamini and Hochberg (1995). For the purpose of comparison, both our proposed approach and the conventional approach based on two-sample t-statistics were applied to a real data set, and the subsets of selected significant features from these two approaches were then evaluated by a supervised classification techniques based on k-nearest neighbors (kNNs). Our results showed that for metabolic NMR data, our proposed LME approach is more consistent with our intuition than the conventional two-sample t-statistics.

email: ymei@isye.gatech.edu

DETECTION OF BIOMARKER FOR FETAL ALCOHOL SYNDROME (FAS) USING HIGH THROUGHPUT (MALDI-TOF) PROTEOMICS DATA

Susmita Datta*, University of Louisville

FAS is characterized by abnormal facial features, mental and physical growth deficiencies, and problems with central nervous system (CNS). People with FAS might also have multiple difficulties with learning, memory, attention span, communication, vision, hearing etc. Detecting FAS as early as possible has major potential benefits. We use peptide profiles of mouse amniotic fluid generated by matrix assisted laser desorption ionization - time of flight (MALDI-TOF) mass spectrometer. We use novel statistical techniques to normalize and denoise the mass spectra and use univariate and multivariate statistical procedures to detect the differentiating signals between the polypeptide profile of fetal amniotic fluids of control and alcohol exposed mice. We use MS/MS techniques and database searches to identify the associated protein.

email: susmita.datta@louisville.edu

CONFIDENCE INTERVALS FOR CROSS-VALIDATED PREDICTION ACCURACY ESTIMATES

Kevin K. Dobbin*, National Cancer Institute/NIH

Cross-validation can be used to obtain a nearly unbiased estimate of a classifier’s accuracy. But constructing a confidence interval for the accuracy is complicated by the existence of overdispersion. Overdispersion may be a particular concern in high dimensional settings with variable selection. In order to better understand the effects of overdispersion in this context, we present a method for constructing confidence intervals for cross-validation based estimates of prediction accuracy. We explore the geometry of the solution subspace for a specified accuracy and methods of assessing variation in overdispersion at different points throughout the subspace. Problems with specifying distributions on the solution space are discussed. We consider application to the setting of high dimensional genomic data.

email: dobbinke@mail.nih.gov

A CLASSIFICATION ALGORITHM FOR THE DEVELOPMENT OF GENOMIC SIGNATURES FROM HIGH-DIMENSIONAL DATA

Songjoon Baek*, National Center for Toxicological Research/FDA
Hojin Moon, National Center for Toxicological Research/FDA
Hongshik Ahn, Stony Brook University
Ralph L. Kodell, National Center for Toxicological Research/FDA
James J. Chen, National Center for Toxicological Research/FDA

Personalized medicine is defined by the use of genomic signatures of patients in a target population to assign more effective therapies for diseases. Classification algorithms can be applied to high-dimensional data for better prediction of disease progression and/or response to therapy to help individualize clinical assignment of treatment. We have developed a robust nonparametric classification algorithm for high-dimensional data based on ensembles of classifiers from the optimal number of random partitions of features. The performance of the biomarker is assessed by cross-validation to obtain a valid measure of prediction accuracy. Using published high-dimensional data with clinical/demographic variables, the performance of our classification algorithm is shown to be consistently good when compared to existing algorithms that were considered. The predictive accuracy can be improved by adding some relevant clinical/demographic measurements to the genomic data.

email: SongJoon.Baek@fda.hhs.gov

47. IMAGING

ROBUST FITTING FOR NEURORECEPTOR MAPPING

Chung Chang*, Columbia University
Robert Todd Ogden, Columbia University

Positron emission tomography (PET) is used in studies to estimate the density of a neuroreceptor at each location throughout the brain by measuring the concentration of a radiotracer over time and modeling its kinetics. There are a variety of kinetic models in common usage and these typically rely on nonlinear least squares (LS) algorithms for parameter estimation. However, PET data often contains artifacts (such as uncorrected head motion) and so the assumptions on which the LS methods are based may be violated. Quantile regression (QR) provides a robust alternative to LS methods and has been used successfully in many applications. We consider fitting various kinetic models to PET data using QR and study the relative performance of the methods via simulation. A data adaptive method for choosing between LS and QR is proposed and the performance of this method is also studied.

email: cc2240@columbia.edu

MODELING THE SPATIAL AND TEMPORAL DEPENDENCE IN FMRI DATA: AN APPLICATION TO AN INHIBITORY CONTROL STUDY OF COCAINE ADDICTS

Gordana Derado*, Emory University
F. D. Bowman, Emory University

Functional neuroimaging technology is important for investigating behavior-related neural processing linked to substance abuse disorders and associated treatment interventions. fMRI and PET studies yield large data sets that contain temporal correlations from repeated scans and complex spatial correlations. Bowman (2005) proposed a two-stage model for the estimation and testing of localized activity in which the second stage accounts for spatial dependence between voxels within the same neural processing cluster (defined by a data-driven cluster analysis). We propose an extension of this model: a two-stage spatio-temporal model which additionally accounts for temporal or repeated-measures type dependence between the multiple experimental effects for a subject. The first stage of our model resembles a GLM for each individual's vector of serial responses. At the second stage, we employ a simultaneous autoregressive model to capture spatio-temporal correlations between the multiple effects at a given location and between pairs of functionally related voxels, which also provides information about functional connectivity. We use maximum likelihood methods to estimate parameters from our spatio-temporal model. We apply our method to fMRI data from a cocaine addiction study.

email: gderado@emory.edu

ROBUST INDEPENDENT COMPONENT ANALYSIS FOR FMRI

Ping Bai*, University of North Carolina at Chapel Hill
Haipeng Shen, University of North Carolina at Chapel Hill
Young Truong, University of North Carolina at Chapel Hill

Independent component analysis (ICA) is an effective exploratory tool for analyzing spatio-temporal data. It has been successfully applied in analyzing functional Magnetic Resonance Imaging (fMRI) data, to recover the interested source signals from different parts of the brain. Due to the high sensitivity of MR scanners, outliers are inevitable in acquiring fMRI datasets while they cause misleading effects for the analysis. In the current literature, no particular method exists yet to handle this problem. In this paper, we propose a robust ICA procedure that is less sensitive to outliers in fMRI analyses. Singular value decomposition (SVD) is commonly used prior to ICA for dimension reduction. We first motivate SVD from a low rank matrix approximation perspective. Regularization through basis expansion is then introduced to the corresponding minimization problem to achieve a regularized low rank approximation. Such regularization performs dimension reduction as well as trims the outlier effect. Our method makes use of the particular designs of fMRI experiments, and is shown very effective in reducing outlier effect in a spatio-temporal simulation study. We also compare our method with two existing ICA packages by analyzing a real fMRI dataset, and our method can detect extra underlying components.

email: pbai@email.unc.edu

MODELING PROGRESSION OF CEREBROVASCULAR DISEASE WITH LONGITUDINAL MRI DATA VIA SPATIO-TEMPORAL TRANSITION MODELS

Qian Weng*, University of California-Davis
Danielle J. Harvey, University of California-Davis
Laurel A. Beckett, University of California-Davis

Cerebrovascular disease (CVD) is associated with increased risk of cognitive decline and other causes of dementia. Monitoring CVD's impact on the brain is of particular interest because CVD is preventable and may be treatable. Magnetic resonance imaging (MRI) enables researchers to visualize tissue damage in the brain. Abnormalities of cerebral white matter, seen as hyperintense signals (WMH) on MRI, are a common manifestation of CVD. My research focuses on the development of models for the progression of CVD via WMH as seen on sequential MRI. Data from a single MRI consist of intensity measurements for several hundred thousand spatially defined voxels, each corresponding to a small volume of the brain. We assume that we know the WMH status for each voxel, and that once a voxel has WMH, it will remain so for subsequent MRI. The likelihood of a voxel developing WMH is assumed to depend on characteristics of the individual, on the spatial location of the voxel, and on the history of WMH in nearby voxels. We present an estimation strategy for the parameters in the model and some simulations that illustrate the properties of these estimates.

email: qweng@ucdavis.edu

LONGITUDINAL IMAGE ANALYSIS VIA STOCHASTIC EM ALGORITHM

Xiaoxi Zhang*, University of Michigan
Timothy D. Johnson, University of Michigan
Roderick R.A. Little, University of Michigan

It has been shown that differential radiation dose can induce differential changes in brain/tumor vascular permeability, which potentially enhances the delivery of large chemotherapeutic agents by increasing tumor permeability relative to the brain. It is not only the extend and magnitude of changes in vascular permeability induced by radiation that is of interest, but also the temporal profile of the changes which are crucial in optimizing the administration of chemotherapeutic agents with respect to radiotherapy. We propose using Gaussian Hidden Markov Random Fields (Potts model), a flexible yet powerful model, in this longitudinal image analysis setting. The latent Markov Random Field (MRF) models the spatial structure of the heterogeneous brain/tumor response to radiation. The image sequences given the MRF are assumed to be multivariate Gaussian. This fits into a missing data problem, where the latent MRF state labels are "missing ". However, the spatial structure makes direct application of standard EM infeasible. We approach this spatial-temporal model via the Stochastic EM algorithm. We present simulation results and a preliminary analysis of the real data. Future works will be analysis on multiple patients.

email: xiaoxi@umich.edu

BAYESIAN IMAGE RECONSTRUCTION USING INVERSE REGULARIZATION

Margaret B. Short*, University of Alaska – Fairbanks
David E. Sigeti, Los Alamos National Laboratory
Derek Armstrong, Los Alamos National Laboratory
Diane Vaughn, Los Alamos National Laboratory
David M. Higdon, Los Alamos National Laboratory

In image reconstruction we usually think in terms of smoothing noisy data. At other times, we are presented with images that arise from a known smoothing operator. In such cases, we cannot simply apply the inverse operator since it is highly unstable. We implement a method for reconstructing the original, noisy image by inverse regularization, in which we model the parameter that controls the amount of smoothness in the image. We use MCMC to carry out the calculations. We present simulated data as well as a pRAD (proton radiography) data set. This represents joint work with David Sigeti, Derek Armstrong, Diane Vaughan and David Higdon.

email: enar@mshort.authorized.yon.net

BAYESIAN HIDDEN MARKOV NORMAL MIXTURE MODEL WITH APPLICATION TO MRI TISSUE CLASSIFICATION

Dai Feng*, The University of Iowa

A variety of statistical and other methods have been used in MRI tissue classification. In most previous studies using statistical models, EM-type algorithms were widely used to perform statistical classifications by mixtures of normal distributions with or without taking spatial correlations into consideration. We consider a Bayesian hierarchical framework of normal mixture modeling with the latent state vector indicating the membership of the components to which each observation belongs. The spatial dependency is addressed by modeling the relation among components of the latent state vector by hidden Markov models---the simple Potts model and its variations. We studied the sampling algorithms of different spatial models and the impact of different models on the final results. Some preliminary theoretical results concerning the impact of spatial models are studied. A new strategy to model the partial volume effect is proposed. The comparison of different models is based on the data from the BrainWeb. In order to faster the simulation, the checker board idea which takes advantage of conditional independence is used.

email: dai-feng@uiowa.edu

48. ROC ANALYSIS

AN ALTERNATIVE APPROACH TO PERMUTATION TESTS FOR COMPARING CORRELATED ROC CURVES

Thomas M. Braun*, University of Michigan
Todd A. Alonzo, University of Southern California

In paired studies of two diagnostic modalities, existing permutation tests for testing a difference in areas under the curve (AUCs) between the receiver operating characteristic curve of each modality are executed by permuting the labels of the two modalities within each diseased and non-diseased subject. Such an approach implicitly assumes that both modalities are exchangeable within-subject and require an appropriate transformation, such as ranks, for modalities differing in scale. We instead propose that permutations can be made among subjects, specifically by shuffling the diseased/non-diseased labels of the subjects. No within-subject exchanges are made to the modalities, thereby requiring no transformation of the original data. We prove that our test is valid under the assumption of equal AUCs and show with numerical examples that our test is comparable in power to other permutation tests.

email: tombraun@umich.edu

MULTI-READER ROC METHODS: EXPLICIT FORMULATIONS AND RELATIONSHIPS BETWEEN DBM, MULTI-WMW, BWC AND GALLAS'S METHODS

Andriy Bandos*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh
Brandon D. Gallas, NIBIB/CDRH
David Gur, University of Pittsburgh

Multi-reader receiver operating characteristic (ROC) analysis is frequently employed when assessing the accuracy of a diagnostic system. The nonparametric estimator of the area under the ROC curve (AUC), or equivalently Wilcoxon-Mann-Whitney (WMW) statistic, is an average performance measure that is often used in such an analysis. Here we consider several commonly used multi-reader methods permitting the use of this summary statistic as a primary index including: the ANOVA analysis of the jackknife pseudovalues (DBM); asymptotic procedure for a vector of U-statistics (multi-WMW); the method utilizing bootstrap-based estimates of the variance components (BWC), and a method based on unbiased estimators of the variance of the WMW-based estimator of reader-averaged AUC recently proposed by Gallas. In this presentation we demonstrate that with the WMW-based estimator of the AUC the resampling-based approaches of DBM and BWC permit explicit solutions. We also discuss the relationship between the test-statistics of these methods and the test-statistics of multi-WMW and Gallas's resampling-free procedures for testing differences between two reader-averaged AUCs under a paired design when readers are treated as fixed or random. Finally, we present the results of a simulation study of type I error and power of the considered procedures.

email: anb61@pitt.edu

STRONG APPROXIMATIONS FOR RESAMPLE QUANTILE PROCESSES AND APPLICATION TO ROC METHODOLOGY

Jiezhun Gu*, North Carolina State University-Raleigh
Subhashis Ghosal, North Carolina State University-Raleigh

The receiver operating characteristic (ROC) curve is defined as true positive rate versus false positive rate obtained by varying a decision threshold criterion. It has been widely used in medical science for its ability to measure the accuracy of diagnostic or prognostic tests. Mathematically speaking, ROC curve is a composition of survival function of one population to the quantile function of another population. In this paper, we study strong approximation for the quantile processes of bootstrap and the Bayesian bootstrap resampling distributions, and use this result to study strong approximations for the empirical ROC estimator, the corresponding bootstrap, and the Bayesian versions in terms of two independent Kiefer processes. The results imply asymptotically accurate coverage probabilities for bootstrap and the Bayesian bootstrap confidence bands, and accurate frequentist coverage probabilities of bootstrap and the Bayesian bootstrap confidence intervals for the area under the curve functional of the ROC.

email: sherrygu2001@yahoo.com

INFERENCES FOR A NONPARAMETRIC COMPETING PROBABILITY

Yongming Qu, Eli Lilly and Company
Yan D. Zhao*, Eli Lilly and Company
Dewi Rahardja, University of Indianapolis

The Wilcoxon-Mann-Whitney (WMW) test has been widely used as a nonparametric method to compare a continuous or ordinal variable between two groups. The WMW test essentially compares the competing probability p = P(X>Y) + 0.5 P(X=Y) with 0.5, where X and Y are independent random variables from two distributions. The competing probability is naturally meaningful and equal to the area under the Receiver Operating Characteristics (ROC) curve. To construct a confidence interval (CI) for p, existing methods focused either only on continuous variables or only on ordinal variables. Furthermore, recently developed methods require complicated computation. In this paper, we propose a unified approach to construct a CI for p where the data can be continuous, ordinal, or a mixture of the two. The new approach gives a closed-form solution which is easy to compute. In addition, we propose a small sample modification which allows for constructing a CI even when the estimator for p is 0 or 1. Finally, simulation shows that the performance of the new method is comparable or superior to the existing methods.

email: qu_yongming@lilly.com

SEMIPARAMETRIC LEAST SQUARES ANALYSIS OF CLUSTERED ROC CURVE DATA

Liansheng Tang*, University of Washington
Xiaohua A. Zhou, University of Washington

The receiver operating characteristic (ROC) curve is an ideal tool to evaluate the accuracy of diagnostic tests with continous measurements. In this paper a semiparametric least square method is proposed to deal with the situation when the baseline function is unknown. Its performance is compared with the parametric counterpart and a semiparametric method in a large scale of simulation studies. The method is illustrated on a well-known cancer biomarker data.

email: lstang@u.washington.edu

COMPARISON OF NONPARAMETRIC CONFIDENCE INTERVALS FOR THE AREA UNDER THE ROC CURVE OF A CONTINUOUS-SCALE DIAGNOSTIC TEST

Gengsheng Qin*, Georgia State University
Lejla Hotilova, Georgia State University

The accuracy of a diagnostic test with continuous-scale results is of high importance in clinical medicine. It is often summarized by the area under the ROC curve (AUC). In this paper, we discuss and compare nine nonparametric confidence intervals of the AUC for a continuous-scale diagnostic test. Simulation studies are conducted to evaluate the relative performance of the confidence intervals for the AUC in terms of coverage probability and average interval length. A real example is used to illustrate the application of the recommended methods.

email: gqin@gsu.edu

EVALUATING MARKERS FOR SELECTING A PATIENT'S TREATMENT

Xiao Song*, University of Georgia
Margaret S. Pepe, University of Washington

Selecting the best treatment for a patient's disease may be facilitated by evaluating clinical characteristics or biomarker measurements at diagnosis. We consider how to evaluate the potential of such measurements to impact on treatment selection algorithms. We propose a graphical display, the selection impact (SI) curve, that shows the population response rate as a function of treatment selection criteria based on the marker. The curve can be useful for choosing a treatment policy that incorporates information on the patient's marker value exceeding a threshold. The SI curve can be estimated using data from a comparative randomized trial conducted in the population as long as treatment assignment in the trial is independent of the predictive marker. Asymptotic distribution theory is used to evaluate the relative efficiencies of the estimators. Simulation studies show that inference is straightforward with realistic sample sizes. We illustrate the SI curve and statistical inference for it with data motivated by an ongoing trial of surgery versus conservative therapy for carpal tunnel syndrome.

email: xsong@uga.edu

49. VARIABLE SELECTION

EVALUATION AND COMPARISON OF REGULARIZATION BASED VARIABLE SELECTION METHODS

Michael C. Wu*, Harvard School of Public Health
Tianxi Cai, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

Prediction of disease outcomes using genomic data is challenging due to the fact that the number of covariates could potentially be much larger than the sample size. One approach to incorporating such high dimensional data is to use variable selection techniques to choose important covariates. This may be achieved by penalization methods including the non-negative garrote, LASSO, SCAD and adaptive LASSO. These methods regularize the minimization of a usual empirical loss function by adding a penalty whose magnitude is controlled by a penalty parameter. For any given penalty parameter, one may obtain a regularized estimator for the regression coefficients and the penalty parameter can then be selected by methods such as cross-validation. These approaches have become popular tools in incorporating high dimensional data because they achieve simultaneous shrinkage and variable selection. However, little is studied in making inference in such variable selection procedures and in approximating the distribution of such estimators. Furthermore, it remains unclear how to choose an appropriate penalty parameter to achieve desirable theoretical and empirical performance of the variable selection procedure. We propose to provide insights to these questions through both empirical and theoretical studies.

email: mwu@hsph.harvard.edu

MODEL ESTIMATION AND SELECTION VIA DOUBLE PENALIZED LEAST SQUARES IN PARTIAL LINEAR MODELS

Xiao Ni*, North Carolina State University
Hao Helen Zhang, North Carolina State University
Daowen Zhang, North Carolina State University

We introduce a new approach to simultaneous model estimation and model selection for partial linear models via minimizing double-penalized least squares. The proposed method selects important linear effects using the smoothly clipped absolute deviation (SCAD) penalty, and at the same time estimates nonparametric effects with smoothing splines. The asymptotic properties of the resulting estimators are investigated. To facilitate computation, we represent the method in a linear mixed model framework and obtain parameter estimates by standard software packages. The smoothing parameter can be estimated as a variance component instead of using grid search approach. We also discuss the selection of the tuning parameter in the SCAD penalty. Finally, we derive the frequentist and Bayesian covariances of the model estimate. Simulation and real data analysis are presented to show effectiveness of our method.

email: xni@stat.ncsu.edu

SPARSE PARTIAL LEAST SQUARES REGRESSION WITH AN APPLICATION TO THE GENOME SCALE TRANSCRIPTION FACTOR ACTIVITY ANALYSIS

Hyonho Chun*, University of Wisconsin - Madison
Sunduz Keles, University of Wisconsin - Madison

Analysis of modern biological data often involves ill-posed problems due to high dimensionality and multicollinearity. Partial least squares (PLS) regression has been an alternative to ordinary least squares for handling multicollinearity. At the core of the PLS methodology lies a dimension reduction technique coupled with a regression model. Although PLS regression can achieve good predictive performance, it is not particularly tailored for variable/feature selection and therefore often produces linear combinations of the original predictors that are hard to interpret due to high dimensionality. We propose a sparse partial least squares (SPLS) formulation which aims to simultaneously achieve good predictive performance and variable selection thereby producing sparse linear combinations of the original predictors. Our SPLS formulation utilizes a semi-definite programming framework and the actual algorithm for finding the SPLS directions involve the smoothing technique for approximating non-smooth functionals. We investigate the prediction and variable selection performance of SPLS regression by a simulation study and apply it to the problem of inferring transcription factor activity by integrating gene expression microarray data and genome-wide binding data.

email: chun@stat.wisc.edu

MODEL COMPLEXITY FOR THE AIC STATISTIC WITH NEURAL NETWORKS

Doug Landsittel*, Duquesne University
Dustin Ferris, Duquesne University

Feed forward artificial neural networks (ANNs) are commonly used for applications related to prediction and classification. The advantage of such models, over standard generalized linear models, is their inherently non-linear structure and implicitly fitted interactions. One limitation of such methods, however, is the lack of straightforward approaches for selecting the optimal predictor variables and the optimal number of hidden units. This is particularly significant since ANNs are known to over-fit observed relationships and thus poorly generalize to other data sets. Although some work has been done to establish criteria for model selection, the current methodologies are either ad hoc in nature, or difficult to translate into practical applications. This study proposes use of the AIC statistic with neural networks using one of two different measures of model complexity. For one measure of complexity, we use a modified version of generalized degrees of freedom for a binary outcome. For the other measure, we present simulation results to characterize model complexity under the null assumption of no association. Results are compared to alternative approaches and applied to cancer biomarker data.

email: landsitteld@duq.edu

SIMULTANEOUS SUBSET SELECTION VIA RATE-DISTORTION THEORY, WITH APPLICATIONS TO GENE CLUSTERING AND SIGNIFICANCE ANALYSIS OF DIFFERENTIAL EXPRESSION

Rebecka J. Jornsten*, Rutgers University

We present a novel model selection methodology, based on the well-known principle of bit-allocation in rate-distortion theory. This simultaneous selection strategy turns the combinatorial subset selection problem, across clusters and variables, into a simple line search. Moreover, in the special case where data objects form their own clusters, the methodology generalizes to subset selection in a multiple testing framework. The simultaneous selection method is applied to the analysis of a time course gene expression data set of two proliferating stem cell lines; one cell line forms neurons, the other glial cells. Our analysis provides sparse and easy-to-interpret model representations of gene clusters (e.g. clusters with no cell line/time interaction). By incorporating subset selection into testing, we show that the power of detection is significantly increased, and we obtain sparse model representations for each gene. Thus, we find that the significant list is dominated by main cell line, and static expression in neurons.

email: rebecka@stat.rutgers.edu

FIXED AND RANDOM EFFECTS SELECTION IN LINEAR AND LOGISTIC MODELS

Satkartar K. Kinney*, Duke University
David B. Dunson, National Institute of Environmental Health Sciences

We address the problem of selecting which variables should be included in the fixed and random components of logistic mixed effects models for correlated data. A fully Bayesian variable selection is implemented using a stochastic search Gibbs sampler to estimate the exact model-averaged posterior distribution. This approach automatically identifies subsets of predictors having non-zero fixed effect coefficients or non-zero random effects variance, while allowing uncertainty in the model selection process. Default priors are proposed for the variance components and an efficient parameter expansion Gibbs sampler is developed for posterior computation. The approach is illustrated using simulated data and an epidemiologic example.

email: saki@stat.duke.edu

RAMIFICATIONS OF PRE-PROCESSING ON FEATURE DISCOVERY IN MICROARRAY EXPERIMENTS

Kouros Owzar*, Duke University Medical Center
Sin-Ho Jung, Duke University Medical Center

A challenging task in clinical studies in cancer is the identification of features (genes) on microarrays which are associated with clinical endpoints such as time to death or time to disease recurrence. From a statistical point of view, the challenge is to devise a gene-discovery procedure incorporating the censoring mechanism and to adequately control the false selection error given the large number of features. There are a number of publications discussing non- and semi-parametric methods which attempt to address these statistical issues. What should be noted is that before the gene discovery procedure can be carried out, the raw molecular data is put through a series of pre-processing steps such as normalization and exclusion of features which are deemed to be absent. These pre-processing steps produce the so called gene expressions which will constitute the set of co-variables in the gene discovery procedure. Most of the methods discussed in the literature do not take into account the ramifications of these pre-processing steps on the final results. What is to be presented in this talk is an illustration of the ramifications of various pre-processing methods on the results when applied to an ubiquitously cited data set involving a microarray experiment in lung cancer.

email: kouros.owzar@duke.edu

50. ADAPTIVE DESIGNS: WHERE ARE WE NOW, AND WHERE SHOULD WE BE GOING?

HANDLING UNCERTAINTY ABOUT THE LIKELY TREATMENT EFFECT: THE ROLES OF GROUP SEQUENTIAL DESIGNS AND ADAPTIVE SAMPLE SIZE RE-ESTIMATION

Christopher Jennison*, University of Bath-U.K.

The uncertainty of sample size assumptions is a well known issue in clinical trials and there has been a lively debate on how to design a study with adequate power when there is disagreement about the likely effect size. The 'start small then ask for more' approach uses a small initial sample size but may increase this in response to interim results. Group sequential designs set a higher maximum sample size at the outset but plan to reduce this by stopping early when data allow this. It is of high practical interest to assess the efficiency of these competing approaches and give recommendations on how to choose between them. This talk will compare and contrast the adaptive and group sequential approaches in the context of designing a clinical trial when several plausible effect sizes need to be considered.

email: cj@maths.bath.ac.uk

ASSESSING AND QUANTIFYING THE OPERATIONAL BIAS OF ADAPTIVE DESIGNS

Armin Koch*, BfArM (German Federal Institute for Drugs and Medical Devices)

An important concern in the practical implementation of adaptive designs are the potential changes that may occur as a consequence of a design adaptation, or an information leak at an interim analysis. These may refer to the patient population, the underlying treatment effect, and/or other study characteristics and may compromise the internal validity of the trial. Although it may be reasonable to hypothesize that such changes can occur, it is often difficult to safely identify them by examining the data. This talk will discuss methods which can be used in practice to investigate the presence and magnitude of operational bias in data resulting from an adaptive design and investigate their properties.

email: a.koch@bfarm.de

COMPARISON OF STATISTICAL METHODS AND CHARACTERISTICS OF ADAPTIVE SEAMLESS PHASE II/III TRIALS

Jeff D. Maca*, Novartis Pharmacueticals

The current statistical literature contains different methods for the design and analysis of adaptive seamless Phase II/III designs. When considered in the context of actual clinical trials, each of these methods comes with its own benefits and drawbacks. Using simulated data and real trial examples, various methods will be discussed and compared, with recommendations on their use in practice being presented.

email: jeff.maca@novartis.com

51. ADVANCES IN METHODOLOGY FOR THE ANALYSIS OF RECURRENT EVENT DATA

ESTIMATION OF MARGINAL CHARACTERISTICS FOR RECURRENT EVENT PROCESSES

Richard J. Cook*, University of Waterloo
Jerry F. Lawless, University of Waterloo

Intensity-based modeling of recurrent event data is called for when the focus is on understanding the dynamics of an event process. In settings such as randomized trials, however, we aim to express treatment and possibly other covariate effects so that they are easily interpreted. In this setting, intensity-based methods are less useful and we prefer to base analyses on marginal features of the event process, such as rates or the distribution of event counts. This talk will discuss the pros and cons of intensity-based methods and estimating function methods for estimating marginal features. We consider issues of efficiency, robustness to adaptive censoring, and the impact of terminal events such as death.

email: rjcook@uwaterloo.ca

FINITE AND ASYMPTOTIC PROPERTIES OF ESTIMATORS FOR A GENERAL RECURRENT EVENT MODEL

Edsel A. Peña*, University of South Carolina

In this talk I will discuss finite and asymptotic properties of semi-parametric estimators for the general recurrent event model proposed by Pena and Hollander (2004) and Pena, Slate and Gonzalez (2006). The general model subsumes many recurrent event models arising in biostatistics, reliability, engineering, and other areas, and it incorporates unique aspects inherent in recurrent event monitoring, such as impact of interventions after each event occurrence, sum-quota accrual scheme effects, covariate effects, and the presence of associated inter-event times. The general weak convergence theorem utilized in obtaining the asymptotic properties may also be useful in other situations. Potential utility of the asymptotic results will be mentioned, such as in construction of confidence bands, in testing hypothesis, in goodness-of-fit testing, and in the detection of outliers and influential observations.

email: pena@stat.sc.edu

A SEMIPARAMETRIC NESTED FRAILTY MODEL FOR CLUSTERED BIVARIATE PROCESSES, WITH APPLICATIONS TO RECURRENT AND TERMINAL EVENTS

Emmanuel Sharef, Cornell University
Robert L. Strawderman*, Cornell University

We propose a model for clustered, paired point processes, e.g., recurrence and death. Conditionally on a pair of correlated random effects and covariates, each process is assumed to follow a modulated renewal process (MRP; Cox, 1972). The bivariate random effects are modeled in a hierarchical fashion, each level being specified only through its first two moments. The resulting semiparametric nested frailty model extends that in Ma, Krewski and Burnett (2003), who proposed an analogous Cox regression model for a single outcome in the presence of two layers of clustering; see also Ma, Wilms and Burnett (2002) for related work on recurrent events. Specifically, given the random effects, we construct two auxiliary Poisson likelihoods, each of which is equivalent to the corresponding partial likelihood generated by the MRP model. An iterative scheme for determining the regression parameter is then proposed in which the random effects are replaced by 'orthodox BLUPS,' formulated and computed under these auxiliary Poisson models. We apply this framework to the problem of modeling clustered recurrent event data in the presence of a terminal event. An application cardiac event data is also considered.

email: rls54@cornell.edu

SEMIPARAMETRIC ANALYSIS OF CORRELATED RECURRENT AND TERMINAL EVENTS

Yining Ye*, Amgen Inc.
Jack D. Kalbfleisch, University of Michigan
Douglas E. Schaubel, University of Michigan

In clinical and observational studies, an event can recur on the same subject, and in a more complicated situation, there exists a terminal event (e.g. death), which stops the recurrent event process. In many instances, the terminal event is strongly correlated with the recurrent event process. I propose a semiparametric method to jointly model the recurrent and terminal event processes. The dependence is modeled by a shared gamma frailty that is included in both the recurrent event rate and terminal event hazard function. Marginal models are used to estimate the regression effects on the terminal and recurrent event processes. Different from the existing frailty approaches, we relax the Poisson assumption. Then we extend the analysis to model multiple types of recurrent events and terminal events. An analysis of hospitalization data for patients in the peritoneal dialysis study is presented to illustrate the proposed method.

email: yye@amgen.com

52. COMPUTING YOUR RISK OF DISEASE: ABSOLUTE RISK MODELS

LUNG CANCER RISK PREDICTION

Peter B. Bach*, Memorial Sloan-Kettering Cancer Center

Lung cancer risk prediction can conceivably be useful in the development and analysis of lung cancer prevention and early detection trials, and in counseling indvidiuals at risk. We developed two lung cancer models - one predicting the risk of incident disease, the other predicting the risk of disease specific death, using data from the Beta-Carotene and Retinol Efficacy Trial (CARET) to generate absolute risk prediction models for lung cancer. We then validated the models internally and externally. The models produce interesting results that lend insight into the design and analysis of prevention trials - we will illustrate the implications of the models for risk variation, for analysis of outcomes in screening trials, and for sample size calculations in the context of study development.

email: bachp@mskcc.org

DEVELOPMENT OF A COLORECTAL CANCER RISK ASSESSMENT TOOL

Ruth Pfeiffer*, National Cancer Institute
Andrew Freedman, National Cancer Institute
Mitchell Gail, National Cancer Institute
Marty Slattery, University of Utah

Several modifiable risk factors have been consistently identified for colorectal cancer (CRC). To date, models to estimate absolute CRC risk only exist for special populations. We developed a model to estimate the probability of first incidence of proximal, distal or rectal cancer in White men and women, age 50 years and older, over a defined period of time. Using data from two large population-based case-control studies of colon and rectal cancer, conducted in Utah, Minnesota and Northern California Kaiser Permenante between 1991 and 2001, we estimated relative risks separately for proximal, distal and rectal cancer for several previously identified risk and protective factors. Age specific incidence rates from SEER and attributable risks were used to compute baseline age-specific hazard rates. National mortality rates were used to estimate the hazard for competing causes of death. Variances for absolute risk estimates were computed. The factors in the model include screening and polyp history, family history of CRC, vigorous exercise, regular aspirin/NSAIDS use, smoking, BMI, HRT use and vegetable intake. Relative risk estimates and risk factors differed among proximal, distal and rectal cancer sites. We developed a short questionnaire (5-8 minutes) to ascertain factors for the model and validated it by cognitive testing.

email: pfeiffer@mail.nih.gov

RISK PREDICTION IN THE FRAMINGHAM HEART STUDY

Lisa M. Sullivan*, Boston University
Michael Pencina, Boston University
Ramachandran Vasan, Boston University
Ralph D'Agostino, Boston University

For almost 50 years, the Framingham Heart Study (FHS) has been a leader in the development of mathematical models relating risk factors to incident cardiovascular events. Multivariable statistical models combining risk factor information have evolved over time, starting with discriminant function analysis and currently involving multivariable proportional hazards modeling. The FHS produced risk functions for predicting initial and recurrent coronary heart disease, stroke and peripheral artery disease; the functions can be used to generate estimates of short- (e.g., 2 year) and longer-term risk (e.g., 10 years). A recent and widely publicized use of the FHS functions is the National Cholesterol Education Program's Adult Treatment Panel (ATP) III report. The ATPIII report includes a risk function which produces an estimate of an individual's 10-year risk of coronary heart disease as a function of age, sex, systolic blood pressure, cholesterol level, smoking and diabetes status. An individual's absolute coronary risk is used to determine the intensity of risk factor reduction therapy. The Framingham functions have been evaluated in numerous demographically and clinically diverse groups. With recalibration, the functions perform well across a range of ethnically diverse groups with different levels of risk factors and with different incidence of coronary disease.

email: lsull@bu.edu

53. HIERARCHICAL MODELING OF LARGE BIOMEDICAL DATASETS

LATENT ASPECTS ANALYSIS FOR GENE EXPRESSION DATA

Edoardo M. Airoldi*, Princeton University
Stephen E. Fienberg, Carnegie Mellon University
Eric P. Xing, Carnegie Mellon University

Discovering salient gene expression patterns that explain a biological process is a central problem in contemporary biology. In a typical SAGE experiment, the observations consist of a large pool of temporal gene expression profiles. In this talk, we discuss hierarchical models that enable the discovery of salient gene expression patterns, in the absence of prior knowledge about their characteristics. The proposed models capture the notion of contagion, defined as dependence among multiple occurrences in the sample of the same gene identifier. Contagion is a convenient analytical formalism to capture semantic themes underlying the observed temporal profiles, such as biological context. Model variants are introduced that are tailored to various properties of SAGE data. A general variational inference scheme for fast, approximate posterior inference is presented. The methodology is validated on both simulated data, and realistic high-throughput gene expression profiles via SAGE. Our results show improved predictions of gene functions over existing methods based on stronger independence assumptions, and demonstrate feasibility of a promising hierarchical Bayesian formalism for soft clustering and latent aspects analysis.

email: eairoldi@cs.cmu.edu

A TWO STAGE METHOD FOR FITTING SEMIPARAMETRIC RANDOM EFFECTS MODELS TO LARGE DATA SETS

Michael L. Pennell*, The Ohio State University
David B. Dunson, NIEHS

In some biomedical studies, such as multi-center studies or large-scale prospective studies, massive amounts of hierarchical data are collected. In these settings, it would be advantageous to use the abundant information to relax model assumptions, such as the normality of random effects. Unfortunately, it can be difficult to fit parametric and semiparametric random effects models to large data sets due to the memory limitations of existing software or the lengthy runtimes needed for some standard methods to converge. Motivated by data from an epidemiologic study of childhood growth, we propose a two stage method for fitting semiparametric random effects models to longitudinal data with many subjects. In the first stage, we use a multivariate clustering method to identify G << N groups of subjects whose data have no scientifically important differences, as defined by subject matter experts. Then, in Stage 2, group-specific random effects are assumed to come from an unknown distribution, which is assigned a Dirichlet process prior, further clustering the groups from Stage 1. We use our approach to model the effects of maternal smoking during pregnancy on growth in 17,518 girls.

email: mpennell@sph.osu.edu

COMPUTATIONALLY EFFICIENT ANALYSIS OF SPATIALLY VARYING DISEASE RATES FOR LARGE DATA SETS

Louise M. Ryan*, Harvard University

The large size of many hospital record and other administrative databases limits the applicability of many reasonable statistical models for disease incidence and mortality rates. We describe an efficient algorithm for fitting Poisson regression, using the statistical principle of sufficiency and the convenient collapsibility property of Poisson models. The algorithm is easily implemented using standard software for fitting generalized linear models and avoids the need for data reduction techniques such as standardization. We extend the algorithm to accommodate spatial correlation using a Conditional Autoregressive (CAR) model and illustrate our results with an analysis that explores the relationship between socio-economic factors and heart disease in New South Wales, Australia. Data on emergency hospital admissions for ischemic heart disease over a 5 year period were linked at the postcode level with the Socio-Economic Indexes for Areas (SEIFA) reported by the Australian Bureau of Statistics. Although our dataset contained approximately 33 million observations, our algorithm resulted in fast, stable analysis and revealed several interesting and previously unknown patterns in the data.

email: lryan@hsph.harvard.edu

54. SEMIPARAMETRIC REGRESSION METHODS FOR LONGITUDINAL DATA ANALYSIS

NONPARAMETRIC/SEMIPARAMETRIC REGRESSION FOR INCOMPLETE LONGITUDINAL DATA USING EM

Xihong Lin*, Harvard School of Public Health
Raymond Carroll, Texas A&M University

There have been considerable recent developments of nonparametric and semiparametric regression methods for complete longitudinal data. We consider nonparametric and semiparametric methods for incomplete longitudinal data using kernel and splines methods and profile methods. Such methods are developed within a likelihood framework using the kernel/spline and profile EM methods. Theoretical properties of these methods are investigated. The proposed methods will be evaluated using simulation studies and applied to several real world data examples.

email: xlin@hsph.harvard.edu

GENERALIZED SEMIPARAMETRIC LINEAR MIXED EFFECTS MODELS FOR LONGITUDINAL DATA

Tatiyana V. Apanasovich*, Cornell University, School of ORIE

Generalized linear mixed effects models are widely used for longitudinal non-Gaussian data analysis to incorporate between-subject and within-subject variations. To weaken model assumption for possible misspecification and to avoid the curse of dimensionality of fully nonparametric regression in the presence of several predictor variables, semiparametric models have been considered. We employ the penalized likelihood regression to estimate the models. We will focus on the efficient computation and the effective smoothing parameter selection. Real-data examples from AIDS studies will be presented to demonstrate the applications of the methodology.

email: tva2@cornell.edu

MARKOV CHAIN MARGINAL BOOTSTRAP FOR LONGITUDINAL DATA

Di Li, University of Illinois at Urbana-Champaign
Xuming He*, University of Illinois at Urbana-Champaign

Generalized estimating equations (GEE) are frequently used in the analysis of longitudinal data. In some cases, it remains a challange to find the 'correct' solution to the GEE for a parameter of interest, and furthermore, the standard large-sample approximation to the variance-covariance matrix of the estimate could lead to poor or unstable inference. We extend the Markov chain marginal bootstrap (MCMB) approach, developed originally for independent data, to the GEE-based inference, and show both theoretically and empirically that it can provide reliable inference when we are unsure about the reliability of standard methods. The MCMB method reduces a high-dimensional problem into several one-dimensional problems before re-sampling is used for inference.

email: x-he@uiuc.edu

ANALYSIS OF LONGITUDINAL DATA WITH SEMIPARAMETRIC ESTIMATION OF COVARIANCE FUNCTION

Jianqing Fan*, Princeton University
Tao Huang, University of Virginia
Runze Li, Pennsylvania State University

Improving efficiency for regression coefficients and predicting trajectories of individuals are two important aspects in analysis of longitudinal data. Both involve estimation of the covariance function. Yet, challenges arise in estimating the covariance function of longitudinal data collected at irregular time points. A class of semiparametric models for the covariance function is proposed by imposing a parametric correlation structure while allowing a nonparametric variance function. A kernel estimator is developed for the estimation of the nonparametric variance function. Two methods, a quasi-likelihood approach and a minimum generalized variance method, are proposed for estimating parameters in the correlation structure. We introduce a semiparametric varying coefficient partially linear model for longitudinal data and propose an estimation procedure for model coefficients by using a profile weighted least squares approach. Sampling properties of the proposed estimation procedures are studied and asymptotic normality of the resulting estimators is established. Finite sample performance of the proposed procedures is assessed by Monte Carlo simulation studies. The proposed methodology is illustrated by an analysis of a real data example.

email: jqfan@princeton.edu

55. MODEL SELECTION AND ASSESSMENT IN GEE

COVARIANCE MODEL SELECTION IN GEE

Vincent J. Carey*, Harvard Medical School
You-Gan Wang, CSIRO
Lin Y. Hin, Private Medical Practitioner

We compare the motivations and performance of a variety of data-based criteria for the selection of working covariance models in generalized estimating equations for clustered responses.

email: stvjc@channing.harvard.edu

SOME ASPECTS OF MODEL ASSESSMENT WHEN USING GEE

Bahjat F. Qaqish*, University of North Carolina at Chapel Hill
John Preisser, University of North Carolina at Chapel Hill

We discuss some issues in the selection and fitting of models using GEE. One class of issues concerns the identification of what parameters GEE estimates in non-standard situations such as non-random sampling and random cluster size. Another class concerns the construction of efficient estimating functions. A third class concerns the related issues of regression diagnostics, residuals and goodness of fit.

email: bahjat_qaqish@unc.edu

CONSISTENT MODEL SELECTION AND DATA-DRIVEN SMOOTH TESTS FOR LONGITUDINAL DATA IN THE ESTIMATING EQUATION APPROACH

Lan Wang, University of Minnesota
Annie Qu*, Oregon State University

An important problem facing marginal regression analysis of longitudinal data, as in the method of generalized estimating equations, is how to choose a marginal regression model from a number of candidate models. Although several methods have been suggested in the literature for practical use, theoretical investigation of the large sample theory is still lacking. We propose a new BIC-type model selection criterion in this paper, and prove that with probability approaching one it selects the most parsimonious correct model. The model selection criterion uses the quadratic inference function proposed by Qu, Lindsay and Li (2000) and does not need to specify the full likelihood or quasilikelihood. This model selection procedure also motivates a data-driven Neyman-type smooth test for checking the goodness-of-fit of a conjectured model. Compared to the classical tests which require the specification of an alternative, such as the GEE Z-test, the new test selects a data-driven alternative based on model selection and leads to increased power performance in general. The finite sample performance of the new model selection and model checking procedures is demonstrated through Monte Carlo studies and analysis of a clinical trial data set.

email: qu@stat.oregonstate.edu

ON THE IMPACT AND LIKELIHOOD OF A VIOLATION OF BOUNDS FOR THE CORRELATION IN GEE ANALYSES OF BINARY DATA FROM LONGITUDINAL TRIALS AND WHAT WE CAN DO TO ADDRESS THIS PROBLEM

Justine Shults*, University of Pennsylvania School of Medicine
Wenguang Sun, University of Pennsylvania School of Medicine
Xin Tu, University of Rochester
Thomas R. Ten Have, University of Pennsylvania School of Medicine

The issue of violation of bounds for the correlation has been the topic of recent discussion in the statistical literature. We consider a clinical trial whose analysis with GEE and quasi-least squares results in a violation of bounds. We then tackle the following questions: (1) Why is this issue of violation of bounds important?; (2) When is a violation of bounds more likely to occur?; (3) Can a violation of bounds ever be beneficial? and (4) If we choose to overcome a violation of bounds in our analysis, which approach is the best? We demonstrate that misspecification of the working correlation structure can lead to a severe violation of bounds, both asymptotically and for small samples. As a result, the potential for violation of bounds might be viewed as an additional useful tool for choosing an appropriate working correlation structure. For a working structure that is otherwise plausible and results in only a minor violation, we propose an iterative algorithm that yields an estimate that satisfies the constraints. We make general recommendations and compare our suggestions with working independence, an ad-hoc approach that has been suggested recently, and a likelihood approach based on a Markovian dependence model. In addition, we discuss an approach for choosing between alternative correlation structures.

email: jshults@cceb.med.upenn.edu

56. RECURRENT EVENTS

ROBUST METHODS FOR ANALYZING RECURRENT EVENTS IN PRESENCE OF TERMINAL EVENTS

Rajeshwari Sundaram*, National Institute of Child Health and Human Development, NIH

Recurrent event data are frequently encountered in biomedical studies: infections in AIDS patients and seizures in epilepsy patients. In addition to loss to follow up, such data are further complicated by the presence of terminal events like death, which obviously precludes subsequent recurrences. Here, we consider a family of semiparametric transformation models for the cumulative mean function of such recurrent events process over time in presence of terminal events. In the proposed method, we first model the survival times of the individuals and using them appropriately, define a class of robust estimators for the regression parameters. The asymptotic properties like consistency and asymptotic normality of the proposed estimators are established. Finite sample properties are examined through extensive simulations. We conclude with a real data example.

email: sundaramr2@mail.nih.gov

ACCELERATED FAILURE TIME MARGINAL MEANS MODELS FOR RECURRENT EVENTS WITH A TERMINAL EVENT

Xiaoyan Wang*, University of North Carolina at Chapel Hill
Jianwen Cai, University of North Carolina at Chapel Hill

Recurrent events with a terminal event such as death arise in many application areas. In the presence of terminal events, we consider a marginal mean model, assuming the accelerated failure time (AFT) model for counting process (Lin, Wei and Ying, 1998) holds for the recurrent event means function marginally across survival status. Estimators for the regression parameters and the baseline mean function are proposed, which are shown to be consistent and asymptotically normal. We investigate the finite-sample properties of the proposed estimators through simulation studies and illustrate the proposed method through an application to recurrent hospitalization data taken from the Studies of Left Ventricular Dysfunction (SOLVD).

email: yanyan@email.unc.edu

ACCELERATED TESTING WITH RECURRENT EVENTS

Alexander C. McLain*, University of South Carolina
Edsel A. Peña, University of South Carolina

In many studies, it is oftentimes of interest to infer about lifetime characteristics at a pre-specified stress level, such as at the operating stress level. To do so in a timely manner, an accelerated testing framework may be implemented. When the number of stress sites is limited a recurrent events framework may increase efficiency and shorten the amount of time needed. Important issues of estimation of parameters, planning of statistically optimal test stresses and allocations will be addressed. Estimation of parameters is approached through maximum likelihood, and their asymptotic properties are obtained. Simulation studies show that in small to moderate sample size's the asymptotic properties of our estimators hold under the given assumptions. Keywords: Accelerated Testing, Recurrent Events, Counting Processes, Martingales.

email: alex.mclain@gmail.com

SEMIPARAMETRIC ESTIMATION OF THE GAP-TIME DISTRIBUTION WITH RECURRENT EVENT DATA UNDER AN INFORMATIVE MONITORING PERIOD

Akim Adekpedjou*, University of South Carolina
Edsel A. Peña, University of South Carolina

We consider a biomedical study which monitors the occurrences of a recurrent event for n subjects over a random observation window for each subject. We assume that the distribution of the random observation window is informative regarding the distribution of event time. The problem of semiparametric estimation of the cumulative hazard and consequently of the gap-time distribution is considered under a model of informative censoring called the Koziol Green model. We derive a Nelson-Aalen and Kaplan-Meier type estimators for the distribution function under the specified model. Asymptotic and small sample properties of the proposed estimators are established. The proposed estimators are compared to the nonparametric estimator proposed in Pena et al. (2001, JASA) to ascertain any efficiency gain achieved by exploiting the Koziol-Green structure.

email: adek@stat.sc.edu

GOODNESS OF FIT FOR COMPOSITE HYPOTHESIS IN THE GENERAL RECURRENT EVENT SETTING

Jonathan T. Quiton*, University of South Carolina
Edsel A. Peña, University of South Carolina

A smooth goodness of fit test for the general recurrent event model is proposed. The method used Neyman's embedding technique on the hazard functions and the stochastic process formulation of the likelihood function. The proposed statistic was the quadratic form of the score test with adjustments due to the nuisance parameter. Asymptotic properties of the test was obtained under a sequence of local alternatives, with the consideration that the nuisance parameter is estimated under the null hypothesis. Closed form expressions for the Homogeneous Poisson Process (HPP) model were obtained, and the procedure was demonstrated in the Migratory Motor Complex (MMC)data.

email: quiton@stat.sc.edu

RECURRENT COMPETING RISKS WITH RANDOM MONITORING PERIOD

Laura L. Taylor*, University of South Carolina
Edsel A. Peña, University of South Carolina

A parametric model for Q competing risks in a recurrent event setting with a random monitoring period and under a completely perfect intervention strategy is studied. We address the problem of estimating a p-dimensional parameter for the cause specific hazard functions associated with the Q competing risks, where p denotes the total dimension of the Q parameter vectors of the cause specific hazard functions combined. Estimation is approached through maximum likelihood, and the asymptotic properties of our estimator are obtained. We present the results of a simulation study demonstrating the small to moderate sample size properties of our estimator, and we apply our methods to a real data set.

email: taylor@stat.sc.edu

A MAXIMUM LIKELIHOOD APPROACH TO ACCOMMODATING CENSORING IN AUTOREGRESSIVE LINEAR MODELS ASSUMING MULTIVARIATE NORMALITY WITH APPLICATIONS TO MODELING HIV RNA CONCENTRATION IN PLASMA

Caitlin Ravichandran*, Harvard University
Victor DeGruttola, Harvard University

We consider a maximum likelihood method for fitting autoregressive models of arbitrary order and covariance structure to response variables that may be censored or missing. Our work is motivated by the investigation of the effect of specific genetic mutations on response of HIV RNA concentration in plasma to treatments provided in the GART Study, a randomized controlled trial that compared treatment guided by HIV viral genotype to treatment guided by clinical history only. We test whether a resistance mutation, m184v, with and without use of a drug associated with this mutation, 3TC, affects RNA response before and after viral rebound. To characterize RNA response, we use an autoregressive model assuming normally distributed errors. To accommodate the censoring of RNA values at 250 copies per ml, an EM algorithm in which the expectation of the complete data log likelihood is calculated by numerical integration is used to maximize the likelihood. We describe methods for hypothesis testing and estimating the information matrix using efficient sampling. Results from the GART study show that m184v is associated with increasing viral load after rebound among those patients not taking 3TC. A simulation study demonstrates the good properties of our estimators with moderately sized samples.

email: cthomas@hsph.harvard.edu

57. DIAGNOSTICS I

TWO NEW MEASURES OF CARDIAC DYSSYNCHRONY BASED ON EIGENVALUE DECOMPOSITIONS OF CORRELATION MATRICES

Peter M. Meyer*, Rush University Medical Center
Matthew Weinberg, Rush University Medical Center
Richard Miller, Rush University Medical Center
Jeffrey Neiger, Rush University Medical Center
Steven B. Feinstein, Rush University Medical Center

New measures of radial strain in heart muscle segments using 2D echocardiography provide insight into cardiac dysfunction. The variation in dyssynchrony among segmental strain reflects differing etiologies and aid the clinician in identifying appropriate treatments. One dyssynchrony measure has been suggested based on the timing variation in peak systolic strain (Kapetanakis et al. Real-time three-dimensional echocardiography: a novel technique to quantify global left ventricular mechanical dyssynchrony. Circulation. 112(7):992-1000, 2005 Aug 16.) We propose two additional dyssynchrony measures based 1) on an eigenvalue decomposition of the correlation matrix of the strain traces and 2) on the angle between the first eigenvector of the correlation matrix and the equiangular vector. In a convenience sample of 15 normal and 10 abnormal echo traces these two measures have an ROC area under the curve of 0.94 for distinguishing between normal and abnormal traces.

email: pmeyer@rush.edu

BAYESIAN ESTIMATION OF SUMMARY ROC CURVE FOR META-ANALYSIS OF DIAGNOSTIC TEST PERFORMANCE

Scott W. Miller*, Medical University of South Carolina
Debajyoti Sinha, Medical University of South Carolina
Elizabeth H. Slate, Medical University of South Carolina
Donald Garrow, Medical University of South Carolina
Joseph Romagnuolo, Medical University of South Carolina

The most commonly used method for meta-analysis of diagnostic test performance is the Summary Receiver Operator Characteristic (SROC) curve approach of Moses, Shapiro and Littenberg. We propose a novel Bayesian model and associated method for the meta-analysis of diagnostic tests to generate a SROC curve when the heterogeneity among the studies cannot be explained using only the unknown study-specific threshold of disease categorization. Our model is conceptually straightforward and accommodates both within and between-study variation, as well as the impact of covariates. Our method can be used effectively for prediction of results from future studies, assessment of the influence of each study on the overall conclusion, and for making detailed evaluation of the nature of the heterogeneity among studies. The method is applied to a data-set comparing endoscopic ultrasound (EUS) to endoscopic retrograde cholangiopancreatography (ERCP) in the detection of biliary obstructions.

email: millersw@musc.edu

EXPLORING PARALLEL AND COMBINED TESTING STRATEGIES FOR MAMMOGRAPHY AND OTHER IMAGING TESTS

Deborah H. Glueck*, University of Colorado at Denver and Health Sciences Center
Molly M. Lamb, University of Colorado at Denver and Health Sciences Center
John M. Lewin, Diversified Radiology of Colorado
Pisano D. Etta, University of North Carolina at Chapel Hill School of Medicine
Keith E. Muller, University of Florida

The correlation between two test results in a parallel testing strategy strongly affects the shape of the ROC curve for a combined test. A combined test is positive if it exceeds the ROC cutoff for test 1, test 2, or both. As an example, we explore the utility of a combined test in the Lewin et al. (2002) study. Lewin et al. conducted a trial in which 6,736 women had both film and digital mammograms. The study detected no significant difference in area under the curve (AUC) between the two mammography techniques. However, the AUC for the combined test was significantly greater than that for either film or digital mammography in a nonparametric analysis. In order to explore this phenomenon, we reformulate the classic binormal model. We account for the potential that the correlation between ROC scores for women with cancer may be different from the correlation between ROC scores for women without cancer. We are able to characterize the combined ROC curve analytically. We demonstrate that low correlation between ROC scores for women with cancer and high correlation between ROC scores for women without cancer creates the maximum AUC for the combined test. This has implications for medical decision-making.

email: Deborah.Glueck@uchsc.edu

MULTIVARIATE MIXTURES OF POLYA TREES FOR MODELING ROC DATA

Tim Hanson, University of Minnesota
Adam Branscum*, University of Kentucky
Ian Gardner, University of California-Davis

Receiver operating characteristic (ROC) curves provide a graphical measure of diagnostic test accuracy. Because ROC curves are determined using the distributions of diagnostic test outcomes for noninfected and infected populations, there is an increasing trend to develop flexible models for these component distributions. We present methodology for joint nonparametric estimation of several ROC curves from multivariate serologic data. We develop an empirical Bayes approach that allows for arbitrary noninfected and infected component distributions that are modeled using Bayesian multivariate mixtures of finite Polya trees priors. Robust, data-driven inferences for ROC curves and the area under the curve are obtained, and a straightforward method for testing a Dirichlet process versus a more general Polya tree model is presented. Computational challenges can arise when using Polya trees to model large multivariate data sets that exhibit clustering. We discuss and implement practical procedures for addressing these obstacles, which are applied to bivariate data used to evaluate the performances of two ELISA tests for detection of Johne's disease.

email: abran3@email.uky.edu

ML INFERENCE ON ROC SURFACES IN THE PRESENCE OF DIFFERENTIAL VERIFICATION

Yueh-Yun Chi*, University of Washington
Xiao-Hua Zhou, University of Washington

In diagnostic medicine, the Receiver Operating Characteristic (ROC) surface is one of the commonly used tools for assessing the accuracy of a diagnostic test in distinguishing three disease states, and the volume under the ROC surface has served as a summary index for diagnostic accuracy. In practice, the selection for definitive disease examination may be based on initial test measurements, and induces verification bias in the assessment. We propose here a nonparametric likelihood-based approach to construct the empirical ROC surface in the presence of differential verification, and to estimate the volume under the ROC surface. Estimators of the standard deviation are derived by both the information and Jackknife method, and their relatively accuracy is evaluated in an extensive simulation study. The methodology is further extended to incorporate discrete baseline covariates in the selection process, and to compare the accuracy of a pair of diagnostic measurements. We apply the proposed method to compare the diagnostic accuracy between Mini-Mental State Examination and clinical evaluation of dementia, in distinguishing among three disease states of Alzheimer's disease.

email: yychi@u.washington.edu

AN EXACT TEST FOR DETECTING INCONSISTENCY IN READER'S INTERPRETATION SCREENING MAMMOGRAMS

Ji-Hyun Lee*, H. Lee Moffitt Cancer Center & Research Institute, The University of South Florida
Steven Eschrich, H. Lee Moffitt Cancer Center & Research Institute, The University of South Florida

Radiologists' interpretation on screening mammograms is measured by accuracy indices such as sensitivity and specificity. We have tested the assumption that individual radiologists are consistent in their image interpretation over time (Lee et al., 2006). This work showed that reader variability across time may be a source of inconsistency in interpretation of screening mammograms. Here, we address the question of consistency across radiologists at distinct time points. In this study, we propose an efficient method to use an exact conditional distribution for testing overdispersion of the binomial assumption. This proposed algorithm for computing the exact test utilizes common characteristics of the problem in a constraint-based search to reduce computational demand. This test is applied to data from a study in reading screening mammograms in a population of 110 U.S. radiologists to determine if inconsistency exists, on average, across all radiologists over time. The exact method is compared analytically with a currently available method based on large sample approximations.

email: leej@moffitt.usf.edu

WHEN SENSITIVITY DEPENDS ON AGE AND TIME SPENT IN PRECLINICAL STATE IN CANCER SCREENING

Dongfeng Wu*, Mississippi State University
Gary L. Rosner, University of Texas, M.D. Anderson Cancer Center
Lyle D. Broemeling, University of Texas, M.D. Anderson Cancer Center

This research extends previous probability models in periodic cancer screening exams. The specific aim is to investigate how screening sensitivity changes with a woman's age, and the time spent in the preclinical state, while the transition probability into the preclinial state is age-dependent. We apply our model to the HIP study of female breast cancer and provide statistical inference for the parameters involved. We hope this will provide more information to policy makers, regarding a screening program's sensitivity, sojourn time distribution and transition probability.

email: dwu@math.msstate.edu

58. HEALTH SERVICES RESEARCH AND MEDICAL COST

A RANDOMIZATION TEST OF RATER AGREEMENT WITH GROUPS OF RATERS

A. John Bailer*, Miami University
Robert B. Noble, Miami University

The assessment of the care needs of nursing home residents may ultimately be linked to reimbursement for their care. These assessments are conducted at facilities by an individual or team of individuals. In this presentation, we consider the agreement in the scoring of individuals by facility values and by an independent rater. Issues that surface in this analysis include: 1) a facility rater provides scores for all residents in a facility; and 2) different independent rates rated different facilities. The analysis of agreement in these ratings needed to reflect these dependencies, and we describe a randomization procedure that implements this.

email: baileraj@muohio.edu

ANALYZING PRESSURE ULCER DEVELOPMENT OF 36 NURSING HOMES USING BAYESIAN HIERARCHICAL MODELING

Jing Zhang*, University of Missouri
Zhuoqiong He, University of Missouri
David R. Mehr, University of Missouri

Pressure ulcer development is an important measurement in judging the quality of nursing service, therefore it is an interesting research topic to find out the effect of nursing homes in the measurements of ulcer development rates. A lot of different methods have been used, including traditional logistic regression, Bayesian Hierarchical modelling and semi-parametric approaches. In this report we use Bayesian hierarchical Models and apply MCMC algorithm to predict the performance of 36 nursing homes in term of pressure ulcer development of the residents in these institutes. Bayesian approach provides posterior distributions besides estimates and predictions, therefore we could develop different criteria for detecting the nursing homes that are providing problematic nursing services. We also performed cross validation to evaluate the predictive ability of the Bayesian Hierarchical model.

email: jztw7@mizzou.edu

DISTRIBUTION OF DATA ENVELOPMENT ANALYSIS EFFICIENCY SCORES: AN APPLICATION TO NURSING HOMES' CARE PLANNING PROCESS

Byron J. Gajewski*, University of Kansas
Robert Lee, University of Kansas
Marjorie Bott, University of Kansas
Ubolrat Piamjariyakul, University of Kansas
Roma Lee Taunton, University of Kansas

Traditionally, data envelopment analysis (DEA) is a deterministic econometric modeling procedure for calculating efficiency using data from an observed set of units. We propose a method for calculating the distribution of efficiency scores estimated by the DEA. Our framework relies on estimating the unobserved data with a fully Bayesian model that is built using the data from the observed units. The model provides posterior predictive data for the unobserved units, thus augmenting the frontier in the DEA which provides researchers a posterior predictive distribution for the efficiency scores. The proposed method results in a stochastic interpretation to DEA. We initially present the method with a simple example using one-input and one-output. We then further explore the method on a more complicated multiple-input and multiple-output DEA model. The data for this example come from a comprehensive examination of how nursing homes complete a standardized, mandatory assessment of residents. To assess the efficiency of this process our research team gathered detailed process and resource use data from a random sample of 107 of nursing homes from a population of 444 nursing homes.

email: bgajewski@kumc.edu

BAYESIAN SAMPLE SIZE CALCULATION FOR COUNTED DATA WITH EXCESS ZEROES

Chunyao Feng*, Amgen Inc.
James Stamey, Baylor University
John Seaman, Baylor University

Hospitalization is the largest portion of the total cost of heart failure patients to the health services. Interest lies in comparing the rate of hospitalization for heart failure patients with and without anemia. We consider a Bayesian method to determine the required sample size to test equality of these rates with a pre-specified power. Generalized Poisson (GP) and Negative binomial (NB) models are considered as the candidates to model the rate of hospitalization. In addition, we compare the performance of these two models via Bayesian model comparison techniques. Finally, a sensitivity analysis is performed to test the robustness of this methodology.

email: amyfeng@gmail.com

USING HIERARCHICAL MODELS TO ESTIMATE A WEIGHTED AVERAGE OF STRATUM-SPECIFIC PARAMETERS

Babette A. Brumback*, University of Florida
Larry H. Winner, University of Florida
George Casella, University of Florida
Allyson Hall, University of Florida
Paul Duncan, University of Florida

Many applications of statistics involve estimating a weighted average of stratum-specific parameters. We consider the case of known weights, motivated by a stratified survey sample of Medicaid participants in which the parameters are population stratum means and the known weights are determined by the population sampling frame. Assuming heterogeneous parameters, it is common to estimate the weighted average with the weighted sum of sample stratum means; under homogeneity, one ignores the known weights in favor of precision weighting. We focus on a general class of estimators, based on hierarchical models, that encompasses these two methods but which also includes adaptive approaches. One basic adaptive approach corresponds to using the DerSimonian and Laird (1986) model for the parameters. We compare this with a novel alternative, which models the variances of the parameters as inversely proportional to the known weights. For two strata, the two approaches coincide, but for three or more strata, they differ.

email: bbrumback@phhp.ufl.edu

EMPIRICAL LIKELIHOOD INFERENCE FOR THE CALIBRATION REGRESSION MODEL WITH LIFETIME MEDICAL COST

Yichuan Zhao*, Georgia State University
Min Lu, Georgia State University

Medical cost has received increasing interest recently in Biostatistics and public health. Statistical analysis and inference of life time medical cost have been challenging by the fact that the survival times are censored on some study subjects and their subsequent cost are unknown. Huang (2002) proposed the calibration regression model which is a semiparametric regression tool to study the medical cost associated with covariates. In this talk, an inference procedure is investigated using empirical likelihood ratio method. The adjusted empirical likelihood confidence regions are constructed for the regression parameters. We compare the proposed empirical likelihood method with normal approximation based method. Simulation results show that the proposed empirical likelihood ratio method outperforms the normal approximation based method in terms of coverage probability. In particular, the adjusted empirical likelihood overcomes the under coverage problem.

email: matyiz@langate.gsu.edu

ADJUSTMENT OF MULTIPLE CARDIOVASCULAR RISK FACTORS WITH A SUMMARY RISK SCORE

Patrick G. Arbogast*, Vanderbilt University
Hua Ding, Vanderbilt University
Wayne A. Ray, Vanderbilt University

To simultaneously adjust for confounding by multiple cardiovascular risk factors, recently-published large pharmacoepidemiologic studies utilized an index of risk of cardiovascular disease (a cardiovascular risk score). This summary measure is a multivariate confounder score created from regression models relating these risk factors to the outcome. It is then used in regression models to adjust for confounding on the exposure of interest. This summary score has a number of advantages. However, there is concern that it may result in underestimation of the standard error of the exposure estimate and thus to an excess of statistically significant results. We conducted simulation studies comparing regression models adjusting for all risk factors directly to models using this summary risk score for large cohort studies. Results indicated that estimated standard errors from the regression models using this summary risk score approximated their empirical standard errors well and were similar to the standard errors from the regression models directly adjusting for all of the risk factors. Based on these simulation results, this summary risk score is a reasonable approach for summarizing many risk factors in large cohort studies.

email: patrick.arbogast@vanderbilt.edu

59. MICROARRAY EXPRESSION ANALYSIS AND ARRAY CGH

HIERARCHICAL MIXTURE MODEL FOR PAIRED MICROARRAY EXPERIMENTS

Haiyan Wu*, Emory University
Ming Yuan, Georgia Institute of Technology
Rama Akondy, Emory University
M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington

Paired microarray experiments are popular among human genetic studies. In this paper, we construct a hierarchical mixture model for paired microarray experiments. Besides a parametric version of the model, we also provide a semiparametric version of the model for a more flexible fitting of the data. In addition, we present a model to deal with paired microarray experiments with missing data. The EM algorithm was used to fit the model. Gene-specific posterior probabilities of upregulation, downregulation, and no changes are obtained to identify differentially expressed genes and to measure the significance of genes. Model diagnostics are provided to investigate the fit of the model to the data. We also present a method to evaluate the power of paired microarray experiments along with sample size calculation for experimental design. We conduct a simulation study to evaluate the hierarchical mixture model and illustrate the proposed model using a human na?ve and effector CD8 T cell comparison experiment.

email: hwu3@sph.emory.edu

VARIABLE SELECTION FOR VARYING COEFFICIENTS MODELS, WITH APPLICATIONS TO ANALYSIS OF MICROARRAY TIME COURSE GENE EXPRESSION DATA

Lifeng Wang*, University of Pennsylvania School of Medicine
Hongzhe Li, University of Pennsylvania School of Medicine

Since many important biological systems or processes are dynamic systems, it is important to study the gene expression patterns over time in a genomic scale in order to capture the dynamic behavior of gene expression. Microarray technologies have made it possible to measure the gene expression levels of essentially all the genes during a given biological process. In order to determine the transcriptional factors involved in gene regulation during a given biological process, we propose to develop a functional response model with time-varying coefficients in order to model the transcriptional effects on gene expression levels, and to develop a penalized procedure for selecting variables with time-varying coefficients. The proposed procedure combines regression spline idea with the smoothly clipped absolute deviation penalty to perform grouped variables selection. Simulation studies indicated that such a procedure is quite effective in selecting the relevant variables and in estimating the time-varying coefficients. The proposed method is applied to mouse breast cancer development MTC data set to identify the transcriptional factors that might be involved in breast cancer development.

email: lifwang@mail.med.upenn.edu

ARE BOEC CELLS MORE LIKE LARGE VESSEL OR MICROVASCULAR ENDOTHELIAL CELLS OR NEITHER OF THEM?

Aixiang Jiang*, Vanderbilt University
Wei Pan, University of Minnesota
Liming, Milbauer, University of Minnesota
Robert Hebbel, University of Minnesota

Because available microarray data of BOEC (human blood outgrowth endothelial cells), large vessel, and microvascular endothelial cells were from two different platforms, a working cross platform normalization method was needed to make these data comparable. With six HUVEC (human umbilical vein endothelial cells) samples hybridized on two-channel cDNA arrays and six HUVEC samples on Affymetrix arrays, 64 possible combinations of a three step normalization procedure were investigated to search for the best normalization method, which was selected based on two criteria measuring the extent to which expression profiles of biological samples of the same cell type arrayed on two platforms were indistinguishable. Next, three discriminative gene lists between large vessel and microvascular endothelial cells were achieved by SAM (Significant Analysis of Microarrays), PAM (Prediction Analysis for Microarrays), and a combination of SAM and PAM lists. The final discriminative gene list was selected by SVM (support vector machine). Based on this discriminative gene list, SVM classification analysis showed that BOEC cells were far from large vessel cells, they either formed their own class, or fell into microvascular class. Based on all the common genes between the two platforms, SVM analysis further confirmed this conclusion.

email: aixiang.jiang@vanderbilt.edu

A METHOD FOR CGH MICROARRAY DATA ANALYSIS

Wenqing He*, The University of Western Ontario
Ian McLeod, The University of Western Ontario
Shahidul Islam, The University of Western Ontario

Genomic DNA copy number alterations are important features for the development of human diseases. CGH microarray is a powerful technique that enables us to search genome-wide for possible regions with DNA copy number alterations. The DNA copy number may be viewed as a sequence along the whole genome, and the alteration regions correspond to the sequence changes. We propose to use a modified wavelet coefficient approach to identify alteration regions. Simulation studies are conducted to evaluate the performance of the proposed method and a real CGH data set is analyzed.

email: whe@stats.uwo.ca

PRACTICAL ISSUES ASSOCIATED WITH THE ANALYSIS OF ARRAY COMPARATIVE GENOMIC HYBRIDIZATION DATA

Daniel P. Gaile*, University at Buffalo New York Center of Excellence in Bioinformatics and Life Sciences
Jeffrey C. Miecznikowski, University at Buffalo New York Center of Excellence in Bioinformatics and Life Sciences
Lori Shepherd, New York Center of Excellence in Bioinformatics and Life Sciences
Norma J. Nowak, Roswell Park Cancer Institute New York Center of Excellence in Bioinformatics and Life Sciences

Array Comparative Genomic Hybridization (aCGH) is an array-based technology which provides simultaneous and spatially correlated spot assays of relative genetic abundance levels at multiple sites across the genome. In the context of multiple hypothesis testing, the spatial correlation of the assays complicates the question of how best to define a discovery and consequently, how best to estimate the false discovery rate (FDR) corresponding to a given rejection region. We present a quick and convenient method for estimating the (chromosome) Arm-specific False Discovery Rate and present the argument that it may sometimes be more appropriate to estimate the aFDR rather than the FDR associated with the traditional definition of a discovery. We also demonstrate that the perfect within sample correlation of the the centering errors induces a positive bias in spot assay by spot assay correlation estimates while negative across-sample-correlations can induce a negative bias in said estimates. We provide analyses of a real and simulated dataset and demonstrate that, under realistic conditions, that the centering errors can induce biases in spot assay by spot assay correlation estimates while not significantly affecting permutation based thresholds for significance; a combination which can lead to a break down in error control.

email: dpgaile@buffalo.edu

ANALYSIS OF DNA COPY NUMBER VARIATIONS USING PENALIZED LEAST ABSOLUTE

Xiaoli Gao*, University of Iowa
Jian Huang, University of Iowa

Deletions and amplifications of the human genomic DNA copy number are the cause of numerous diseases such as cancer. Therefore, the detection of DNA copy number variations (CNV) is important in understanding the genetic basis of human diseases. Various techniques and platforms have been developed for genome-wide analysis of DNA copy number, such as array-based comparative genomic hybridization (aCGH), SNP arrays, and high-resolution mapping using high-density tiling oligonucleotide arrays. Since complicated biological and experimental processes are involved in these platforms, data can be contaminated by outliers. Inspired by the robustness property of the LAD regression, we propose a penalized LAD regression with the fused lasso penalty for detecting CNV. This method incorporates the spatial dependence and sparsity of CNV into the analysis and is computationally feasible for high-dimensional array-based data. We evaluate the proposed method using simulation studies and demonstrated it on two real data examples.

email: xiaoli-gao@uiowa.edu

SMARTER CLUSTERING METHODS FOR HIGH-THROUGHPUT SNP GENOTYPE CALLING

Yan Lin*, University of Pittsburgh
George C. Tseng, University of Pittsburgh
Lora J.H. Bean, Emory University
Stephanie L. Sherman, Emory University
Eleanor Feingold, University of Pittsburgh

Many high-throughput genotyping technologies for single nucleotide polymorphism (SNP) markers have been developed. Most use clustering methods to 'call' the SNP genotypes, but standard clustering methods are not optimal in distinguishing the genotype clusters of a SNP because they do not take advantage of a number of specific features of the genotype calling problem. In particular, when family data are available, pedigree information is mostly ignored. Furthermore, when prior information about the distribution of the measurements for each cluster is available, choosing an appropriate model-based clustering method can significantly improve the genotype calls. In this paper, we discuss the impact of incorporating external information into clustering algorithms to call the genotypes. We also propose two new methods to call genotypes using family data. The first method is a modification of the K-means method which uses the family information by updating all members from a family together. The second method is a likelihood-based method that combines the Gaussian or beta mixture model with pedigree information. We compare the performance of these two methods and some other existing methods using simulation studies. We also compare the performance of these methods on a real dataset generated by the Sequenom platform (www.sequenom.com).

email: yal14@pitt.edu

60. MULTIPLE TESTING PROCEDURES, ICLUDING GATEKEEPING AND FDR

TREE-STRUCTURED GATEKEEPING PROCEDURES FOR MULTIPLE ENDPOINTS

Ajit C. Tamhane*, Northwestern University
Alex Dmitrienko, Eli Lilly and Company
Brian Wiens, Myogen
Xin Wang, Sanofi-Aventis

Parallel and serial gatekeeping procedures have been recently proposed (Westfall and Krishen 2001, and Dmitrienko, Offen and Westfall 2003) for testing hierarchically ordered families of hypotheses. We generalize these procedures to what we call tree-structured gatekeeping procedures. This generalization is necessary to deal with problems involving hierarchically ordered multiple objectives subject to logical restrictions, e.g., in the analysis of multiple endpoints in dose-control studies and in superiority-equivalence testing. The proposed approach is based on the closure principle of Marcus, Peritz and Gabriel (1976) and uses weighted Bonferroni tests for intersection hypotheses. In special cases of parallel or serial tree structures the closed testing procedure can be shown to be equivalent to stepwise procedures, which are easy to implement. Two illustrative clinical trial examples are given.

email: ajit@iems.northwestern.edu

GATEKEEPING TESTING PROCEDURES BASED ON THE SIMES TEST WITH CLINICAL TRIAL APPLICATIONS

Alex Dmitrienko*, Eli Lilly and Company
Ajit C. Tamhane, Northwestern University
Xing Wang, Sanofi-Aventis

This presentation deals with an extension of multiple testing procedures with a hierarchical structure (known as gatekeeping procedures) that find applications in a wide variety of clinical studies with multiple objectives, including studies with multiple primary/secondary endpoints and dose-finding studies. Existing gatekeeping procedures are based mainly on the Bonferroni test and this talk introduces a family of more powerful gatekeeping procedures derived from the Simes test. This family includes gatekeeping procedures for clinical trials with ordered endpoints and trials with logical restrictions. The talk also discusses the use of resampling methods to improve the power of Simes-based gatekeeping procedures. The performance of the proposed procedures is compared to that of Bonferroni-based procedures via Monte-Carlo simulations.

email: alexei@lilly.com

SIMULTANEOUS INFERENCE FOR MULTIPLE TESTING AND CLUSTERING VIA A DIRICHLET PROCESS MIXTURE MODEL

David B. Dahl*, Texas A&M University
Qianxing Mo, Texas A&M University
Marina Vannucci, Texas A&M University

We propose a Bayesian nonparametric regression model which exploits clustering for increased sensitivity in multiple hypothesis testing. We build on Dahl and Newton (2005) who showed that this was feasible by modeling the dependence among objects through clustering and then estimating hypothesis-testing parameters averaged over clustering uncertainty. We propose several improvements. First, we separate the clustering of the regression coefficients from the part of the model that accommodates for heteroscedasticity. Second, our model accommodates a wider class of experimental designs, such as permitting covariates and not requiring independent sampling. Third, we provide a more satisfactory treatment of nuisance parameters and some hyperparameters. Finally, we do not require the arbitrary designation of a reference treatment. The proposed method is compared in a simulation study to ANOVA and the BEMMA method of Dahl and Newton (2005).

email: dahl@stat.tamu.edu

BONFERRONI-BASED CORRECTION FACTOR FOR MULTIPLE, CORRELATED ENDPOINTS

Rickey E. Carter*, Medical University of South Carolina
Amy E. Herrin, Medical University of South Carolina
Sharon D. Yeatts, Medical University of South Carolina

Use of multiple endpoints and repeated hypothesis testing is common in biomedical research examining efficacy between two groups (e.g., treatment vs. control). For this setting, many correction factors have been developed to control the overall Type I error rate. A common approach is to use a Bonferroni-like correction to ensure the appropriate experiment-wise significance level. For J correlated endpoints, the strict Bonferroni correction is too conservative and sacrifices power, particularly if J is large. In this research, a new Bonferroni-like correction factor, based on assessment of the underlying dimensionality of multiple endpoints, determined using principal components analysis, is proposed. The 'effective dimensionality', g*, is the number of principal components required to explain at least 80% of the variability in the endpoints. In the equi-correlated case where a common intraclass correlation (rho) can be reliably estimated, the correction factor for J endpoints is g*={(J+1) – [1+(J-1)rho]}. In the event the J endpoints are orthogonal, the proposed correction factor reduces to J, the Bonferroni correction factor. Similarly, as the correlation of J endpoints increases, the correction factor decreases to 1.0 (i.e., no inflation to p-values). The methods are demonstrated using a simulation study and are applied to a substance abuse study.

email: carterre@musc.edu

ORACLE AND ADAPTIVE COMPOUND DECISION RULES FOR FALSE DISCOVERY RATE CONTROL

Wenguang Sun*, University of Pennsylvania
Tony Cai, University of Pennsylvania

We develop a compound decision theory framework for multiple testing problems and derive an oracle rule based on the z-values that minimizes the false non-discovery rate (FNR) subject to a constraint on the false discovery rate. It is shown that many commonly used multiple testing procedures, which are p-value based, are inefficient. We introduce an adaptive procedure that asymptotically attains the performance of the z-value oracle procedure and show that it outperforms the traditional approaches based on p-values. We demonstrate our method in an analysis of microarray data from a HIV study that involves testing thousands of hypotheses simultaneously.

email: wsun@cceb.med.upenn.edu

QUICK CALCULATION FOR SAMPLE SIZE WHILE CONTROLLING FALSE DISCOVERY RATE WITH APPLICATION TO MICROARRAY ANALYSIS

Peng Liu*, Iowa State Universsity
J.T. G. Hwang, Cornell University

Sample size estimation is important in microarray or proteomic experiments since biologists can typically afford only a few repetitions. Classical procedures to calculate sample size are based on controlling type I error, e.g., family-wise error rate (FWER). In the context of microarray and other large-scale genomic data, it is more powerful and more reasonable to control false discovery rate (FDR) or positive FDR (pFDR)(Storey03 and Tibshirani, 2003). However, the traditional approach of estimating sample size is no longer applicable to controlling FDR, which has left most practitioners to rely on haphazard guessing. We propose a procedure to calculate sample size while controlling false discovery rate. Two major definitions of the false discovery rate (FDR in Benjamini and Hochberg, 1995, and pFDR in Storey, 2002) vary slightly. Our procedure applies to both definitions. The proposed method is straightforward to apply and requires minimal computation, as illustrated with two sample t-tests and F-tests. We have also demonstrated by simulation that, with the calculated sample size, the desired level of power is achievable by the q-value procedure (Storey, Taylor and Siegmund, 2004) when gene expressions are either independent or dependent.

email: pliu@iastate.edu

ON GENERALIZED FDR

Fei Zou*, University of North Carolina-Chapel Hill
Fred A. Wright, University of North Carolina-Chapel Hill
Jianhua Hu, University of North Carolina-Chapel Hill

FDR has become increasingly attractive in genetic studies where hundreds of thousands of tests are performed simultaneously and traditional Bonferroni correction for multiple comparisons is often extremely conservative. Besides its appealing features, there are several disadvantages of FDR. First, FDR measures the false rejection rate among total rejected events, which is well defined when the parameter space of interests δ can be appropriately partitioned into the null and alternative. However, when the parameter space is continuous and π_0, the proportion of null hypotheses, euqals 0 and therefore, with its current definition, FDR will be 0. Second, FDR involves π_0, which is often unknown in practice. The accuracy of FDR estimate depends on the accuracy of the estimation of π_0. It turns out that estimating π_0 accurately is not an easy task. To overcome these drawbacks, alternatively, we propose a new measure which is analogy to FDR but is more flexible in dealing with continuous parameter space and may be more accurately estimated. Further, the proposed measure has intuitive interpretations appealing to scientists.

email: fzou@bios.unc.edu

61. SPATIAL/TEMPORAL METHODOLOGY AND APPLICATIONS

SPATIALLY ADAPTIVE INTRINSIC GAUSSIAN MARKOV RANDOM FIELDS

Yu Yue*, University of Missouri at Columbia
Paul L. Speckman, University of Missouri at Columbia
Dongchu Sun, University of Missouri at Columbia

We propose a new class of priors for spatially adaptive smoothing splines and thin-plate splines on a lattice by fully Bayesian hierarchical inference. These priors extend intrinsic Gaussian Markov random fields (IGMRFs) by adaptively modeling the precision parameter. We show how the spatially adaptive IGMRFs are effective in nonparametric regression. We also discuss computational issues for Gibbs sampling using sparse matrices and blocking techniques.

email: yytc9@mizzou.edu

DECOMPOSITION OF REGRESSION ESTIMATORS TO EXPLORE THE INFLUENCE OF UNMEASURED CONFOUNDERS IN SPACE AND TIME

Yun Lu*, Johns Hopkins University
Scott L. Zeger, Johns Hopkins University

In environmental epidemiology, exposure X(s,t) and health outcome Y(s,t) vary in space s and time t. We present a method to diagnose the possible influence of unmeasured confounders U(s,t) on the estimated effect of X on Y and to propose an approach to robust estimation. The idea is to use space and time as proxy measures for the unmeasured factors U. We start with the time series case where Xt and Yt are continuous variables at equally-spaced times and assume linear model. We define component regression coefficients b(u)'s that correspond to pairs of observations with specific lag u. Controlling for a smooth function of time, U(t), using a kernel estimator is roughly equivalent to estimating b with a linear combination of the b(u)'s with weights that involve two components: the assumptions about the smoothness of U(t) and the normalized variogram of the X process. If U is a stationary process, independent of X, the b(u)'s are unbiased estimator of b. The excess variation in the distance-specific coefficients is evidence of confounding by U. We use b(u)'s to diagnose the influence of U on the effect of X and Y and use appropriate linear combination of b(u)'s to obtain a more robust estimator. The methods will be illustrated with time series analysis of air pollution and mortality in U.S. cities.

email: ylu@jhsph.edu

HIERARCHICAL MULTIRESOLUTION APPROACHES FOR DENSE POINT-LEVEL BREAST CANCER TREATMENT DATA

Shengde Liang*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Bradley P. Carlin, University of Minnesota

The analysis of point-level (geostatistical) data has historically been plagued by computational difficulties, owing to the high dimension of the nondiagonal spatial covariance matrices that need to be inverted. This problem is greatly compounded in hierarchical Bayesian settings, since these inversions need to take place at every iteration of the associated Markov chain Monte Carlo (MCMC) algorithm. This paper offers an approach for modeling the spatial correlation at two separate scales. This reduces the computational problem to a collection of lower-dimensional inversions that remain feasible within the MCMC framework. We illustrate the importance and applicability of our methods using a collection of dense point-referenced breast cancer data collected over the mostly rural northern part of the state of Minnesota. Substantively, we wish to discover whether women who live more than a 60-mile drive from the nearest radiation treatment facility tend to opt for mastectomy over breast conserving surgery (BCS, or ``lumpectomy'), which is less disfiguring but requires 6 weeks of follow-up radiation therapy. Our hierarchical multiresolution approach resolves this question while still properly accounting for all sources of spatial association in the data.

email: shengdel@biostat.umn.edu

A DISAGGREGATION APPROACH TO BAYESIAN SPATIAL MODELING OF CATEGORICAL DATA

Eric C. Tassone*, Duke University
Alan E. Gelfand, Duke University
Marie L. Miranda, Duke University

We develop novel methodology for spatial modeling of subgroups within areal units. Typical Bayesian mapping models use areal unit counts aggregated over subgroups, with subgroups sometimes partially accounted for via covariates or expected counts. Our approach disaggregates these aggregated counts, using individual-level data to form areal unit subgroup cell counts in a spatially-smoothed, multilevel loglinear model. Advantages include: a richer class of available models; dimension reduction; and not having to set specify a 'response' variable, i.e., not being confined to conditional probability statements. Indeed, joint probability modeling enables inference regarding arbitrary joint, marginal, and conditional probabilities. We illustrate our approach with county-level North Carolina Detailed Birth Record data from 1999–2003, focusing on several categorical variables, including low birth weight (LBW), maternal race, maternal tobacco use, and infant sex. Findings highlight geographic differences across counties in their association and in overall and subgroup-specific rates of LBW and racial disparities in LBW incidence. Simulation results support the benefit of our approach.

email: eric.tassone@duke.edu

SMOOTHING APPROACHES FOR EVALUATING ECOLOGICAL BIAS: AN APPLICATION TO RACIAL DISPARITY IN MORTALITY RATES IN THE U.S. MEDICARE POPULATION

Yijie Zhou*, Johns Hopkins Univerity
Francesca Dominici, Johns Hopkins Univerity
Thomas A. Louis, Johns Hopkins Univerity

The use of aggregated data (e.g., county-level average exposure and total death count) to infer associations at individual-level can result in ecological bias. When the outcome and exposure data vary continuously in space, aggregation based on standard geography may not capture the spatial characteristics of both the exposure and health outcome, and therefore be more subject to bias. We introduce a spatial smoothing method for studying the effects of spatial aggregation on association estimates from a Generalized Linear Model. Both regressors and outcomes are spatially 'aggregated' using a kernel smooth. Ecological bias is evaluated as a function of the kernel bandwidth parameter which measures the degree of aggregation. This approach facilitates flexible investigation of ecologic association as a function of the form and the degree of aggregation. We apply the approach to the study of racial disparity in mortality rates in the U.S. medicare population. Specifically, we investigate how the relation between race and mortality risk depends on the form of the kernel and value of the smoothing parameter (degree of aggregation). Analyses based on disaggregated data at the individual level provide the gold standard from which bias is computed.

email: yijzhou@jhsph.edu

MIXTURES OF POLYA TREES FOR FLEXIBLE SPATIAL FRAILTY SURVIVAL MODELING

Luping Zhao*, University of Minnesota
Timothy E. Hanson, University of Minnesota
Bradley P. Carlin, University of Minnesota

Mixtures of Polya trees offer a very flexible, nonparametric approach for modeling time-to-event data. Many such settings also feature spatial association that requires further sophistication, either at a point (geostatistical) or areal (lattice) level. In this paper we combine these two aspects within three competing survival models, obtaining a data analytic approach that remains computationally feasible in a fully hierarchical Bayesian framework thanks to modern Markov chain Monte Carlo (MCMC) methods. We illustrate the usefulness of our proposed methods with an analysis of spatially oriented breast cancer survival data from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute. Our results indicate appreciable advantages for our approach over previous, competing methods that impose unrealistic parametric assumptions, ignore spatial association, or both.

email: lupingz@biostat.umn.edu

SPATIAL MODELING OF THE RELATIONSHIP BETWEEN AIR QUALITY AND MYOCARDIAL INFARCTION ADJUSTED FOR SOCIO-DEMOGRAPHIC STATUS

Jie Yang*, University of Florida
Linda J. Young, University of Florida
Carol A.G. Crawford, National Center for Environmental Health, CDC
Greg Kearney, Division of Environmental Health, Florida Department of Health
Chris Duclos, Division of Environmental Health, Florida Department of Health

Increasingly studies are conducted to assess the potential relationship between environmental factors or changes in these factors and public health. Here we studied the relationship between air quality and myocardial infarction (MI) for the state of Florida. Daily ozone and pm 2.5 measurements were linked with hospital visit records and income and education levels at the county level. A spatially-varying coefficient model was implemented to allow the relationship between rate of MI and air quality, adjusted for socio-demographic variables , to vary smoothly through space. In this talk we will also discuss the statistical challenges in the spatial modeling process, including change of support problem when combining environmental data, health and socio-economic data spatially, issues of combining environmental data, health and socio-economic data temporally, hypothesis testing, and prediction problems in geographical weighted regression modeling.

email: jyang81@ufl.edu

62. PRESIDENTIAL INVITED ADDRESS

Success, Change, and the Future: The Evolving role of a Statistician in Research and Development

Frank Rockhold, GlaxoSmithKline

email: frank.w.rockhold@gsk.com

63. IMS MEDALLION LECTURE

PREDICTION BY SUPERVISED PRINCIPAL COMPONENTS

Robert Tibshirani*, Stanford University

In regression problems where the number of predictors greatly exceeds the number of observations, conventional regression techniques may produce unsatisfactory results. We describe a technique called "supervised principal components" that can be applied to this type of problem. Supervised principal components is similar to conventional principal components analysis except that it uses a subset of the predictors that are selected based on their association with the outcome. Supervised principal components can be applied to regression and generalized regression problems such as survival analysis. It compares favorably to other techniques for this type of problem, and can also account for the effects of other covariates and help identify which predictor variables are most important. We also provide asymptotic consistency results to help support our empirical findings. These methods could become important tools for DNA microarray data, where they may be used to more accurately diagnose and treat cancer. If time permits, we will also discuss "Pre-conditioning" for consistent feature selection.

This is joint work with Eric Bair, Trevor Hastie and Debashis Paul.

email: tibs@stat.stanford.edu

64. COVARIANCE SELECTION AND VARIANCE COMPONENT TESTING IN MODELING LONGITUDINAL DATA

BAYESIAN VARIABLE SELECTION UNDER VARIOUS MODEL SETTINGS

Mahlet G. Tadesse*, University of Pennsylvania
Sinae Kim, University of Michigan
Naijun Sha, University of Texas El Paso
Marina Vannucci, Texas A&M University

I will start with a brief review of the Bayesian stochastic search variable selection procedure in linear models. I will then present extensions of these variable selection techniques in the context of classification and clustering. In classification, we build the variable selection procedure in a multinomial probit model. In clustering, we search for discriminating variables while uncovering the group structure of the experimental units by specifying mixture models.

email: mtadesse@cceb.med.upenn.edu

LIKELIHOOD RATIO TESTS FOR ZERO VARIANCE IN LINEAR MIXED MODELS

Ciprian M. Crainiceanu*, Johns Hopkins University

We propose and investigate the properties of the Likelihood Ratio Tests (LRTs) for zero variance components in Linear Mixed Models (LMMs). In the case of LMMs with one variance component our finite sample and asymptotic results are different from standard results, which assume that the response vector can be partitioned into a large number of independent sub-vectors. We show that the problem persists in the case of LMMs with more than one variance component and propose an accurate finite sample approximation of the null distribution of the LRTs for testing for zero variance components. Our method is relatively fast and provides corrected p-values for mixed model software such as the lme function in R or S-plus, the MIXED procedure in SAS and the xtmixed function in STATA.

email: ccrainic@jhsph.edu

REGULARIZED ESTIMATION OF COVARIANCE MATRICES FOR LONGITUDINAL DATA THROUGH SHRINKAGE AND SMOOTHING

Jianhua Huang*, Texas A&M University
Linxu Liu, Columbia University

We propose a penalized likelihood method for producing statistically efficient estimator of a covariance matrix for longitudinal data. The approach parameterizes the covariance matrix through the modified Cholesky decomposition of its inverse. For longitudinal data, the entries of the lower triangular and the diagonal matrix associated with the modified Cholesky decomposition have meaning as regression coefficients and prediction variances when regressing a measurement on its predecessors. We develop a regularization scheme for covariance matrix estimation based on two observations: First, the lower triangular is likely to have many off-diagonal elements that are zero or close to zero; Second, there is usually some kind of continuity among neighboring elements in the lower triangular. We employ an L_1 or L_2 penalty for shrinkage and selection and employ a roughness penalty to impose smoothness in the rows or subdiagonals of the lower triangular. It is demonstrated through simulations that combination of shrinkage and smoothing does better and sometimes much better than using shrinkage or smoothing alone. The proposed covariance matrix estimator is applied to efficient estimation of regression coefficients for longitudinal data and to best linear forecast with a high-dimensional predictor vector.

email: jianhua@stat.tamu.edu

65. DIAGNOSTIC MEDICAL IMAGING

THRESHOLDS ON TEST POSITIVITY AND REFERENCE STANDARD: EFFECTS ON TEST PREDICTIVE VALUE AND ACCURACY

Constantine Gatsonis*, Brown University
Shang Ying Shiu, Brown University and Academia Sinica

The effect of the threshold for test positivity on test sensitivity and specificity has been thoroughly investigated in recent years, as reported in the extensive literature on ROC analysis. However, considerably less attention has been given to the study of the effect of this threshold on the positive and negative predictive values of a test. The interplay between these two measures of predictive performance turns out to be more complex and far less amenable to succinct mathematical characterization than the corresponding interplay between sensitivity and specificity. In the first part of this presentation we will discuss recent work leading to the formulation of the predictive analogue of an ROC curve. In the second part of the presentation we will consider a different extension of the binary test/binary disease model, in which the binary disease status is defined by applying a threshold on a measure of (possible) disease. In this setting, we will examine the effect of the disease threshold on diagnostic performance using a semi-parametric model linking disease status to test result.

email: gatsonis@stat.brown.edu

EVALUATING THE POTENTIAL OF NOVEL PHASE II IMAGING ENDPOINTS

Lori E. Dodd*, National Cancer Institute

New imaging technologies offer the potential to assess response to therapy earlier and more accurately than current methods. In the context of phase II testing of therapies in cancer, for example, some have suggested replacing traditionally used endpoints with newer imaging endpoints. Phase II trials establish preliminary efficacy data quickly in order to screen out less promising drugs to be evaluated in more definitive phase III trials. Most phase II endpoints in cancer research have been based on the assumption that a reduction in tumor size was a necessary, but not necessarily sufficient, condition for achieving a survival benefit. Functional imaging techniques measure other features of a tumor and may more accurately depict the tumor's response to therapy. However, there has been little attention given to what is needed to demonstrate their role as a good phase II endpoint. Clearly, a better phase II endpoint will be evaluated earlier and/or will more reliably predict whether an agent will be successful in a phase III trial. Predictive values and diagnostic accuracy (sensitivity/specificity) have a role here, although each has potential problems. In this talk, I will discuss these problems and what might be needed to establish a new method as a good phase II endpoint.

email: doddl@mail.nih.gov

ASSESSING COMPUTER-AIDED DIAGNOSIS ALGORITHMS: ALGORITHM ARCHITECTURE, STATISTICAL TOOLS AND STUDY DESIGNS

Nicholas Petrick*, U.S. Food and Drug Administration
Brandon D. Gallas, U.S. Food and Drug Administration
Frank W. Samuelson, U.S. Food and Drug Administration
Sophie Paquerault, U.S. Food and Drug Administration

Computer-aided diagnosis (CAD) algorithms are quickly becoming integral aids to the clinician across medicine. However, the clinical utility of these new computer tools, and indeed even the utility of some novel imaging systems, relies primarily on the ability of the clinician to interact correctly with these software tools. This human/computer interaction complicates device evaluation because reader skill and reader variability are integral to the evaluation process and utility of the tools. In this presentation, we will provide a brief introduction to some of the more common types of computer-aided detection and characterization software devices available in the medical imaging field. We will introduce common pre-clinical and clinical study designs, and the statistical tools available for assessing device performance. This will include a discussion of both receiver operating characteristic (ROC) and free-response ROC methodologies, and address methods to account for both case and reader variability in common clinical study paradigms.

email: nicholas.petrick@fda.hhs.gov

OVERVIEW OF CONTRAST AGENTS AND SPECIFIC STATISTICAL CHALLENGES FOR MEDICAL IMAGING

Kohkan Shamsi, Symbiotic Pharma Research
Suming W. Chang*, Berlex Inc.
Joerg Kaufmann Schering AG

Various types of contrast agents are used that complement and overcome the limitations of imaging devices like MRI and CT. Complementary development of machines and contrast agents leads to optimization of images and help in diagnosis and management of patients. An overview of current and new contrast agents will be addressed. The different types of contrast agents with different targeted applications will be presented. Issues specific to development of contrast agents will also be discussed. The choice of one or more diagnostic indices for a contrast agent remains a challenging topic. The accuracy is the most widely used performance index. The sensitivity and specificity play an important role in demonstrating the efficacy of a contrast agent in diseased and non-diseased cases. But, the claims based solely on sensitivity and specificity may not be accepted in the absence of a significant effect of overall accuracy. We propose to use a weighted accuracy along with simultaneous comparisons of sensitivity and specificity for a primary endpoint. The problem of adjusting for multiple comparisons, the issues with markedly unbalanced subgroups, and the choices of different weighting schemes will also be discussed.

email: suming_chang@berlex.com

66. NEW STRATEGIES IN DESIGNING COMBINED-PHASE CLINICAL TRIALS

A BAYESIAN APPROACH TO JOINTLY MODELING TOXICITY AND BIOMARKER EXPRESSION IN A PHASEI/II DOSE-FINDING TRIAL

B. Nebiyou Bekele*, The University of Texas M. D. Anderson Cancer Center
Yu Shen, The University of Texas M. D. Anderson Cancer Center

In this presentation we propose a Bayesian approach to phase I/II dose-finding oncology trials by jointly modeling a binary toxicity outcome and a continuous biomarker expression outcome. We apply our method to a clinical trial of a new gene therapy for bladder cancer patients. In this trial, biomarker expression indicates that the therapy is biologically active. For ethical reasons, the trial is conducted sequentially, with the dose for each successive patient chosen using both toxicity and activity data from patients previously treated in the trial. The modeling framework that we use naturally incorporates correlation between the binary toxicity and continuous activity outcome via a latent Gaussian variable. The dose-escalation/de-escalation decision rules are based on the posterior distributions of both toxicity and activity. A flexible state-space model is used to relate the activity outcome and dose. Extensive simulation studies show that the design reliably chooses the preferred dose using both toxicity and expression outcomes under various clinical scenarios.

email: bbekele@mdanderson.org

PRACTICAL RECOMMENDATIONS FOR PHASE I/II DESIGNS

Rick Chappell*, University of Wisconsin-Madison
Ken YK Cheung, Columbia University

In the conventional paradigm of developing a new cancer treatment, Phase I trials are safety studies which estimate the maximum tolerated dose based on dose-limiting clinical toxicities, whereas phase II trials examine the potential efficacy of the new therapy. Although this paradigm allows each trial to have a clear and achievable objective, the phase I/II distinction is artificial because there is always interest in looking at efficacy data collected from the former and evaluating safety in the latter. Thus, there is scientific rationale for conducting a combined phase I/II trial in which both toxicity and efficacy outcomes are considered as primary endpoints. This also reduces administrative burden. A typical combined phase I/II design performs dose escalation in a phase I portion to determine the maximum tolerated dose, whose efficacy is evaluated in an expanded cohort of patients in a subsequent phase II portion of the trial. Our objective is two-fold. We first discuss the advantages and limitations regarding this design, make recommendations to mend its limitations, and review alternative designs in the statistical literature. We then give recommendations on the analysis of safety and efficacy outcomes associated with a combined phase I/II trial, and illustrate the approach with two vaccine trials.

email: chappell@stat.wisc.edu

SEAMLESS PHASE II & III: SMALLER, STRONGER, QUICKER

Scott M. Berry*, Berry Consultants
Donald A. Berry, M.D. Anderson Cancer Center

We describe adaptive Phase II dose finding studies that morph in to phase III confirmatory studies. The phase II design results in better information, in a shorter time, with fewer subjects. The phase II design is based entirely on the goals of the phase III design including endpoints and adverse events. Several example trials will de discussed and the challenges of each discussed.

email: scott@berryconsultants.com

A BAYESIAN DECISION-THEORETIC APPROACH BASED ADAPTIVE DESIGNS FOR CLINICAL TRIALS

Yi Cheng*, Indiana University-South Bend
Yu Shen, M.D. Anderson Cancer Center

A Bayesian adaptive design is proposed for a comparative two-armed clinical trial using decision-theoretic approaches. A loss function is specified, based on the cost for each patient, and the costs of making incorrect decisions at the end of a trial. At each interim analysis, the decision to terminate or to continue the trial is based on the expected loss function while concurrently incorporating efficacy, futility and cost. The maximum number of interim analyses is determined adaptively by the observed data. We derive explicit connections between the loss function and the frequentist error rates, so that the desired frequentist properties can be maintained for regulatory settings. The operating characteristics of the design can be evaluated on frequentist grounds.

email: ycheng@iusb.edu

67. RECENT INNOVATIONS IN DYNAMIC TREATMENT REGIMES

AN INTRODUCTION TO DYNAMIC TREATMENT REGIMES

Marie Davidian*, North Carolina State University

A ‘dynamic treatment regime’ or ‘adaptive treatment strategy’ is a set of rules that together dictate how to make sequential decisions on treatment of a patient over time. Each rule corresponding to a particular time at which a decision is to be made on changing, modifying, augmenting, stopping, or starting treatment takes as input information on the patient up to that point and, based on this information, outputs the next treatment action. Thus, dynamic treatment regimes are algorithms that allow treatment decisions to be ‘individualized’ through a principled, evidenced-based set of rules that attempt to operationalize the way clinicians manage patients in practice. In this talk, we provide an introduction to thinking about dynamic treatment regimes and use the setting of evaluation of courses of cancer therapy to provide a concrete but simple example of how inference may be made on dynamic treatment regimes based on suitable clinical trials. This talk will provide background for the subsequent talks in this session and is meant to be accessible to statisticians with no prior exposure to dynamic treatment regimes.

email: davidian@stat.ncsu.edu

ESTIMATING MEAN RESPONSE AS A FUNCTION OF TREATMENT DURATION IN AN OBSERVATIONAL STUDY, WHERE DURATION MAY BE INFORMATIVELY CENSORED

Anastasios A. Tsiatis*, North Carolina State University

In a recent clinical trial ‘ESPRIT’ of patients with coronary heart disease who were scheduled to undergo percutaneous coronary intervention (PCI), patients randomized to receive Integrilin therapy had significantly better outcomes than patients randomized to placebo. The protocol recommended that Integrilin be given as a continuous infusion for 18--24 hours. There was debate among the clinicians on the optimal infusion duration in this 18--24-hour range, and we were asked to study this question statistically. Two issues complicated this analysis: (i) The choice of treatment duration was left to the discretion of the physician and (ii) treatment duration must be terminated (censored) if the patient experienced serious complications during the infusion period. To formalize the question, ‘What is the optimal infusion duration?’ in terms of a statistical model, we developed a framework where the problem was cast using ideas developed for adaptive treatment strategies in causal inference. The problem is defined through parameters of the distribution of (unobserved) potential outcomes. We then show how, under some reasonable assumptions, these parameters could be estimated. The methods are illustrated using the data from the ESPRIT trial.

email: tsiatis@stat.ncsu.edu

HYPOTHESIS TESTING & DYNAMIC TREATMENT REGIMES

Susan Murphy*, University of Michigan
Lacey Gunter, University of Michigan
Bibhas Chakraborty, University of Michigan

Dynamic treatment regimes are individually tailored treatments that mimic the adaptive nature of clinical practice. These regimes repeatedly adapt treatment to the patient based on his/her outcomes. We discuss the construction of dynamic treatment regimes from data in which patients are randomized repeatedly at a series of "critical" decision points. An important methodological difficulty concerns measures of confidence necessary for conducting hypothesis tests; in general the bootstrap and usual asymptotic arguments do not provide valid hypothesis tests. We provide an alternate approach and illustrate this using data on patients with treatment resistant depression.

email: samurphy@umich.edu

68. STATISTICAL MODELING IN ECOLOGY

MODELLING NON-LINEAR MIXTURE EXPERIMENTS IN STREAM ECOLOGY

Mary C. Christman*, University of Florida
Christopher Swan, University of Maryland

In stream ecology, the effects of mixtures of leaves from different tree species on the bacterial and insect ecosystem in the stream is of major interest due to rapid change in land use near waterways in recent years. A primary research area is the interaction of the leaf species composition and shredders (insects that physically shred the leaf) on the rate of decay of leaves in the stream. We developed two competing models of the effect of the composition of leaf species and of shredders on decay rate. The first model is a non-linear model ("multi-exponential" model) in which main effects and interaction of mixtures and shredders are incorporated as a linear combination of exponential and offset terms. The error terms are assumed to be independent normal variates with mean zero and constant variance. The second model is a log-linear model with multiplicative error in which the response variable is transformed using natural logarithms so that the resulting model is intrinsically linear. Here the error terms have unequal variances that depend on the number of days since start of the experiment. The analysis proceeds as in regular mixture experiments with models that are linear in the parameters. We compare the statistical behavior of the two models using simulation to determine the robustness of the models to failure to identify the true state before analysis.

email: mcxman@ufl.edu

ESTIMATING RARE BUTTERFLY ABUNDANCE BY COMBINING MULTIPLE DATA TYPES WITH SIMPLE POPULATION MODELS

Kevin Gross*, North Carolina State University
Nick M. Haddad, North Carolina State University
Brian R. Hudgens, Institute for Wildlife Studies

Many butterfly populations are monitored by regular surveys in which observers count the number of individuals seen while walking established transects. For species with non-overlapping generations, each adult flight period yields a time series of counts. Current methods combine these transect count data with models of the processes governing butterfly dynamics to estimate an index of total butterfly abundance. Here, we modify these methods to accommodate both stochastic population dynamics and data from other sources that provide information about detectability. These modifications produce an absolute measure of abundance instead of an index, as well as more accurate measures of the uncertainty in the estimated abundance. We illustrate our method with monitoring data for St. Francis satyr (Neonympha mitchelli), a rare butterfly occurring in the sandhills region of south-central North Carolina, USA.

email: gross@stat.ncsu.edu

INVASIONS, EPIDEMICS, AND BINARY DATA IN A CELLULAR WORLD

Mevin B. Hooten*, Utah State University
Christopher K. Wikle, University of Missouri

Statistical models for invasions and other propagating phenomena (e.g., epidemics) on partitioned domains are often desired in situations where only binary data are available. Here, the dynamics of a binary spatio-temporal process are estimated in terms of probabilistic cellular automata. A simple set of probabilistic neighborhood-based transition rules allows for very flexible space-time behavior including anisotropy and non-stationarity. Additionally, a model is proposed that utilizes covariates to create an environmental preference field, throughout which the phenomenon will propagate from areas of low preference to areas of high preference. An attraction model, motivated by a partial differential equation in the form of a set of directional gradient fields, transforms the covariate-based preference field into parameters governing the direction and magnitude of dispersal.

email: mevin.hooten@usu.edu

HIERARCHICAL MODELING AND INFERENCE IN METACOMMUNITY SYSTEMS

Jeffrey A. Royle*, USGS Patuxent Wildlife Research Center

Inference about animal community structure is typically based on surveys that yield observations of species presence or absence on a sample of spatial units. A critical consideration in conducting inference about community structure is that species are detected imperfectly. That is, a species may be present but go undetected during sampling. This leads to fewer species observed in the sample than actually exist in the community, and biases summaries of community structure and dynamics. This consideration has led to the development of methods for modeling animal community structure that are analogous to classical capture/recapture models. While this approach has led to much progress in understanding the structure and function of animal communities, it does not preserve species identity among spatial samples, and thus is limiting from the perspective of developing predictive models of community composition. We describe a modeling strategy based on species-specific models of occurrence, from which estimates of important summaries of community structure are derived by aggregating indicators of occurrence for all species observed in the sample, and for the estimated complement of unobserved species. We use a data augmentation approach for Bayesian analysis of the model. Examples are given.

email: aroyle@usgs.gov

69. MICROARRAY ANALYSIS II

EVALUATING THE POSITION EFFECT OF SINGLE-BASE-PAIR MISMATCH PROBES IN AFFYMETRIX GENECHIP®

Fenghai Duan*, University of Nebraska Medical Center

The advancement of the microarray technology has greatly increased the development of biomedical research and health sciences. Of varieties of microarray platforms, the high-density short-oligonucleotide Affymetrix GeneChip® is becoming a standard tool in the research of molecular biology. There are lots of arguments about the way to incorporate the information of mismatch probes into the analysis. Very few studies have been conducted to examine the effect of different mismatch positions on the signal estimates from a probe-level base, particularly for the mismatch probes on the Affymetrix GeneChip®. By establishing a cross-hybridization system and compiling a set of probe pairs of perfect match (PM) and mismatch (MM) with the single base mismatch occurring in each of 25 positions, we were able to evaluate the effect of different mismatched positions on various aspects of the microarray experiments using the technology of Affymetrix GeneChip®.

email: fduan@unmc.edu

ENHANCED QUANTILE APPROACH FOR ASSESSING DIFFERENTIAL GENE EXPRESSION

Huixia Wang*, North Carolina State University
Xuming He, University of Illinois at Urbana-Champaign

Due to the small number of replicates in typical gene microarray experiments, the power of statistical inference is often unsatisfactory without information-sharing across genes. In this paper, we propose an enhanced quantile rank score test (EQRS) for detecting differential expression in GeneChip studies by analyzing the quantiles of gene intensity distributions through probe level measurements. A measure of sign correlation, δ, plays an important role in the rank score tests. By sharing information across genes, we develop a calibrated estimate of δ, which reduces the variability at small sample sizes. We compare the EQRS test with three other approaches for determining differential expression: the gene-specific quantile rank score test, the quantile rank score test assuming a common δ, and a modified t-test using summarized probe set level intensities (SAM). The proposed EQRS is shown to be favorable in our simulation study. In addition, we illustrate the value of the proposed approach using a GeneChip study comparing gene expression in the livers of mice exposed to chronic intermittent hypoxia versus those exposed to intermittent room air.

email: wang@stat.ncsu.edu

META-ANALYSIS FOR MICROARRAY EXPERIMENTS

Shiju Zhang*, University of Alabama at Birmingham
Grier P. Page, University of Alabama at Birmingham

Expression microarrays are a powerful technology for scanning tens of thousands of genes simultaneously. They are useful in tumor classification, biological pathway modeling, and QTL detection. But there are also challenges in the use of expression microarray. The arrays are subject to a variety of sources of variation including biological and technical. The number of replicates (sample size) is quite small, and small relative to the number of genes. In order to generate reproducible results, microarray experiments must be carefully designed and executed. In order to overcome these issues it may be useful to pool data or synthesize results from different microarray experiments through a meta-analysis. In a meta-analysis, common genes, networks or patterns among studies can be preserved, while patterns occurring only in some of the experiments may be discovered. We review various meta-analysis methods that have been developed for microarray studies, including model-based and model-free methods such as ANOVA, Fisher~{!/~}s P-value, and Rank Product. We present new methods based upon mixture models and permutation approaches which are compared with existing ones via simulation and real data.

email: shiju.zhang@gmail.com

MINIMAX ESTIMATION OF MEANS WITH APPLICATION TO MICROARRAY EXPERIMENTS

Tiejun Tong*, Yale University
Liang Chen, University of Southern California
Hongyu Zhao, Yale University

The development of microarray technology has revolutionized biomedical research, and microarrays have become a standard tool in biological studies. Due to the cost and/or experimental difficulties, it is common that thousands of genes are measured only with a small number of replications. In particular, the standard gene-specific estimators for means and variances are not reliable and the corresponding tests have low power and/or high false positive rate. Recently, various approaches to improving the standard gene-specific t-test or F-test have emerged by improving the variance estimation. However, little attention has been paid to improving the gene-specific mean estimation. In this paper we study shrinkage estimators for means with unknown variances. We prove that our proposed estimator is minimax under the quadratic loss function. We conduct simulations to evaluate the performance of our proposed shrinkage estimator and compare it with some existing estimators, and show that our estimator is robust in practice. Finally, we construct a shrinkage-based t-like statistic to detect differentially expressed genes and evaluate its performance through simulations and a case study.

email: tiejun.tong@yale.edu

INTEGRATING QUANTITATIVE INFORMATION FROM CHIP-CHIP EXPERIMENTS INTO MOTIF FINDING

Heejung Shim*, University of Wisconsin-Madison
Sunduz Keles, University of Wisconsin-Madison

Identifying binding locations of transcription factors within long segments of non-coding DNA is a challenging task. Recent ChIP-chip experiments utilizing tiling arrays are especially promising for this task since they provide high resolution genome-wide maps of the interactions between the transcription factors and DNA. A two step paradigm is commonly used for performing motif searches based on ChIP-chip data. First, candidate bound sequences that are in the order or 500-1000 base pairs are identified from ChIP-chip data. Then, motif searches are performed among these sequences. These two steps are typically carried out in a disconnected fashion in the sense that the quantitative nature of the ChIP-chip information is ignored in the second step. We develop a conditional two component mixture model (CTCM) which adaptively intergrates ChIP-chip information into motif finding. The performance of the new and existing methods are compared using simulated data and ChIP-chip data from recently available ENCODE studies. These studies indicate that CTCM efficiently utilizes the information available in the ChIP-chip experiments and has superior sensitivity and specificity when the motif of interest has low abundance among the ChIP-chip bound regions and/or has low information content.

email: shim@stat.wisc.edu

BAYESIAN CHANGE POINT ANALYSIS OF GENETIC INSTABILITY IN HIGH-THROUGHPUT EXPERIMENTS

David L. Gold*, Texas A&M University
Choa Sima, Texas A&M University
Daniel P. Gaile, The State University of New York
Norma Nowak, Roswell Park Cancer Institute
Bani Mallick, Texas A&M University

Genetic instability has been linked to diverse forms of cancer and is the topic of ongoing cancer research. One difficulty in modeling high-throughput human chromosomal data is that genetic instability can be subject specific and therefore disease populations are heterogeneous. We investigate Bayesian Change Point Analysis (BCPA) for borrowing strength and accounting for change point uncertainty in high-throughput experiments from heterogeneous patient populations. BCPA shows an increase in power for detecting copy number changes across heterogeneous disease populations, in simulation and applied to Wilms Tumor BAC arrays.

email: dlgold@stat.tamu.edu

DYNAMIC NETWORK ANALYSIS OF TIME COURSE MICROARRAY EXPERIMENTS

Donatello Telesca*, University of Washington
Lurdes YT Inoue, University of Washington

Time-course gene expression data consist of RNA expression from a common set of genes, collected at different time points. Such time series are usually thought to describe some underlying biological processes developing over time. Gene expression profiles over time are modeled as a hierarchical random functional transformation of a reference curve. We propose measures of functional similarity and time order based on the estimated warping functions. This allows for novel inferences on genetic networks which takes full account of the timing structure of functional features associated with the gene expression profiles. We discuss the application of our model to simulated and time-course microarray data arising from animal models on prostate cancer progression.

email: telesca@stat.washington.edu

70. FUNCTIONAL DATA ANALYSIS

CARDS: CLASSIFICATION AFTER REDUCTION OF DIMENSION WITH SCAD PENALTY

Woncheol Jang*, University of Georgia
Jeonyoun Ahn, University of Georgia
Cheolwoo Park, University of Georgia

With the power of modern technology, it is increasingly common to collect functional data in scientific studies that are beyond the capabilities of traditional statistical methods. Here data are considered as functional if data are regularly measured on a fine grid. Dimension reduction or feature selection is a key issue to make statistical methods most effective in functional data such as profiles and curves and estimating profiles can be considered as making inferences for infinite dimensional objects. We develop a new classification method for functional data which we shall call CARDS, standing for Classification After Reduction of Dimension with Smoothed clipped absolute deviation penalty. This proposed method is novel because it can be used for both prediction and feature selection by achieving sparsity at each step. More specifically, we want to keep as many nonzero coefficients as possible in the function estimation step yet with a spare representation and achieving sparsity in the classification step for the classification purpose. Examples in proteomics will be presented.

email: jang@uga.edu

REGULARIZED FUNCTIONAL LINEAR MODELS

Qi Long*, Emory University
Ming Yuan, Georgia Institute of Technology

We consider analysis of data when some predictor (s) is of functional form, for example, an observed profile over a spatially measured biomarker. Generlized functional linear models have been proposed to study the effect of functional predictors on various types outcomes (Muller, 2004; James 2002). In particulary, James (2002) proposed to use basis functions to nonparametrically model functional predictor(s) and use ML approach to obtain parameter estimates defining its effect on outcomes of interest. We propose a different approach. We propose to use smooth splines to model functional predictor(s) and therefore a regularized objective function to estimate its effect on outcomes of interest. We illustrate our methods using a dataset from a cancer epidemiology study, which was conducted to study various biomarkers and their prognostic values in early detection of colon cancer.

email: qlong@sph.emory.edu

MULTISCALE ANALYSIS ON FMRI DATA

Cheolwoo Park*, University of Georgia

We conduct a thorough investigation of the null hypothesis distribution for functional magnetic resonance imaging (fMRI) data using multiscale analysis. Current approaches to the analysis of fMRI data assume temporal independence, or, at best, simple models for temporal (short term or long term dependence) structure. Such simplifications are to some extent necessary due to the complex, high dimensional nature of the data, but to date there has been no systematic study of the temporal structures under a range of possible null hypotheses, using data sets gathered specifically for that purpose. We aim to address these shortcomings by analyzing fMRI data with a long enough time horizon to study possible long range temporal dependence, and by analyzing those data via multiscale methods, SiZer (Significance of Zero Crossings of the Derivative) and wavelets.

email: cpark@stat.uga.edu

DIAGNOSTICS FOR NONLINEAR DYNAMICS

Giles Hooker*, Cornell University

Systems governed by deterministic differential equations have traditionally received little attention in statistics. Current methods for developing systems of differential equations rely either on a priori knowledge or on entirely nonparametric methods. This talk presents a suite of diagnostic tools for analyzing lack of fit in already-identified models. These tools are based on the estimation of unobserved forces for the proposed differential equation and I show how these may be used to identify the presence of unobserved dynamical and forcing components, and the mis-specification of model terms.

e-mail: gjh27@cornell.edu

SMOOTHING PARAMETER SELECTION IN PENALIZED SIGNAL REGRESSION

Philip T. Reiss*, New York University

Studies in biomedical and other disciplines often produce data in the form of functions, i.e. curves or images, and associated scalar outcomes that can be regressed on those functions to produce an estimated coefficient function. Such functional linear models can be fitted by restricting the coefficient function estimate to the span of a B-spline basis, and minimizing a criterion defined as the sum of squared errors plus a roughness penalty. Borrowing ideas from the nonparametric regression literature, we present novel methods for improved selection of the parameter controlling the strength of the roughness penalty. These methods are illustrated using a psychiatric data set.

e-mail: reissp01@med.nyu.edu

NONPARAMETRIC ESTIMATION FOR RELATING DOSE DISTRIBUTIONS TO OUTCOME MEASURES

Matthew J Schipper*, University of Michigan
Jeremy MG Taylor, University of Michigan

Normal tissue complications are a common side effect of radiation therapy. They are the consequence of the dose of radiation received by the normal tissue (usually a gland or organ) near to the tumor. The dose received is typically not uniform, and can be represented as a distribution function. Dose-damage-injury models relate this dose distribution to the observed injury/complication in two steps. The first step relates dose distribution to an unobserved scalar called damage, and the second step relates damage to the observed injury. We present a model in which both of these relations are modeled nonparametrically. Regarding the density of the dose distribution as a curve, a summary measure of damage is obtained by integrating a weighting function of dose (W(d)) over the dose density. Similar to a generalized additive model, the linear predictor in our model includes a nonparametric function of damage, H(damage) which relates damage to injury. Both W(.) and H(.) are written as regression splines. For biological reasons, both nonparametric functions should be monotonic. We discuss identifiability conditions and illustrate our method with data from a head and neck cancer study in which the irradiation of the parotid gland results in loss of saliva flow.

e-mail: mjschipp@med.umich.edu

ESTIMATING COEFFICIENTS AND LINEARLY INDEPENDENT SOLUTIONS OF A LINEAR DIFFERENTIAL OPERATOR WITH COVARIATES FOR FUNCTIONAL DATA

Seoweon Jin*, The University of Texas at El Paso
Joan G. Staniswalis, The University of Texas at El Paso

It is often believed that the curve data correspond to a physical process described well by a differential equation, that is, each data curve is a sum of a random error term and a smooth curve in the null space of a linear differential operator. Ramsay(1996) first proposed the method of regularized principal differential analysis for fitting a differential equation to a collection of noisy data curves. Once the coefficients of the linear differential operator are estimated, a basis for the null space is computed using iterative methods from numerical analysis for solving linear differential equations. Ultimately, a smooth low dimensional approximation to the data curves is obtained by regressing the data curves on the null space basis. We extend principal differential analysis to allow for the coefficients in the linear differential equation to smoothly depend upon covariates. The estimating equations for coefficients in the principal differential analysis with covariates are derived; these are implemented in Splus. The estimators are studied on simulated data, prior to analyzing the evoked potential curves in a study of cochlear implants.

e-mail: seoweonjin@hotmail.com

71. LATENT VARIABLE APPLICATIONS, INCLUDING STRUCTURAL EQUATIONS AND FACTOR ANALYSIS

SEMIPARAMETRIC RELATIONSHIP AMONG LATENT VARIABLES IN THE STRUCTURAL EQUATION MODELS

Bo Ma*, University of South Carolina
Andrew B. Lawson, University of South Carolina

Structural equation models (SEMs) are commonly assumed to have parametric relationships among latent variables. It is well-known that the parametric framework often fails to uncover the true relationship, especially for the latent variables whose values could not be observed directly. To improve the ability to uncover the relationship that suggested from the data, we have developed Bayesian SEMs in which the semiparametric regression was introduced to model the relationship between latent variables. We have chosen low-rank thin-plate splines due to the computation reduction enabled by the low-rank smoothers and the smaller posterior correlation enabled by the thin-plate splines. Furthermore, penalized splines models yield the best linear predictor in the linear mixed model, which is easy to implement in the Markov Chain Monte Carlo method. A simulation study is presented. Two scenarios of data with a quadratic and a threshold relationship between latent variables respectively were simulated, each having two normally distributed latent variables and two normally distributed continuous indicators for each latent variable. Our proposed model framework could uncover the true relationship as well as the true model. Finally, we apply the proposed modeling framework to a real life example.

email: mab@gwm.sc.edu

RESIDUAL-BASED DIAGNOSTICS FOR STRUCTURAL EQUATION MODELS

Brisa N. Sanchez*, University of Michigan
E. Andres Houseman, Harvard School of Public Health
Louise M. Ryan, Harvard School of Public Health

We prosose residual-based diagnostics for structural equation models that assess distributional and linearity assumptions. Recent developments in diagnostic methodology for linear mixed models provide a theoretical framework for the tests proposed herein. We adapt such methods to the structural equations framework, and evaluate some properties of those tests not previously explored given the added complexity of structural equation models in contrast to linear mixed models. Further, we demonstrate the use of these residuals for outlier detection and for assessing the need to include additional covariates in the model. We demonstrate the methods by applying them to the study of in-utero lead exposure.

email: brisa@umich.edu

APPLICATION OF COVARIANCE SHRINKAGE TO FACTOR ANALYSIS

Sock-Cheng Lewin-Koh*, Eli Lilly and Company
Nicholas J. Lewin-Koh, Eli Lilly and Company

Factor analysis is widely used as a multivariate tool to model high dimensional data with a smaller set of underlying, unobservable variables. Maximum likelihood is a common estimation approach in factor analysis modeling. Based on the idea of covariance shrinkage, we propose a modified covariance in the estimation of the factor analysis model, with the goal of improving the estimation accuracy. Some results from a simulation study will be presented.

email: sockcheng@lilly.com

L ^ 2-BASED HOMOGENEITY TESTS FOR MIXTURES WITH STRUCTURAL PARAMETERS

Hongying Dai*, Columbus State University
Richard Charnigo, University of Kentucky

One of the obstacles for testing a homogeneous model versus a finite component mixture model is that the mixture distribution is not generally identifiable. How to ascertain the number of components in a mixture distribution from a standard parametric family with a single parameter has been thoroughly studied in numerous papers, including the Chen, Chen, and Kalbfleisch (2001) work on a modified likelihood ratio test and the Charnigo and Sun (2004) work on an L ^ 2-based homogeneity test called the D-test. However, a mixture distribution from a parametric family with both a parameter of central interest (e.g., a location parameter) and a structural parameter (e.g., a scale parameter) is often a more realistic model. In this work, we extend the D-test to such situations and characterize some asymptotic properties of the D-test statistic. Simulation studies examine the performance of the D-test versus likelihood-ratio-based competitors.

email: dai_hongying@colstate.edu

SUPERVISED BAYESIAN LATENT CLASS MODELS FOR HIGH-DIMENSIONAL DATA

Stacia M. DeSantis*, Harvard University
E. Andres Houseman, Harvard University
Brent A. Coull, Harvard University
Rebecca A. Betensky, Harvard University

High grade gliomas, the most common primary brain tumors in adults, are diagnosed using a large number of immunohistological variables. One goal of diagnosis is survival prognosis, yet clinical subsets based on these variables alone may not correlatewell with survival. We propose two penalized supervised latent class models for high-dimensional binary data, which are estimated using Bayesian MCMC techniques. Penalization and model selection are easily incorporated in this setting by including prior distributions on the unknown parameters. In simulations, these new methods provide parameter estimates under conditions in which standard supervised latent class models break down. Resulting latent classes correlate well with survival. We illustrate our new methodologies in a study of glioma, for which identifiable parameter estimates cannot be obtained without penalization. With penalization, the resulting latent classes not only correlate very well with clinical tumor grade but offer additional information on survival prognosis that may not be captured by clinical diagnosis alone.

email: sdesanti@hsph.harvard.edu

BAYESIAN MODELING OF EMBRYONIC GROWTH USING LATENT VARIABLES

James C. Slaughter*, University of North Carolina at Chapel Hill
Amy H. Herring, University of North Carolina at Chapel Hill
Katherine E. Hartmann, Vanderbilt University Medical Center

In a growth model, individuals move progressively through a series of states where each state is indicative of their developmental status. Interest lies in estimating the rate of progression through each state while incorporating covariates that might affect the transition rates. We develop a Bayesian discrete time multistate growth model for inference from cross-sectional data with unknown initiation times. For each subject, data are collected at only one time point at which we observe the state as well as covariates that measure developmental progress. We link the developmental progress variables to an underlying latent growth variable that can affect the transition rates. We also examine the association between latent growth and the probability of future events. We use a Markov chain Monte Carlo algorithm for posterior computation and apply our methods to a novel study of embryonic growth and pregnancy loss in which we were able to find evidence in favor of a previously hypothesized but unproven association between slow growth early in pregnancy and increased risk of future loss.

email: jslaught@bios.unc.edu

BAYESIAN MULTIVARIATE GROWTH CURVE LATENT CLASS MODELS FOR MIXED OUTCOMES

Benjamin E. Leiby*, Thomas Jefferson University
Mary D.Sammel, University of Pennsylvania
Thomas R. Ten Have, University of Pennsylvania
Kevin G. Lynch, University of Pennsylvania

In many clinical studies, the disease of interest is multi-faceted and multiple outcomes are needed to adequately characterize the disease or its severity. When the disease of interest has an unknown etiology and/or is primarily a symptom-defined syndrome, there is potential for the study population to be heterogeneous with respect to their symptom profiles. Identification of these subgroups is of interest as it may enable clinicians to provide targeted treatments or develop accurate prognoses. We propose a multivariate growth curve latent class model that group subjects based on multiple outcomes measured repeatedly over time. These latent classes are characterized by distinctive longitudinal profiles of a latent variable which is used to summarize the multivariate outcomes at each point in time. The mean growth curve for the latent variable in each class defines the features of the class. We develop this model for any combination of continuous, binary, ordinal or count outcomes within a Bayesian hierarchical framework. Simulation studies are used to validate the estimation procedures. We apply our models to data from a randomized clinical trial evaluating the efficacy of Bacillus Calmette-Guerin in treating symptoms of Interstitial Cystitis where we are able to identify a class of subjects where treatment was effective in reducing symptoms over an eight-month period.

email: bleiby@mail.jci.tju.edu

72. GENETIC EPIDEMIOLOGY/STATISTICAL GENETICS

EFFICIENT ASSOCIATION MAPPING OF QUANTITATIVE TRAIT LOCI WITH SELECTIVE GENOTYPING

Bevan E. Huang*, University of North Carolina at Chapel Hill
Danyu Lin, University of North Carolina at Chapel Hill

Genetic association studies offer valuable insight into relationships between genotype and phenotype. However, large samples are necessary to detect moderate effects, and the cost required to achieve acceptable power is often prohibitive. Selective genotyping (i.e. genotyping only those individuals with extreme phenotypes) can greatly improve the power to detect and map quantitative trait loci. Because selection depends on the phenotype, the resulting data cannot be properly analyzed by standard statistical methods. We provide appropriate likelihoods for assessing the effects of genotypes and haplotypes on quantitative traits under selective genotyping designs. We demonstrate that the likelihood-based methods are highly effective in identifying causal variants and are substantially more powerful than existing methods.

email: behuang@email.unc.edu

A GENERAL QUANTITATIVE GENETIC MODEL FOR CONSTRUCTING THE NETWORK OF HAPLOTYPE-HAPLOTYPE INTERACTIONS IN GENETIC ASSOCIATION STUDIES

Song Wu*, University of Florida
Jie Yang, University of Florida
Rongling Wu, University of Florida

Increased availability of genotyping single nucleotide polymorphisms (SNPs) from the genome has provided an excellent opportunity to identify and estimate specific DNA sequence variants that control a complex trait. Such sequence variants have been characterized by choosing a so-called risk haplotype that is different from the rest of haplotypes. However, this simple treatment may lose some information about haplotype effects, because it fails to estimate actions and interactions of individual haplotypes. In this talk, we will propose a general quantitative genetic model for studying haplotype effects on a complex trait by considering all possible haplotype combinations and constructing a network of haplotype-haplotype interactions. We derive a series of closed form for the EM algorithm to estimate and test the additive and non-additive effects of haplotypes. A model selection procedure is implemented to detect the optimal number and pattern of haplotype combinations that best explain the data. The statistical properties and utilization of the model are demonstrated through simulation studies and analyses of real data from a pharmacogenetic project. The model proposed will find its immediate applications to genetic association studies of complex traits.

email: swu@stat.ufl.edu

ESTIMATING A MULTIVARIATE FAMILIAL CORRELATION USING JOINT MODELS FOR CANONICAL CORRELATIONS

Hye-Seung Lee*, University of South Florida
Myunghee Cho Paik, Columbia University
Joseph H. Lee, Columbia University

Analyzing multiple traits can provide a better information than single trait to understand the underlying genetic mechanism of a common disease. To accommodate multiple traits in familial correlation analysis adjusting for confounders, we develop a regression model for canonical correlation parameters and propose joint modelling along with mean and scale parameters. The proposed method is more powerful than the regression method modelling pairwise correlations since it captures familial aggregation manifested in multiple traits through maximum canonical correlation.

email: leeh@epi.usf.edu

PERFORMANCE OF STATISTICAL PROCEDURES FOR THE DETECTION OF INTERLOCUS INTERACTIONS IN GENOME-WIDE ASSOCIATION STUDIES: STATISTICAL POWER AND TYPE I ERROR RATES

Solomon K. Musani*, University of Alabama at Birmingham
Amit Patki, University of Alabama at Birmingham
Hemant K. Tiwari, University of Alabama at Birmingham

Searching for and estimating the effects of gene × gene interactions (or epistasis) has recently aroused great interest among investigators, largely due to increasing empirical evidence from model organisms and human studies suggesting that interactions among loci contribute substantially to the genetics of complex human diseases, and availability of highly polymorphic SNP markers across the entire genome. Full assessment of interaction among loci has been hampered by high computational burden and multiple testing. A variety of statistical methods have been developed to address these issues, the leading of which is Multifactor Dimensionality Reduction (MDR) and two-step logistic regression. In this study, we compare the performance of MDR, two-stage logistic regression and the recently proposed Focused Interaction Testing Framework (FITF) in terms of statistical power and type I error rates assuming a variety of interlocus models.

email: smusani@ms.soph.uab.edu

USING DUPLICATE GENOTYPED DATA IN GENETIC ANALYSES: TESTING ASSOCIATION AND ESTIMATING ERROR RATES

Nathan L. Tintle*, Hope College
Stephen J. Finch, Stony Brook University
Derek Gordon, Rutgers University

Typically, researchers use duplicate genotyped (or re-genotyped) data for calculating an inconsistency rate. The relationship between this inconsistency rate and the genotyping error rate is not well specified. Additionally, the assumptions that are made when using the inconsistency rate as an ad-hoc measure of genotyping error have not been detailed. We present a model to demonstrate the relationship between the inconsistency rate and the genotyping error rate. Additionally, we present software that easily incorporates duplicate genotyped data into a genetic test of association with potential power gains when compared to methods that do not use duplicate genotyped data. The test of association is implemented as both an asymptotic test and a permutation test for small samples.

email: tintle@hope.edu

TAGGING SNP SELECTION WITH SUPERVISED SUPPORT VECTOR SCORE TEST AND RECURSIVE FEATURE ADDITION ALGORITHM IN CASE-CONTROL ASSOCIATION STUDIES

Yulan Liang*, University at Buffalo
Arpad G. Kelemen, University at Buffalo
Qingzhong Liu, New Mexico Institute of Mining and Technology

Given large scale SNPs markers of the human genome that are now available in public databases, a non-redundant subset of tagging SNPs identifications from massive genomic data have become the main task to be tackled with. In this paper,we propose a Supervised Support Vector Score Test and Recursive Feature Addition Algorithm for the identification of a subset of independent and highly predictive SNPs that are associated with complex diseases. The proposed feature selection scheme utilizes supervised learning and statistical similarity measures and it is independent of assumptions such as prior block-partitioning or boundary, bi-allelic SNPs and so on. Furthermore, we implement and evaluate the performance of our proposed approach in two classification scenarios: one ignores gene-gene/ gene-environment interactions such as naive classifier and the other utilizes these interactions with a Dynamical Evolutionary Neuro-Fuzzy Inference System. We apply our approaches to two real data sets in case-control designs (1) candidate gene cardiovascular data; (2) a genome-wide rheumatoid arthritis data. Preliminary results show that on the average, our supervised support vector based recursive feature addition with the use of different learning classifiers, outperforms the popular machine learning approaches such as Support Vector Machine Recursive Feature Elimination and Logic Regression for SNP identifications. Also, with the inclusion of gene-gene and gene-environment interactions the Dynamical Evolutionary Neuro-Fuzzy Inference System further improved classifications of genetic association study.

email: akelemen@buffalo.edu

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENE CATEGORIES IN MICROARRAY STUDIES USING MULTIVARIATE NONPARAMETRIC ANALYSIS

Dan Nettleton*, Iowa State University
Justin Recknor, Eli Lilly and Company
James M. Reecy, Iowa State University

We present a new method for identifying gene categories whose joint expression distribution differs across two or more experimental conditions. A multivariate nonparametric test statistic based on within-condition sums of distances between expression vectors is used to identify categories of interest. A permutation testing approach is used to obtain approximate control of the false discovery rate when testing multiple categories. The method provides an alternative to using Fisher's exact test on gene lists, gene set enrichment analysis (GSEA), significance analysis of function and expression (SAFE), and other approaches for finding gene categories of interest. We discuss some advantages of our approach relative to other methods and illustrate its use on an example experiment aimed at identifying gene categories whose joint expression distribution differs between wild type and myostatin-null mice.

email: dnett@iastate.edu

73. SURVIVAL DATA: VARIABLE SELECTION AND COMPETING RISKS

VALIDATION OF CLINICAL PROGNOSTIC MODELS FOR SUICIDE ATTEMPT

Hanga Galfalvy*, Columbia University
Maria A. Oquendo, Columbia University
John J. Mann, Columbia University

Suicidal behavior in patients with mood disorders is a complex phenomenon, and while many individual risk factors have been identified, simple models generally have limited predictive accuracy due to the low base rate for suicide attempts, and because risk factors can modify each other's effect. The aim of this project was to assess the utility of statistical validity measures as a tool for model selection in an observational study of suicide attempt with a relatively large number of candidate predictor variables. We evaluated the predictive accuracy of a number of prognostic models, ranging from basic survival models based on a few predictors to more advanced models using many terms and complex functional forms. We conclude that there are many models of similar predictive accuracy, and argue that other considerations (theoretical, clinical and practical) have an essential part in refining the list of optimal models.

email: hcg2002@columbia.edu

DOUBLY PENALIZED BUCKLEY-JAMES METHOD FOR SURVIVAL DATA WITH HIGH-DIMENSIONAL COVARIATES

Sijian Wang* (Van Ryzin Award Winner), University of Michigan
Bin Nan, University of Michigan
Ji Zhu, University of Michigan
David Beer, University of Michigan

Recent interest in cancer research focuses on predicting patients' survival by investigating gene expression profiles based on microarray analysis. We propose a doubly penalized Buckley-James method for the semiparametric accelerated failure time model to relate high-dimensional genomic data to censored survival outcomes, which uses a mixture of L1-norm and L2-norm penalties. Similar to the elastic-net method for a linear regression model with uncensored data, the proposed method performs automatic gene selection and parameter estimation, where highly correlated genes are able to be selected (or removed) together. The two-dimensional tuning parameter is determined by cross-validation and uniform design. The proposed method is evaluated by simulations and applied to the Michigan squamous cell lung carcinoma study.

email: sijwang@umich.edu

ROBUST REGRESSION FOR CENSORED RESPONSE FOR HIGH DIMENSIONAL COVARIATES

Huiliang Xie*, University of Iowa
Jian Huang, University of Iowa

The least-absolute-deviations (LAD) estimation is a robust alternative to the least squares method. Several approaches have been proposed in the literature to extend the LAD method to the accelerated failure time (AFT) model with censored data. However, it is difficult to adapt the existing censored LAD methods to problems with high-dimensional covariates, such as studies investigating the relationship between survival and gene expressions using microarrays. We propose a penalized weighted LAD method for censored data with high dimensional covariates. This method uses the Kaplan-Meier weights to account for censoring and imposes a lasso penalty on the regression coefficients. The proposed estimator can be computed using the existing programs for LAD regression. We use the V-fold cross validation and a deviation score for tuning parameter selection, and a bootstrap approach for variance estimation. We show that the proposed estimator is consistent and derive its asymptotic distribution. The proposed method is evaluated using simulations an demonstrated with two real data examples.

email: huiliang-xie@uiowa.edu

COMBINING MULTIPLE BIOMARKER MODELS IN SURVIVAL ANALYSIS

Zheng Yuan*, University of Michigan
Debashis Ghosh, University of Michigan

In medical research, there is great interest in developing methods for combining biomarkers. We argue that selection of markers should also be considered in the process. Traditional model/variable selection procedures ignore the underlying uncertainty due to the instability after the final model is selected. In this work, we propose a model combining method, Adaptive Regression by Mixing with Adaptive Screening (ARMAS), for censored survival data in biomarker studies. It works by considering weighted combinations of various Cox proportional hazard models; three different weighting schemes are developed. To reduce the set of models for combining in high-dimensional data, we propose an adaptive screening procedure in the ARMAS method based on the idea of adaptive penalty from adaptive model selection. In addition, we propose a modified version of our combining method based on the imputed data from multiple imputations, called ARMAS-impute, which does not rely on the proportional hazard assumption. Simulation studies are performed to assess the finite-sample properties of the proposed model combining methods. We compare the predictions of our proposed combining methods with the conventional predictions from a single model. It is illustrated with an application to a real data set from an immunohistochemical study in prostate cancer.

email: yuanz@umich.edu

SELECTION OF EXPERIMENTS FOR WEIGHTED DISTRIBUTIONS

Broderick O. Oluyede*, Georgia Southern University

Weighted distributions occur naturally for some sampling plans in biometry, survival analysis and reliability. In this paper, important results on inequalities and bounds for weighted distributions in general and length-biased distributions in particular are proved for monotone hazard functions and mean residual life functions. Results on sampling and selection of experiments from weighted distributions as opposed to the parent or original distributions are also presented.

email: boluyede@georgiasouthern.edu

SEMIPARAMETRIC ANALYSIS OF MIXTURE REGRESSION MODELS WITH COMPETING RISKS DATA

Wenbin Lu*, North Carolina State University
Limin Peng, Emory University

We present a novel regression approach for competing risks data, where the effects of covariates are formulated separately for the probability of occurrence of an event and the distribution of the time of occurrence of the event. The special mixture structure of competing risks data motivates the adaption of cure model methods for univariate survival data to the framework of competing risks. In this article, we develop an inference procedure based on generalised estimating equations, which does not rely on the independence between the censoring variable and covariates. We establish the consistency and asymptotic normality of the resulting estimators. A simple resampling method is proposed to approximate the distribution of the estimated parameters. Simulation studies and an analysis of a real example demonstrate that our method performs well with realistic sample sizes and is appropriate for practical use.

email: lu@stat.ncsu.edu

CUMULATIVE INCIDENCE FUNCTION UNDER THE SEMIPARAMETRIC ADDITIVE RISK MODEL

Seunggeun Hyun*, National Institute of Child Health and Human Development
Yanqing Sun, University of North Carolina at Charlotte
Rajeshwari Sundaram, National Institute of Child Health and Human Development

In analyzing competing risks data, a quantity of considerable interest is the cumulative incidence function. Typically one models the effect of covariates on the cumulative incidence function via using the proportional hazards model for the cause-specific hazard function. However, the proportionality assumption under the proportional hazards model may be too restrictive in practice. Motivated by this, we consider the more flexible additive hazards model of McKeague and Sasieni (1994). This model allows the effect of some covariates to be specified nonparametrically and some parametrically. We show how to construct confidence intervals and bands for the cumulative incidence function for patients with certain covariates. The finite sample property of the proposed estimators is investigated through simulations. We will illustrate our method with data from malignant melanoma study.

email: hyunseun@mail.nih.gov

74. BAYESIAN METHODS

COLLAPSED STICK-BREAKING PROCESSES FOR SEMIPARAMETRIC BAYES HIERARCHICAL MODELS

Mingan Yang*, National Institute of Environmental Health Science,NIH
David B. Dunson, National Institute of Environmental Health Science,NIH

In parametric hierarchical models, it is standard practice to place mean and variance constraints on the latent variable distributions for sake of identifiability and interpretability. Because incorporation of such constraints is challenging in semiparametric models that allow latent variable distributions to be unknown, previous methods either constrain the median or avoid constraints. In this article, we propose a collapsed stick-breaking process (CSBP), which induces mean and variance constraints on an unknown distribution in a hierarchical model. This is accomplished by viewing an unconstrained stick-breaking process as a parameter-expanded version of a CSBP. Theoretical properties are considered, an efficient blocked Gibbs sampler is developed for posterior computation, and the methods are illustrated through simulated and real data examples. Key Words: Dirichlet process; Latent variables; Moment constraints; Nonparametric Bayes; Random effects.

email: yangm2@niehs.nih.gov

CONSTRUCTING DEPENDENT POLYA TREES WITH COPULAS

Song Zhang*, University of Texas M.D. Anderson Cancer Center
Peter Muller, University of Texas M.D. Anderson Cancer Center

In many applications there are random variables such that although their distributions are unknown, they are believed to be dependent. Polya tree models are a type of Bayes nonparametric method to model unknown probability distributions. Compared with the popular Dirichlet process model, Polya tree models can be constructed to give probability one to the set of continuous or absolutely continuous probability measures. We develop dependent Polya trees to model dependent random distributions. We use copula to introduce the desired dependence. A simulation study and real data analysis are presented.

email: yszhang@wotan.mdacc.tmc.edu

THE NESTED DIRICHLET PROCESS

Abel Rodriguez*, Duke University
David B. Dunson, National Institute of Environmental Health Science, NIH
Alan E. Gelfand, Duke University

In multicenter studies, subjects in different centers may have different outcome distributions. This article is motivated by the problem of nonparametric modeling of these distributions, borrowing information across centers while also allowing centers to be clustered. Starting with a stick-breaking representation of the Dirichlet process (DP), we replace the random atoms with random probability measures drawn from a DP. This results in a nested Dirichlet process (nDP) prior, which can be placed on the collection of distributions for the different centers, with centers drawn from the same DP component automatically clustered together. Theoretical properties are discussed, and an efficient MCMC algorithm is developed for computation. The methods are illustrated using a simulation study and an application to quality of care in US hospitals.

email: abel@stat.duke.edu

SMOOTHING ANOVA FOR GENERAL DESIGNS

Yue Cui*, University of Minnesota
James S. Hodges, University of Minnesota

Analysis of variance (ANOVA) builds a model attributing variation in a response to predictors and to their interactions. There are different ways to determine which interactions to include in the model. A recent paper by Hodges et al (2006; henceforth HCSC) developed a Bayesian method called smoothed ANOVA (SANOVA), which neither includes nor excludes interactions, but smooths them, i.e., mostly removes small effects, mostly retains large ones and partly keeps middling ones. A simulation study in HCSC compared SANOVA to non-smoothed ANOVA, including ordinary least squares ANOVA and two-step ANOVA, which uses significance tests to drop interactions. SANOVA shows advantages if some interactions are actually absent. HCSC used a particular extended notion of degrees of freedom (DF) proposed by Hodges & Sargent (2001; henceforth H&S) to describe the extent of smoothing in the fitted values, and also used priors on DFs to induce priors on smoothing parameters. However, they considered smoothing only for balanced, single-error-term ANOVAs. This paper generalizes SANOVA to a broader class of linear models, including unbalanced ANOVA and ANOVA with random effects, but otherwise enables all the same methods demonstrated in HCSC.

email: yuecui@biostat.umn.edu

BAYESIAN DYNAMIC LATENT CLASS MODELS FOR MULTIVARIATE LONGITUDINAL CATEGORICAL DATA

Bo Cai*, University of South Carolina
David B. Dunson, National Institute of Environmental Health Science, NIH
Joseph B. Stanford, University of Utah

Dynamic latent class models provide a flexible framework for studying biologic processes that evolve over time. Although hidden Markov models (HMMs) are widely used, the Markov assumption is often violated. Motivated by studies of markers of the fertile days of the menstrual cycle, we propose a more general discrete-time dynamic latent class framework, allowing transitionrates to depend on time, fixed predictors and random effects. Observed data consist of multivariate categorical indicators, which change dynamically in a flexible manner according to latent class status. Given the flexibility of the framework, which incorporates semiparametric components using mixtures of betas, identifiability constraints are needed to define the latent classes. Such constraints are most appropriately based on the known biology of the process. Bayesian methods are developed for inference, and the approach is illustrated using mucus symptom data from a study of women using natural family planning.

email: bocai@gwm.sc.edu

BAYESIAN METHODS FOR HIGHLY CORRELATED DATA

Richard F. MacLehose*, National Institute of Environmental Health Sciences, NIH
David B. Dunson, National Institute of Environmental Health Sciences, NIH
Amy H. Herring, University of North Carolina
Jane A. Hoppin, National Institute of Environmental Health Sciences

Studies that include individuals with multiple highly correlated exposures are common in epidemiology. Because standard maximum likelihood techniques often fail to converge in such instances, hierarchical regression methods have seen increasing use. Bayesian hierarchical regression places prior distributions on exposure-specific regression coefficients to stabilize estimation and incorporate prior knowledge, if available. A common parametric approach in epidemiology is to treat the prior mean and variance as fixed constants. An alternative parametric approach is to place distributions on the prior mean and variance to allow the data to inform their values. As a more flexible semi-parametric option, one can place an unknown distribution on the coefficients while simultaneously clustering exposures into groups using a Dirichlet process prior. We also present a semi-parametric model with a variable-selection prior to allow clustering of coefficients at zero. We compare these four hierarchical regression methods and demonstrate their application in an example estimating the association of pesticides on retinal degeneration.

email: maclehoser@niehs.nih.gov

BAYESIAN HIERARCHICAL MODELS FOR THE TWO-COMPONENT MEMORY PROCESS

Xiaoyan Lin*, University of Missouri-Columbia
Dongchu Sun, University of Missouri-Columbia

The concept that human memory consists of several separate components is popular. One component is called recollection, which refers to the conscious recollection. The other component, automatic activation, reflects the automatic, unconscious activation of previously encountered material. One of the approaches to measure these two memory components is the Process Dissociation Procedure (PDP) proposed by Jacoby (1991). Bayesian hierarchical models are proposed. It features two probit links under Jacoby's PDP experiments. Each probit link has a form of linear mixed model with additive components that reflect the participant effects and item effects. Objective priors are used for accurate inference with participant and item variability. By including some latent variables, we have proven the posterior distributions of participants and items parameters are proper using constant prior and another special objective prior. Nonidentifiability problem of parameters is discussed in the project. Simulations are done for illustration.

email: xlzt3@mizzou.edu

75. ANALYSIS OF VERY LARGE GEOSTATISTICAL DATASETS

GAUSSIAN PREDICTIVE PROCESS MODELS FOR LARGE SPATIAL DATASETS

Sudipto Banerjee*, University of Minnesota
Alan E. Gelfand, Duke University
Andrew O. Finley, University of Minnesota
Huiyang Sang, Duke University

With inceased accessibility to geocoded locations through Geographical Information Systems (GIS), investigators are increasingly turning to spatial process models for modelling scientific phenomena. Over the last decade hierarchical models have become especially popular for spatial modelling, given their flexibility and power to estimate models that would be infeasible otherwise. However, fitting hierarchical spatial models involves expensive matrix decompositions whose computational complexity increases exponentially with the number of spatial locations, rendering them infeasible for large spatial data sets. The situation is exacerbated in multivariate settings with several spatially dependent response variables. Here we propose a predictive process derived from the original spatial process that projects process realizations to a lower-dimensional subspace thereby reducing the computational burden. We discuss attractive theoretical properties of this predictive process as well as its greater modelling flexibility compared to existing methods. A computationally feasible template that encompasses these diverse settings will be presented and illustrated.

email: sudiptob@biostat.umn.edu

APPROXIMATE LIKELIHOOD FOR LARGE IRREGULARLY SPACED SPATIAL DATA

Montserrat Fuentes*, North Carolina State University

Likelihood approaches for large irregularly spaced spatial datasets are often very difficult, if not infeasible, to implement due to computational limitations. Even when we can assume normality, exact calculations of the likelihood for a Gaussian spatial process observed at n locations requires O(n ^ 3) operations. We present an approximation to the Gaussian log likelihood for spatial regular lattices with missing values and for irregularly spaced datasets. This method requires O(nlog_2 n) operations and does not involve calculating determinants. We present simulations and theoretical results to show the benefits and the performance of the spatial likelihood approximation method presented here for spatial irregularly spaced datasets and lattices with missing values. We apply these methods to estimate the spatial structure of sea surface temperatures (SST) using satellite data with missing values.

email: fuentes@stat.ncsu.edu

GAUSSIAN PROCESS MODELS FOR HIGH DIMENSIONAL SPACES

Dave Higdon*, Los Alamos National Laboratory
Brian J. Williams, Los Alamos National Laboratory

Gaussian process models have proven to be very useful in modeling computer simulation output. This is because many computer codes are essentially noiseless and respond very smoothly to changes in input settings. In a typical setting, the simulation output is a function of a p-dimensional input vector x. In some applications, p may be as large as 60, however for the application of interest, the output is typically affected by a much smaller subset of these p inputs. After a fixed number of simulations are carried out, a GP model can be used to predict the simulation output at untried settings. When the number of simulations carried out to train the GP model are much over 1000, standard modeling and fitting approaches become computationally burdensome. In this talk, we describe strategies we've found useful for fitting such models (with large p) when the number of sussian process models for high dimensional spacesimulation runs is large.

email: dhigdon@lanl.gov

76. DIAGNOSTICS FOR MIXED MODELS

THE LINEAR MODEL HAS THREE BASIC TYPES OF RESIDUAL

John Haslett*, Trinity College, Dublin, Ireland
Steve J. Haslett, Massey University, Palmerston North-New Zealand

We consider residuals for the linear model with a general covariance structure. In contrast to the situation where observations are independent there are several alternative definitions. We draw attention to three quite distinct types of residuals: the marginal residuals, the model specified residuals and the full-conditional residuals. We adopt a very broad perspective including linear mixed models, time series and smoothers as well as models for spatial and multivariate data. We concentrate on defining these different residual types and discussing their inter-relationships. This work is in press with ISR, 2007. We show that for very many models the model specified residuals are natural, and are closely connected to the full-conditional residuals. But for important classes of models including spatial and multivariate models there are no such residuals. In these circumstances the marginal and full-conditional residuals play a basic joint role. Furthermore, the Delete=Replace principle shows that deletion diagnostics may most easily be derived from the full-conditional residuals. One application (see Haslett and Dillane, JRSSB, 2004) is to deletion-diagnostics for variance estimates in linear mixed models.

email: jhaslett@tcd.ie

EXTENDING THE BOX-COX TRANSFORMATION TO THE LINEAR MIXED MODEL

Matthew J. Gurka*, University of Virginia
Lloyd J. Edwards, University of North Carolina at Chapel Hill
Keith E. Muller, University of Florida
Lawrence L. Kupper, University of North Carolina at Chapel Hill

For a univariate linear model, the Box-Cox method helps choose a response transformation to ensure the validity of a Gaussian distribution and related assumptions. The popularity and relative simplicity of this transformation method has led to its use in more sophisticated models of longitudinal data. The desire to extend the method to a linear mixed model raises many vexing questions. Most importantly, how do the distributions of the two sources of randomness (pure error and random effects) interact in determining the validity of assumptions? For an otherwise valid model, we prove that the success of a transformation may be judged solely in terms of how closely the total error follows a Gaussian distribution. Hence the approach avoids the complexity of separately evaluating pure errors and random effects. The extension of the transformation to the mixed model requires an exploration of its potential impact on estimation and inference of the model parameters. Additional topics such as inference for the transformation parameter itself as well retransformation issues will be discussed. Analysis of longitudinal pulmonary function data and Monte Carlo simulations illustrate the proposed methodology.

email: mgurka@virginia.edu

RESIDUAL DIAGNOSTICS FOR LATENT VARIABLE MIXTURE MODELS

Chen-Pin Wang*, University of Texas Health Science Center at San Antonio
C. H. Brown, University of South Florida
Karen Bandeen-Roche, Johns Hopkins University

This talk presents graphical diagnostics to detect misspecification in latent variable mixture models regarding the number of latent classes, and class-specific means and covariance structures. Each type of model misspecification is quantified by a different empirical Bayes residual. Our procedure begins by imputing multiple independent latent classes for each individual based on his/her posterior class probability. Then the so-called pseudo-class adjusted diagnostic residuals are formed, their asymptotic property is assessed, and their empirical distributions are examined. A criterion for determining the minimum number of pseudo-class draws is derived. These methods are justified in simulation studies as well as in real data applications.

email: wangc3@uthscsa.edu

FORMAL AND INFORMAL MODEL SELECTION AND ASSESSMENT WHEN DATA ARE INCOMPLETE

Geert Molenberghs*, Hasselt University-Diepenbeek, Belgium

Every statistical analysis should ideally start with data exploration, proceed with a model selection strategy, and assess goodness of fit. Graphical tools are an important component of such a strategy. In the context of longitudinal data, oftentimes analyzed using mixed models, these steps are not necessarily straightforward, and the issues are compounded when data are incomplete. Indeed, some familiar results from complete (balanced) data, such as the desire for observed and expected curves to be close to each other, or the well known equivalence between OLS and normal-based regression, do not hold as soon as data are incomplete, unless in very specific cases. Such facts challenge the statistician's intuition and great care may be needed when exploring, building, and checking a longitudinal or multivariate model for incomplete data.

email: geert.molenberghs@uhasselt.be

77. DISCOVERING STRUCTURE IN MULTIVARIATE DATA USING LATENT CLASS AND LATENT FEATURE MODELS

LATENT CLASS MEASUREMENT OF HEALTH STATES LACKING A GOLD STANDARD

Karen Bandeen-Roche*, Johns Hopkins Bloomberg School of Public Health
Jeannie-Marie Sheppard, Johns Hopkins Bloomberg School of Public Health
Jing Ning, Johns Hopkins Bloomberg School of Public Health

Psychiatric disorders in mental health, and frailty and aging in gerontology, exemplify well-recognized health states that subject area specialists consider recognizable and theoretically well grounded, but that lack a gold standard. Thus, diagnosis must be driven by a combination of theory on the underpinnings and clinical presentation of the syndrome and utility for classification or risk prediction. This paper proposes methodology for the measurement of such health states that synthesizes approaches grounded in standard latent class modeling—i.e. focused on internal validity—with approaches focused on heightening concordance with criteria thought to be externally or concurrently validating. Performance properties of the methodology are evaluated and compared with more standard approaches applying individual validation criteria in turn, in analytical and simulation studies and application to health data. The paper aims to accomplish improved methodology for the measurement of health states lacking a gold standard.

email: kbandeen@jhsph.edu

FUNCTIONAL DATA ANALYTIC APPROACH TO PROCESSING MASS SPECTROMETRY DATA WITH AN APPLICATION TO SELDI-TOF MS ANALYSIS

Jaroslaw Harezlak*, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health
Shan Jiang, Tsinghua University-Beijing

In high-throughput mass spectrometry (MS) proteomic experiments, we can simultaneously detect and quantify a large number of peptides/proteins. Such detection techniques have good potential for discovery of new biomarkers related to diseases. Resulting data (spectra) from MS experiments are large and can be treated as finely sampled functions. Most of the existing MS analysis techniques involve multiple ad-hoc sequential methods for pre-processing the MS data, such as baseline subtraction, truncation, normalization, peak detection and peak alignment. We will discuss challenges in analyzing MS proteomic data and propose a unified statistical framework for pre-processing and post-processing mass spectra using advanced nonparametric regression and functional data analytic techniques in conjunction with statistical learning methods. We stress that pre-processing is critical in analysis of mass spectrometry proteomic data. We apply the methodology to a motivating data set obtained from a study of lung cancer patients whose plasma samples were collected and processed using a surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) MS instrument.

email: jharezla@hsph.harvard.edu

MODELING MULTIPLE LATENT CLASSES

Melanie M. Wall*, University of Minnesota

Latent class analysis typically involves modeling a set of different observed measurements via a single underlying (latent) categorical (class) variable which is meant to capture the associations found amongst the observed variables. This latent class variable can then be modeled as either an outcome or predictor variable to address some research question of interest. As applications of this type of modeling of a single latent class variable are becoming more common, it is natural to consider models involving multiple latent class variables. In particular, structural equation models (SEM) of latent class variables will be considered, differing from traditional SEM in that all the latent variables are categorical rather than continuous. In addition to basic main effects type models, models involving interactions effects between different latent class variables on outcomes will be demonstrated as well as structural model relationships between multiple latent class processes (over time). An example application relating social, familial, environmental and personal factors with adolescent obesity will be used.

email: melanie@biostat.umn.edu

78. PANEL DISCUSSION: RETHINKING THE FDA

PANEL DISCUSSION

Susan Ellenberg, University of Pennsylvania
Gary Koch, The University of North Carolina at Chapel Hill
Steve Snapinn, Amgen
Dalene Stangl, Duke University

79. STATISTICAL METHODS FOR INTERPRETING AND ANALYZING PROTEIN MASS-SPECTROMETRY DATA

INTRODUCTION TO MASS SPECTROMETRY BASED PROTEOMICS

Christopher J. Mason*, Mayo Clinic College of Medicine

Proteomics is the study of all of the proteins expressed in an organism or tissue. Because proteins are the molecules that most directly carry out biological function and sometimes cause disease, the human proteome is a rich source of potential diagnostic or prognostic biomarkers. However, the number of new protein-based clinical assays has been steadily declining in recent years to less than one per year. Several complications can explain this discrepancy: the large patient-to-patient variability of protein expression (requiring validation studies that have received less funding), the enormous dynamic range of protein expression in blood serum/plasma (from 35-50 x 10^9 pg/mL for serum albumin to 0-5 pg/mL for interleukin 6), the large number of protein forms (thought to be in the many millions), and the extreme diverse chemical properties of these molecules (proteins being composed from a diverse menu of 20 amino acids, as compared with the four bases of DNA). In spite of these challenges, there exists great potential for new protein biomarkers to reduce the burden of disease. Mass spectrometry, combined with advanced separation techniques, is a technique for unraveling the enormous complexity of the human proteome. Mass spectrometers can determine the mass-to-charge (m/z) ratios of molecules such as proteins in complex mixtures, can isolate component molecules from these mixtures based on their m/z ratios, and can fragment the selected molecules in order to further identify them. We will describe three types of mass spectrometers: time-of-flight (TOF), Fourier-transform ion cyclotron resonance (FT-ICR), and Orbitrap. We will discuss the form of the data (introduce concepts such as mass accuracy, resolving power, isotopic distributions), preprocessing steps, and give an overview of several common biomarker discovery experiments.

email: Mason.Christopher@mayo.edu

SPOT DETECTION AND QUANTIFICATION FOR 2-D PROTEOMIC DATA

Jeffrey S. Morris*, The University of Texas M. D. Anderson Cancer Center

For most proteomic assays, the protein signals are manifest as peaks in 1-dimensional assays (e.g. MALDI-TOF, SELDI-TOF) or spots in 2-dimensional assays (2-d gel electrophoresis, LC-MS). Effective peak and spot detection methods are needed to extract the proteomic information from these data. In this talk, I will briefly describe methods we previously developed for peak detection for 1-d assays, and then describe in more detail a new method for spot detection, matching, and quantification in 2-d assays. This method is automatic and quick, and yields spot quantifications that are reliable and precise. It incorporates a spot definition that is based on simple, straightforward criteria rather than complex arbitrary definitions, and results in no missing data. In the context of 2-d gels, we performed dilution series experiments and demonstrated the method’s superiority to two widely-used commercial software packages, yielding more precise and accurate spot quantifications. I will conclude by discussing how this procedure can be adapted to detect relevant features on LC-MS data.

email: jeffmo@mdanderson.org

AN APPLICATION OF BI-LINEAR MODELS TO A PROTEOMICS PROBLEM

Terry M. Therneau*, Mayo Clinic

In working with proteomic data, we have found that understanding exactly how the instrumentation works is crucial. The data at hand come from an experiment using paired samples, one processed in the presence of 18-oxygen water and the other in standard 16-oxygen water. The two samples are combined and fractionated into 8,000 sub-samples using chromatography; each of the resulting fractions is subjected to mass spectrometry. Each resulting spectrum will contain the 'fingerprints' of several dozen peptides (the compounds of interest). A peptide will appear in multiple sub-samples. We find that the ensemble can be expressed as a bi-linear model, y = b0 + a X b; where X is a known matrix based on natural isotopic distributions, b0 is the background, and a and b are vectors of unknown coefficients, a bilinear model. The traditional analysis (i.e. vendor's software) treats each spectrum and peptide separately, forms a sample1 vs. sample2 ratio and then averages the ratios' fractions. Advantages of the model based approach are the computation of standard errors for the parameters, the ability to account for non-uniform variance in y, and more efficient and reliable results.

email: therneau@mayo.edu

80. SURVIVAL ANALYSIS AND ITS APPLICATIONS IN GENETICS/GENOMICS

MODELING ASSOCIATION OF BIVARIATE COMPETING RISKS

Yu Cheng*, University of Pittsburgh
Jason P. Fine, University of Wisconsin-Madison

Frailty models are frequently used to analyze clustered survival data and evaluate within-cluster associations. However, they are seldom used in multivariate competing risks settings because of the challenge imposed by dependence structure between the primary and competing risks. To address this challenge, we focus on a nonparametrically identifiable quantity: cumulative incidence function (CIF). Frailty models are constructed expressing the bivariate CIF in terms of its marginals based on some improper random variables whose distribution functions corresponding to CIFs. Estimating equations are proposed to estimate the unknown association parameter involved in frailty models. The large sample properties of the association parameter estimators are established using empirical processes techniques and their practical performances are studied by Monte-Carlo simulations. We illustrate their practical utility by an analysis of dementia in the Cache County Study.

email: yucheng@pitt.edu

SEMIPARAMETRIC VARIANCE-COMPONENT MODELS FOR LINKAGE AND ASSOCIATION ANALYSIS OF CENSORED TRAIT DATA

Guoqing Diao*, George Mason University
Danyu Lin, University of North Carolina at Chapel Hill

Variance-component (VC) models are widely used for linkage and association mapping of quantitative trait loci in general human pedigrees. Traditional VC methods assume that the trait values within a family follow a multivariate normal distribution and are fully observed. These assumptions are violated if the trait data contain censored observations. Applying traditional VC methods to censored trait data would inflate type I error and reduce power. We present valid and powerful methods for the linkage and association analysis of censored trait data. Our methods are based on a novel class of semiparametric VC models, which allows an arbitrary distribution for the latent trait values. We construct appropriate likelihood for the observed data, which may contain left or right censored observations. The maximum likelihood estimators are approximately unbiased, normally distributed and statistically efficient. We develop stable and efficient numerical algorithms to implement the corresponding inference procedures. Extensive simulation studies demonstrate that the proposed methods outperform the existing ones in practical situations. An application to a real dataset is provided.

email: gdiao@gmu.edu

MULTIVARIATE SURVIVAL ANALYSIS FOR CASE-CONTROL FAMILY DATA

Li Hsu*, Fred Hutchinson Cancer Research Center

Multivariate survival data arise from case--control family studies in which the ages at disease onset for family members may be correlated. In this talk we consider a multivariate survival model with the marginal hazard function following the proportional hazards model. We use a frailty-based approach in the spirit of Glidden and Self (1999) to account for the correlation of ages at onset among family members. Specifically, we first estimate the baseline hazard function nonparametrically by the innovation theorem, and then obtain maximum pseudo-likelihood estimators for the regression and correlation parameters plugging in the baseline hazard function estimator. We establish a connection with a previously proposed generalized estimating equation-based approach (Shih and Chatterjee, 2002). Simulation studies and an analysis of case--control family data of breast cancer illustrate the methodology's practical utility. We will also discuss extensions of this approach to several directions.

email: lih@fhcrc.org

REGULARIZED ESTIMATION IN PATHWAY-BASED CENSORED DATA REGRESSION MODELING OF GENOMIC DATA

Hongzhe Li*, University of Pennsylvania
Zhi Wei, University of Pennsylvania

High-throughout genomic data provide an opportunity for identifying pathways and genes that are related to various clinical phenotypes, including censored survival phenotypes. Besides these genomic data, another valuable source of data is the biological knowledge about genes and pathways that might be related to the phenotypes of many complex diseases. Databases of such knowledge are often called the metadata. In this talk, we present two pathway-based censored data regression models, including both the pathway-based linear Cox regression model and the nonparametric pathways-based regression (NPR) for censored survival data to efficiently integrate genomic data and metadata. Such pathway-based models consider multiple pathways simultaneously and can allow complex interactions among genes within the pathways. We present a group penalized estimation procedure and a pathway-based gradient descent boosting procedure for identifying the pathways that are related to survival phenotypes. Applications to gene expression data sets on breast cancer distant metastasis/survival are presented. Extension to incorporate information contained in the pathway structures in the framework of spectral graph theory will also be discussed.

email: hli@cceb.med.upenn.edu

81. DENSITY ESTIMATION AND EMPIRICAL LIKELIHOOD

PROBLEMS RELATED TO EFFICACY MEASUREMENT AND ANALYSIS

Sibabrata Banerjee*, New Jersey Institute of Technology
Sunil Dhar, New Jersey Institute of Technology

Comparing two treatments on the basis of their primary efficacy variable is a situation which is commonly encountered in clinical research. More specifically, the quantity P(Y>X) categorized as the probabilistic index for the Effect Size (ES) is a matter of interest in clinical statistics. Here X and Y are the effects of the administered drugs A and B (replace A and B with two treatments). Our main goal was to come up with an efficacy measure that would let us compare these drugs more informatively and objectively. Also the methodology presented here is generalizable to other areas of application. Kernel Density Estimation is a very useful non-parametric method and has been relatively less explored as an applied statistical tool, mainly due to its computational complexity. We have found that this method is robust even under correlation structures that arise during the computation of all possible differences. We have also found that many of the bandwidth selection methods in the estimation process exhibit self similar statistical patterns, some of which can be statistically explained. Also, the same Kernel methods can be applied to the estimation of ROC (Receiver Operating Characteristic) curve, and to implement non-parametric regression of ROC. The area below the ROC curve (AUC) is also explored in this study. This area is exactly equal to the quantity P (Y>X).

email: sb95@njit.edu

NONPARAMETRIC BAYES STRUCTURAL EQUATION MODELS FOR MULTIVARIATE DATA

Ju-Hyun Park*, National Institute of Environmental Health Sciences
David B. Dunson, National Institute of Environmental Health Sciences

Structural equation models (SEMs) provide a broad framework for modeling of relationships in multivariate data, while reducing dimensionality. However, typical assumptions of normality and linearity may not be supported by the data. This article proposes a flexible nonparametric Bayesian SEM framework, allowing latent response variables to change nonparametrically with latent predictors, which are in turn modeled nonparametrically. The conditional distribution of the latent response variables is characterized using a mixture of seemingly unrelated regression (SUR) models, with the mixture distribution changing nonparametrically with predictors through a weighted mixture of Dirichlet processes (WMDP). The latent predictor distribution is then assigned a Dirichlet process mixture of normals. Identifiability issues are discussed, properties are considered, and an MCMC algorithm is developed for efficient posterior computation. The methods are illustrated using simulated data, and a reproductive epidemiology application.

email: juhyunp@email.unc.edu

ESTIMATING THE SLOPE IN LINEAR REGRESSION MODELS USING KERNEL DENSITIES

Thomas Jaki*, Cleveland State University

A class of estimators for the slope parameter, β_1, in simple linear regression based on likelihoods formed from kernel density estimates at varying bandwidths will be introduce. Via simulation, it is shown that the proposed estimator at an optimal bandwidth is decidedly superior to the ordinary least squares estimator for heavy tailed symmetric distributions. A computational method to estimate the optimal bandwidth will be implemented and evaluated.

email: jaki.thomas@gmail.com

EMPIRICAL LIKELIHOOD INFERENCE IN PRESENCE OF NUISANCE PARAMETERS

Mi-Ok Kim*, Cincinnati Children's Hospital Medical Center

Empirical likelihood (EL) is a nonparametric inference method with results that are in general similar to those about likelihood ratio tests and Wilk's theorem in the parametric model. In this talk we formulate EL inference in the frame work of ordinary parametric likelihood and show an analogy to the parametric likelihood. In particular we discuss EL inference in presence of nuisance parameters. We propose an approximate conditional EL inference and extend this proposition to censored case.

email: miok.kim@cchmc.org

BAYESIAN SMOOTHING OF DENSITY ESTIMATION VIA HAZARD RATES

Luyan Dai*, University of Missouri-Columbia
Dongchu Sun, University of Missouri-Columbia

The problem considered in this study is to estimate a bounded density by Bayesian smoothing techniques via hazards rates.The well-known intrinsic autoregressive process prior helps toconstruct frequentists' smoothness by Bayesian methods. Instead ofdirectly model on density, we investigate several transformationsbased hazards for the purpose of estimation. Beyond this, we proposed adaptive variance components model based on IAR priors to improve estimation. Under the proposed model, simulation studies are performed to evaluate estimation. The Bayesian computation can be realized via MCMC, implemented with Gilk's adaptive sampler and ratio of uniform sampler.

email: ld9n9@mizzou.edu

DENSITY ESTIMATION IN INFORMATIVE CENSORING

Bin Wang*, University of South Alabama

Survival data contain missing values due to censoring or trunca tion. Most of the existing statistical methods assume the missing information is missing at random -- the censoring is non-informative. However, in many clinical trials, the non-informative censoring assumption are frequently violated. In this paper, we will study the mechanisms of the typical informative censoring schemes and their effect on the estimate. A new non-parametric estimator will be proposed to estimate from informative censoring the its performance will be compared with some existing methods.

email: bwang@jaguar1.usouthal.edu

DIAGNOSTIC MEASURES FOR EMPIRICAL LIKELIHOOD OF GENERAL ESTIMATING EQUATIONS

Hongtu Zhu*, University of North Carolina at Chapel Hill
Niansheng Tang, Yunnan University, Kunming-People's Republic of China
Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Heping Zhang, Yale University School of Medicine

The method of general estimating equations is a general framework for modeling correlated data. To estimate the parameters in the estimating equations, empirical likelihood is often used thanks to its statistical efficiency and computational advantage. However, few diagnostic measures have been developed for assessing influential level of individual observations in the use of empirical likelihood for the estimating equations. The aim of this paper is to systematically develop diagnostic measures (e.g., residual) for assessing the influential level of each observation and use these diagnostic measures to construct goodness-of-fit statistics for testing the possible misspecification in the estimating equations. Our diagnostic measures include the case-deletion measures, the local influence measures, and the pseudo residuals. Our goodness-of-fit statistics includes the sum of the local influence measures and the processes of the pseudo residuals. Simulation studies are conducted to evaluate our methods, and real datasets are analyzed to illustrate the use of our diagnostic measures and goodness-of-fit statistics.

email: hzhu@bios.unc.edu

82. MEASUREMENT ERROR AND SURROGATE ENDPOINTS

ESTIMATION OF GENERALIZED PARTIALLY LINEAR MODELS WITH MEASUREMENT ERROR USING SUFFICIENCY SCORES

Lian Liu*, Texas A&M University

We study the partially linear model in logistic and other types of canonical exponential family regression when the explanatory variable is measured with independent normal error. We develop a backfitting estimation procedure to this model based upon the parametric idea of sufficiency scores so that no assumptions are made about the latent variable measured with error. We derive the method's asymptotic properties and present a numerical example and a simulation study.

email: lian@stat.tamu.edu

RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION OF A MEASUREMENT ERROR MODEL VIA THE MONTE CARLO EM ALGORITHM: ARE-ANALYSIS OF ENZYME ACTIVITY DATA

Antara Majumdar*, University at Buffalo
Randy L. Carter, University at Buffalo

A structural modeling approach to estimate the parameters of a general measurement error model that utilizes the Monte Carlo EM algorithm (MCEM) is described. In particular, a method to obtain restricted maximum likelihood estimators corrected for bias due to measurement error is presented. An instrumental variable is used to identify the model parameters. Common problems, available techniques and issues involved with implementations of MCEM are also discussed, particularly those associated with cases where a variance parameter is zero. The enzyme activity data of Carter (Biometrics, 1981) is re-visited to study the effects of this boundary value problem on MCEM methods and to otherwise illustrate the methods. We compare the MCEM results to the restricted maximum likelihood estimates obtained from a closed form solution by Carter.

email: aroy3@buffalo.edu

FLEXIBLE EVALUATION OF MIXED CONTINUOUS-BINARY SURROGATE ENDPOINTS USING INFORMATION THEORY AND SIMPLIFIED MODELING STRATEGY BASED ON META-ANALYTIC APPROACH

Pryseley N. Assam*, Hasselt University
Abel Tilahun, Hasselt University
Ariel Alonso, Hasselt University
Geert Molenberghs, Hasselt University

Surrogate endpoints can reduce the duration and hence the cost of a study. They may also lead to an increase in effectiveness and reliability of research, by reducing non-compliance and missingness. In this study we use the Meta-analytic approach by Buyse et al. (2000) on mixed continuous true and binary surrogate endpoints. We focus primarily on investigating, through a simulation study, the performance of simplified modeling strategies introduced by Tibaldi et al. (2003) using two-stage models. The information theory perspective to surrogate endpoint evaluation of Alonso et al. (2004) was implemented to quantify the individual level association. Secondary objectives of the simulation study involve comparison of the performance of penalized splines and linear regression at the second stage. Furthermore interest lies in the comparison of bootstrap and asymptotic confidence intervals, derived by Alonso et al. (2004), for the association measures. Some Key Words: Simplified modeling strategies; Meta-analysis; Surrogate endpoint; Information theory.

email: pryseley.assamnkouibert@uhasselt.be

FLEXIBLE SURROGATE MARKER EVALUATION FROM SEVERAL RANDOMIZED CLINICAL TRIALS WITH BINARY ENDPOINTS USING SAS

Abel Tilahun*, Hasselt University
Pryseley N. Assam, Hasselt University
Ariel Alonso, Hasselt University
Geert Molenberghs, Hasselt University

Surrogate endpoints come into play in a number of contexts in place of the endpoint of interest, commonly referred to as the true or main endpoint. In this study, we apply the meta-analytic approach of Buyse et al. (2000) for binary outcomes and introduce the use of the information theory perspective to quantify the individual level association as suggested by Alonso et al. (2004). Advantages and problems regarding the use of simplified approaches in line with the suggestions made by Tibaldi et al. (2003) are highlighted by means of a simulation study. The simulation study indicated the existence of discrepancy between associations at the observed and latent scales at both individual and trial levels. The modeling procedure involved a latent scale approach for the full bivariate model using a probit formulation and a direct modeling of the observed binary outcomes for the simplified models. One of the critical issues for the broad adoption of methodology like the one presented here is the availability of flexible implementations in standard statistical software. We have therefore developed generically applicable SAS macros at the reader's disposal. Some Key Words: Hierarchical model; Simplified model; Meta-analysis; Latent scale; Random-effects model; Surrogate endpoint

email: abel.tilahuneshete@uhasselt.be

ON SOME ASPECTS OF COVARIATES' MODELS WITH MEASUREMENT ERRORS IN A COMPLETELY RANDOMIZED DESIGN

Karabi Sinha*, University of Illinois at Chicago

Analysis of Covariance [ANCOVA] models are known to possess the special characteristics of blending the twin features of linear models for traditional [varietal] designs on one hand and the regression designs on the other. The design matrices are typically matrices involving the incidence patterns of assignable causes such as block effects or row-column effects, apart from the varietal or treatment effects. The incidence of such effects is reflected by the binary nature of the incidence matrices. On the other hand, the regression matrices reflect the extent of the regressors which are generally non-stochastic and continuous in nature. In this talk we propose to develop the general theory and related data analysis techniques for a covariates' model in situations wherein some or all of the covariates are subject to measurement errors. In particular, we confine to a CRD set-up and present detailed results from power considerations involving one covariate with measurement errors in a structural model.

email: karabi@uic.edu

ANALYSIS OF TISSUE MICROARRAY DATA USING MEASUREMENT ERROR MODELS

Ronglai Shen*, University of Michigan
Debashis Ghosh, University of Michigan
Jeremy MG Taylor, University of Michigan

Tissue Microarrays (TMAs) provide a proteomic platform for validating cancer biomarkers emerging from large-scale DNA microarray studies. Within-tumor repeated sampling results in TMA core-level expression measures that harbor substantial biological and experimental variability. The majority of TMA studies considers such expression variation negligible when associating the core-level repeated measures with patient survival outcome. Contrary to this common conception, the within-replicate variances are in fact substantial. In this article we propose to analyze TMA data in a measurement error model framework. A Latent Expression Index (LEI) is introduced as a robust TMA core-level summary expression index. We compare several estimators of LEI: an Empirical Bayes (EB), a Full Bayes (FB), and a Varying Replicate Number (VRN) estimator as the surrogate expression estimate to associate with patient survival outcome. In addition, we jointly model survival and TMA core-level data via a shared random effects model. Estimation is carried out through Bayesian posterior inference via implementing Markov Chain Monte Carlo (MCMC) methods. Simulation study is used to compare the performances of the various methods. Using two published TMA data sets, we re-evaluate the association of two prostate cancer biomarkers: AMACR - an enzyme in fatty acid synthesis, and BM28 - a chromosome maintenance protein involved in genome replication. We compare their expression estimates under different modeling approaches in differentiating risks of developing Prostate-Specific Antigen (PSA) failure among surgically treated prostate cancer patients. We found that implementing the measurement error models led to rather different conclusions about the prognostic values of these two biomarkers.

email: rlshen@umich.edu

IMPROVING POSTURAL INSTABILITY ONSET TIME MEASURE FOR PARKINSON'S DISEASE

Peng Huang*, Medical University of South Carolina
Ming Hui Chen, University of Connecticut
Debajyoti Sinha, Medical University of South Carolina

Onset of postural instability is a milestone of Parkinson's disease (PD) progression. However, current postural instability (PI) measures are largely affected by the inconsistent performance from both raters and patients. Statistical conclusions based on existing definitions of PI onset time often disguise and appear to contradict the degenerative nature of PD. This makes statistical findings difficult to interpret and unusable in designing subsequent studies. We propose a new measure of PI onset time using a Brownian motion-based latent process model of PD progression and provide a more clinically interpretable analysis of PD data. Theoretical and computational properties of the proposed model and the new PI onset time measure are examined. An application to a PD clinical trial dataset is presented.

email: huangp@musc.edu

83. SURVIVAL DATA: FRAILTY MODELS & CURE RATES

GAUSS QUADRATURE ESTIMATION IN FRAILTY PROPORTIONAL HAZARDS MODELS

Lei Liu*, University of Virginia

In this paper we apply Gauss quadrature techniques to estimation in frailty proportional hazards models. We approximate the unspecified baseline hazard by a piecewise constant baseline hazard, resulting in a parametric model which can be fitted conveniently by Gauss quadrature tools in standard software like SAS Proc NLMIXED. We conduct simulation studies to show that such approximation yields satisfactory results for both normal and Gamma frailty models. We apply our method to the analysis of diabetic retinopathy data.

email: liulei@virginia.edu

FRAILTY MODEL WITH SPLINE ESTIMATED BASELINE HAZARD

Nan Lin*, Washington University in St. Louis
Pang Du*, Virginia Tech

Frailty model has been a popular tool to model heterogeneity of individuals in different subpopulations. In the model, an individual's hazard rate depends partly on a frailty term, which is an unobservable random variable and supposed to act multiplicatively on the hazard. In this presentation, we propose a frailty model where the baseline hazard function is estimated nonparametrically by smoothing splines. The frailty is assumed to follow a log normal distribution, whose parameters are estimated jointly with the baseline hazard through the minimization of a penalized likelihood functional. The significance of frailty is checked by a procedure developed from the geometry of Kullback-Leibler distance. The performance of the proposed techniques is demonstrated in empirical studies and then applied to a real data example.

email: pangdu@vt.edu

BAYESIAN CASE INFLUENCE DIAGNOSTICS FOR SURVIVAL DATA

Hyunsoon Cho*, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Debajyoti Sinha, Medical University of South Carolina
Hongtu Zhu, University of North Carolina at Chapel Hill

We propose Bayesian case influence diagnostics for complex survival models. We develop case deletion influence diagnostics on both the joint and marginal posterior distributions based on the Kullback-Leibler divergence. We present a simplified expression for computing the Kullback-Leibler divergence between the posterior with complete data and the posterior based on case deletion, as well as investigate its relationships to the Conditional Predictive Ordinate (CPO). All the computations for the proposed diagnostic measures can be easily done by using MCMC samples from the complete data posterior distribution. The considered survival models are Cox's model with a gamma process prior on the cumulative baseline hazard, the frailty model, and the Cox model with a beta process prior in the presence of grouped survival data. We present a connection between our case-deletion based diagnostics and diagnostics based on Cox's partial likelihood. We present a real data example to demonstrate the methodology.

email: hscho@email.unc.edu

SHARED FRAILTY MODELS FOR JOINTLY MODELLING GROUPED AND CONTINUOUS SURVIVAL DATA

Denise A. Esserman*, University of North Carolina
Andrea B. Troxel, University of Pennsylvania School of Medicine

Quality of life (QOL) has become an important outcome in understanding treatment effects, especially in the palliative care setting. Thus, patients making treatment decisions should consider QOL as well as survival. To facilitate this assessment, we propose a shared frailty model that jointly models grouped (QOL) and continuous survival data. We focus on the univariate normal distribution to model the random frailty. Via simulations, we explore the robustness of the model to misspecification of the frailty variance (only in the fixed setting) and the frailty distribution when the frailty variance is both fixed and estimated using maximum likelihood (ML) and residual maximum likelihood (REML) methods. In addition, we explore the impact on parameter estimates of different amounts of censoring for the continuous outcomes. We present an example using a clinical trial in which both survival and QOL data were collected.

email: esserman@med.unc.edu

A NEW LATENT CURE RATE MARKER MODEL FOR SURVIVAL DATA

Sungduk Kim, University of Connecticut
Yingmei Xi*, University of Connecticut
Ming-Hui Chen, University of Connecticut

We propose a new mixture model via latent cure rate markers for survival data with a cure fraction. In the proposed model, the latent cure rate markers are modeled via a multimonial logistic regression. The proposed model assumes that the patients may be classified into several risk groups based on their cure fractions. Based on the nature of the proposed model, a posterior predictive algorithm is also developed to classify patients into different risk groups. This approach is well motivated by the current clinical practice. For example, prostate cancer patients are often classified as low, intermediate, and high risk groups based on their serum PSA levels, biopsy Gleason scores, and 1992 AJCC clinical tumor categories. The proposed model not only bears more biological meaning, but also fits the data much better than several existing competing cure rate models based on the popular LPML measure. In addition, we develop necessary theories of the proposed models and efficient Markov Chain Monte Carlo algorithms for carrying out Bayesian computation. A real data set from a prostate cancer clinical trial is analyzed in detail to further demonstrate the proposed methodology.

email: yingmei@stat.uconn.edu

A NEW THRESHOLD REGRESSION MODEL FOR SURVIVAL DATA WITH A CURE FRACTION

Sungduk Kim*, University of Connecticut
Ming-Hui Chen, University of Connecticut
Dipak K. Dey, University of Connecticut

Survival models incorporating cure rate fraction are becoming increasingly popular in analyzing time-to-event data. Due to the fact that certain fraction of the population suffering a particular type of disease get cured because of advanced medical treatment and health care system, we develop a very general class of models to incorporate cure fraction in presence of bacterial and viral load. To account for different immune systems for different individuals, we develop regression models for the number N of infected cells caused by bacteria or viral infection and the antibody level r of immune system. Various properties of the proposed models are carefully examined and efficient Markov Chain Monte Carlo methods are developed for carrying out Bayesian computation. LPML and DIC are used for comparing the proposed models to the existing competing models. A real data set from a prostate cancer clinical trial is analyzed in detail to further demonstrate the proposed methodology.

email: sdkim@stat.uconn.edu

ANALYSIS OF SMOKING CESSATION PATTERNS USING A STOCHASTIC MIXED EFFECTS MODEL WITH A LATENT CURED STATE

Sheng Luo*, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University
Thomas A. Louis, Johns Hopkins University
Nilanjan Chatterjee, National Cancer Institute, National Institutes of Health

We develop a mixed model to capture the complex stochastic nature of tobacco abuse and dependence. This model describes transition processes among addiction and non-addiction stages. An important innovation of our model is allowing an unobserved cure state, or permanent quitting, in contrast to transient quitting. This distinction is necessary to model data from situations wherein censoring prevents unambiguous determination that a person has been 'cured.' We apply our methodology to the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) Study, a large (29,133 participants) longitudinal cohort study. These data are used to model smoking cessation patterns using a discrete-time stochastic mixed-effect model with three states: smoking, transient cessation and permanent cessation (absorbent state). Random participant specific transition probabilities among these states are used to account for participant-to-participant heterogeneity. Another important innovation is to design computationally practical methods for dealing with the size of the data set and complexity of the models. This is achieved using the marginal likelihood obtained by integrating over the Beta distribution of random effects.

email: sluo@jhsph.edu

84. CANCER APPLICATIONS, INCLUDING SPATIAL CLUSTER DETECTION

INTRODUCING SPATIAL CORRELATION IN THE SPATIAL SCAN STATISTICS

Zhengyuan Zhu*, University of North Carolina at Chapel Hill
Ji Meng Loh, Columbia Univeristy

The spatial scan statistic is widely used in epidemiology and medical studies as a tool to identify hotspots in the incidence of disease cases. In common uses of the spatial scan statistic, the underlying model for the disease cases assumes independence between cases in different locations. In many instances, it is more likely that there is some positive spatial correlation. Even though covariates may be used to capture some of the spatial correlation, there is often residual correlation due to unmeasured covariates. We provide some theoretical findings that show that ignoring this spatial correlation results in an increased rate of false positives. This is verified through a simulation study. The study also shows that the presence of overdispersion can increase the rate of false alarms. We propose a modification procedure that aims to reduce the rate of false alarms by modeling any spatial correlation present in the data using a spatial generalized linear mixed model. Simulation studies show that this procedure can substantially reduce the rate of false alarms. These findings are also illustrated using data example involving brain cancer cases in New Mexico.

email: zhuz@email.unc.edu

EVALUATING SPATIAL METHODS FOR CLUSTER DETECTION OF CANCER CASES

Lan Huang*, National Cancer Institute
Barnali Das, National Cancer Institute
Linda Pickle, National Cancer Institute

Power and sample size requirements have been developed for independent observations but not for spatially correlated data. We are developing such requirements for a test of spatial clustering and cluster detection for cancer cases with Poisson distribution. We compared global clustering methods including Moran's I, Tango's MEET, extended Tango's MEET and Besag-Newell R, and cluster detection methods including circular and elliptic spatial scan statistic (SaTScan), flexibly shaped spatial scan statistics, Turnbull's cluster evaluation permutation procedure (CEPP), local indicator of spatial autocorrelation (LISA) and upper level set scan statistics (ULS). We identified eight geographic patterns that are representative of patterns of mortality due to various types of cancer in the United States from 1995-2000. We then evaluated the selected spatial methods based on county level data simulated from these different spatial patterns in terms of geographic locations and relative risks, and varying samples using the 2000 population in each county. The comparison provides insight into the power, precision of cluster detection and computing cost of the spatial methods when applying the cancer count data.

email: huangla@mail.nih.gov

RADON LEUKEMIA STUDY: A HIERARCHICAL POPULATION RISK MODEL FOR SPATIALLY CORRELATED EXPOSURE MEASURED WITH ERROR

Brian J. Smith*, The University of Iowa
Lixun Zhang, Yale University
William R. Field, The University of Iowa

We present a Bayesian model that allows for the joint prediction of county average radon levels and estimation of the associated leukemia risk. The methods are motivated by radon data from an epidemiologic study of residential radon in Iowa that include 2,726 outdoor and indoor measurements. Prediction of county average radon is based on a geostatistical model for the radon data which assumes an underlying continuous spatial process. In the radon model, we account for uncertainties due to incomplete spatial coverage, spatial variability, characteristic differences between homes, and detector measurement error. The predicted radon averages are, in turn, included as a covariate in Poisson models for incident cases of acute lymphocytic (ALL), acute myelogenous (AML), chronic lymphocytic (CLL), and chronic myelogenous (CML) leukemias reported to the Iowa cancer registry from 1973-2002. Since radon and leukemia risk are modeled simultaneously in our approach, the resulting risk estimates accurately reflect uncertainties in the predicted radon exposure covariate. Posterior mean (95% Bayesian credible interval) estimates of the relative risk associated with a 1 pCi/L increase in radon for ALL, AML, CLL, and CML are 0.91 (0.78-1.03), 1.01 (0.92-1.12), 1.06 (0.96-1.16), and 1.12 (0.98-1.27), respectively.

email: brian-j-smith@uiowa.edu

A WEIGHTED KAPLAN-MEIER APPROACH FOR ESTIMATION OF RECURRENCE OF COLORECTAL ADENOMAS

Chiu-Hsieh Hsu*, University of Arizona
Jeremy Taylor, University of Michigan
Qi Long, Emory University
David Alberts, University of Arizona
Patricia Thompson, University of Michigan

The effect of a colorectal polyp prevention trial is often evaluated on the status of recurrence at the end of the trial. Due to non-compliance from some participants, the data can be considered as current status data. The non-compliance could be informative of status of recurrence. In this paper we use mid-point imputation to handle interval censored observations. For the imputed data, we then perform a weighted Kaplan-Meier method on the compliance status to adjust for potential informative non-compliance and a risk score of recurrence to improve efficiency in estimation of recurrence rate at the end of the trial. The risk score is derived from a working logistic regression model for recurrence. In a simulation study, we show that the weighted Kaplan-Meier method can produce reasonable estimates of recurrence rate and can improve efficiency under an informative non-compliance situation with prognostic covariates compared to conventional logistic regression and Kaplan-Meier estimator. The method described here is illustrated with an example from a colon cancer study.

email: phsu@azcc.arizona.edu

NATURAL HISTORY MODEL OF METASTATIC PROGRESSION APPLIED TO LUNG CANCER

Maksim A. Pashkevich*, Stanford University
Bronislava M. Sigal, Stanford University
Sylvia K. Plevritis, Stanford University

We propose a parametric stochastic model of the natural history of cancer that predicts the distribution of the tumor size at the moment of metastatic transition, which is defined as the moment metastases become detectable by the common medical tests. Symptomatic detection of the primary tumor depends on the tumor size and stage of the disease. We derive maximum likelihood estimator for the parameters of the proposed model and apply it to data from the Mayo Lung Project, which was a clinical trial of x-ray screening for lung cancer among males. The estimator distinguishes between the cancer cases that were detected by symptoms and cases detected by medical testing prior to the onset of symptoms. For non-small cell lung cancer, we find that the hazard of symptomatic detection is proportional to the tumor volume, and is 12.8 times higher for tumors with clinically detectable metastases than for tumors without clinically detectable metastases. The median diameter of non-small lung cancer when the disease first involves clinically detectable metastases is 4.1 cm. Additional computer simulations were performed to investigate the sensitivity of the obtained inference to model assumptions.

email: Maksim.Pashkevich@stanford.edu

‘SMOOTH’ REGRESSION ANALYSIS OF ARBITRARILY CENSORED CLUSTER-CORRELATED TIME-TO-EVENT DATA

Lihua Tang*, North Carolina State University
Marie Davidian, North Carolina State University

Regression analysis of correlated, censored time-to-event data is of interest in family studies, litter-matched tumorgenesis studies, and other settings where the survival times may be thought of as arising in groups or ‘clusters, ‘ and the correlation among survival times in each cluster must be taken into account. A natural way to address such dependence is through incorporation of cluster-specific random effects. We propose an accelerated failure time model for such data that involves normally-distributed, mean zero random effects and a within-cluster ‘error’ term that is assumed to have distribution with a density satisfying mild ‘smoothness’ conditions. We approximate the smooth density by the ‘seminonparametric’ (SNP) representation of Gallant and Nychka (1987). This representation facilitates likelihood-based inference on the regression parameter, random effects variance components, and the density, which we implement by a Monte Carlo expectation-maximization algorithm; and we choose the tuning parameter and ‘kernel’ using standard information criteria. Moreover, arbitrary censoring patterns may be accommodated straightforwardly. We illustrate the approach via simulations and by application to data from the Diabetic Retinopathy Study.

email: ltang@ncsu.edu

85. VARIABLE SELECTION METHODS AND APPLICATIONS

VARIABLE SELECTION PROCEDURES FOR GENERALIZED LINEAR MIXED MODELS IN LONGITUDINAL DATA ANALYSIS

Hongmei Yang*, North Carolina State University
Daowen Zhang, North Carolina State University
Hao Helen Zhang, North Carolina State University

For high-dimentional non-Gaussian longitudinal data analysis, model selection is fundamental. But little research is focused on generalized linear mixed models (GLMMs) to date. We propose four procedures for model selection and parameter estimation in GLMMs: Full Likelihood approach (FL), Penalized Quasi-Likelihood approach (PQL), Approximate Marginal Likelihood approach(AML) and Two-stage Penalized Quasi-Likelihood approach(TPQL). Among them, FL and PQL have the feature of selecting informative variables and estimating regression parameters simultaneously. A robust estimator of standard deviation is derived based on a sandwich formula and tested through simulations for FL and PQL. A bias correction is proposed to improve the estimation accuracy of PQLMS. Simulations are used to evaluate the performance of four proposed procedures. In terms of model selection, PQLMS and TPQLMS perform the best, and the other two work very well. As for parameter estimation, FLMS, AMLMS and TPQLMS yield close results. Compared with FLMS, PQLMS is computationally preferable by using Laplace approximation to avoid intractable numerical integration.The flexible two-stage design of AMLMS and TPQLMS not only gains computational efficiency by avoid integration, but brings good estimators.

email: hyang3@ncsu.edu

BRIDGE LOGISTIC REGRESSION BASED ON ROC CRITERION

Guoliang Tian*, University of Maryland Greenebaum Cancer Center
Zhenqiu Liu, University of Maryland Greenebaum Cancer Center
Hongbin Fang, University of Maryland Greenebaum Cancer Center
Ming Tan, University of Maryland Greenebaum Cancer Center

It is well known that the bridge regression gives asymptotically unbiased estimates of the nonzero regression parameters while shrinking the estimates of zero (or small) regression parameters to zero, implying potentially better predictive performance. However, to our knowledge, there is a general lack of corresponding computational methods for modeling even for bridge linear regression. In this article, we first propose a new criterion defined by the receiver operating characteristic (ROC) curve to choose the appropriate penalty parameter instead of the conventional generalized cross-validation criterion. The model selected by the ROC criterion is shown to have better diagnostic accuracy while achieving variable selection at the same time. We then develop a fast EM-like algorithm for non-linear optimization with positivity constraints for model fitting. This algorithm is further applied to bridge regression where the regression coefficients are constrained with Lp norm with $p < 1$ for binary responses. Simulations and examples of prognostic factors and gene selection are presented to illustrate the proposed method.

email: gtian2@umm.edu

VARIABLE SELECTION USING RANDOM FORESTS

Andrejus Parfionovas*, Utah State University
Adele Cutler, Utah State University

This work demonstrates the efficiency of variable selection for multivariate analysis using Random Forests (RF) analysis. We compare the performance of RF against Logistic Regression (LR) which is de facto standard for variable selection in biological and medical studies. By using numerical simulations we observe higher success rate for choosing statistically important variables using RF on highly correlated and/or noisy data. The RF results are also more stable than those from LR. Our further studies of the variable selection mechanisms show the essential difference between the two approaches. Namely, the RF strategy is to assign the importance based on its explanatory value of the variable, including the correlated ones without loosing their importance. LR, on the other hand, cannot handle highly correlated variables. It focuses only on a subset of the variables that provide sufficient explanatory effect, thus completely ignoring other variables of possible interest. By using a number of tests we conclude that RF variable selection provides more informative insight of the data, is more robust and stable. Finally, based on the Kolmogorov-Smirnov test we propose a comprehensible and easy to use method for comparing the variables importance, which is illustrated on a real-life dataset for a cardiac events prediction problem.

email: andrej@cc.usu.edu

MODEL SELECTION FOR MULTIVARIATE SMOOTHING SPLINES WITH CORRELATED RANDOM ERRORS

Eren Demirhan*, North Carolina State University
Hao Helen Zhang, North Carolina State University

Model selection in nonparametric regression is a difficult problem, which becomes more challenging for correlated data such as longitudinal data and repeated measurements. Little work has been done on variable selection for nonparametric models with correlated errors. In the framework of smoothing spline analysis of variance, we propose a unified approach to simultaneously selecting variables and estimating model parameters and covariance structures. The new method, as a generalization of the component selection and smoothing operator (Lin and Zhang 2006), imposes a soft-thresholding penalty on functional components for sparse estimation and take into account covariance structure at the same time. One crucial issue is to adaptively choose two tuning parameters in the model: one controlling smoothness of the fitted surface and the other controlling the size of the model. We use the connection between the smoothing splines and mixed-effects models (Wang, 1998) and estimate the first tuning parameter as a variance component. The second parameter is tuned by criteria such as GML, GCV, and UBR. An efficient algorithm is developed for optimization, which solves the smoothing spline with correlated errors and a quadratic programming iteratively. The performance of the new method is demonstrated through simulations and real examples.

email: edemirh@ncsu.edu

NONPARAMETRIC BAYES LOCAL REGRESSION AND VARIABLE SELECTION

Yeonseung Chung*, University of North Carolina at Chapel Hill
David B. Dunson, NIEHS

Flexibly characterizing the relationship between a response and multiple predictors has been a great interest in many applications of regression smoothing and curve fitting. In this paper, we propose an infinite mixture of regression models, with the mixture weights varying with predictors. This is accomplished through a species sampling mixture (SSM) model for the joint distribution of the response and the predictors. In order to allow uncertainty in the predictors to be included, both within a local region and globally, we incorporate a hierarchical variable selection mixture structure in the base measure of the SSM prior for the regression coefficients. A blocked Gibbs sampler stochastic search algorithm is proposed for posterior computation. This algorithm allows not only for the estimation of the conditional mean and density, but also for the inference on the effects of the individual predictors through the computation of both global and local posterior inclusion probabilities for predictors. Theoretical properties are discussed and the methods are illustrated being applied to a simulated data and an epidemiologic problem.

email: chungy@email.unc.edu

VARIANCE COMPONENT SELECTION FOR MULTILEVEL PARTIALLY-REDUCED DOSE-RESPONSE CURVES IN CELL-CULTURE BIOASSAY

Carrie G. Wager*, Lansky Consulting

A common objective in cell-culture bioassay is to estimate the relative potency of a test relative to reference sample. Given a pair of sigmoidal dose-response curves with the same shape, relative potency is estimated as the horizontal distance between curves. Typically, several test samples are assayed together with a single reference sample across three or more blocks. Each block may have a distinct curve shape, while relative potency is expected to be consistent across all blocks. Some test samples will be similar to the reference sample, while others will not. We fit models to these data using a two-phase model selection strategy. The fixed effects structure is determined by equivalence assessments of each test versus reference pair: if confidence intervals for a parameter difference are within an indifference zone, the model is reduced to use a common parameter for test and reference samples. Variance components in the model are selected by AIC from among all subsets of the random effects imposed by the design structure. Alternating between these two phases occasionally fails to converge to a model where the fixed effects structure agrees with the equivalence assessment. We will discuss why this happens and our evolving approach to these issues.

email: carrie@lanskyconsulting.com

CORRECTION FOR MODEL SELECTION BIAS USING A MODIFIED MODEL AVERAGING APPROACH FOR SUPERVISED LEARNING METHOD APPLIED TO EEG EXPERIMENTS

Kristien Wouters*, Universiteit Hasselt-Belgium
José Cortiñas, Universiteit Hasselt-Belgium
Geert Molenberghs, Universiteit Hasselt-Belgium
Abdellah Ahnaou, Johnson & Johnson Pharmaceutical Research and Development-Belgium
Wilhelmus Drinkenburg, Johnson & Johnson Pharmaceutical Research and Development-Belgium
Luc Bijnens, Johnson & Johnson Pharmaceutical Research and Development-Belgium

A limited number of technologies exist for measuring brain activity. A graphical record of electrical activity of the brain via electroencephalography (EEG) is one of them. EEG experiments have been used in the past for research as well as clinical purposes, our particular interest is in preclinical pharmaco-electroencephalographical studies aiming at characterizing psychotropic drug effects on the basis of spectral EEG analysis. By using EEG-defined sleep-wake behavior in conjunction with electromyogram (EMG) and movement monitoring, clearly defined states of vigilance can be separated out and used to classify psychotropic agents. Typically, six sleep-wake stages are distinguished: Active and Passive Wake, Light, Deep, Intermediate Stage and REM Sleep. A two-step approach called doubly hierarchical discriminant analysis (DHDA) has been proposed to deal with such problems. It turned out to perform well even when the sample size of the training sample was rather small. In this approach the parameters of fractional polynomial mixed models built in the first step, are used in a second step in order to establish discriminant rules in a hierarchical fashion. However some problems related to model selection bias arise. Therefore here we proposed a modified model averaging approach that combined with the DHDA will provide more robust classification results.

email: kristien.wouters@uhasselt.be

86. GENERAL METHODS AND APPLICATIONS

UTILIZING UNSCHEDULED REPORTS OF DISEASE IN ESTIMATING RATES: A CONCEPTUAL FRAMEWORK FOR USE IN EVALUATING HEALTH OF WORKERS IN THE WORLD TRADE CENTER CLEANUP

Sylvan Wallenstein*, The Mount Sinai School of Medicine
Carol A. Bodian, The Mount Sinai School of Medicine
Jeanne M. Stellman, Mailman School of Public Health, Columbia University; and The Mount Sinai School of Medicine

Approximately 40,000 rescue and recovery workers were exposed to caustic dust and toxic pollutants following the Sept. 11, 2001 attack on the World Trade Center. A multi-center clinical program was established to provide free standardized care to responders [Hermert, 2006]. At present, more than 17,500 non-FDNY workers have undergone an initial evaluation. These workers are asked to return at regular intervals for standardized examinations. Some workers spontaneously report a diagnosis to the Screening Program personnel between regular examinations – for example at outreach programs, over the telephone after seeing press reports, or as part of a newly-funded Treatment Program. Currently, there is no context for workers to report the absence of disease between scheduled examinations. We examine the extent to which spontaneous reports are useful in estimating incidence rates of disease, focusing on the information that would otherwise be lost because a worker does not return for scheduled visits. We describe six possible types of response patterns, four of which are observable. We evaluate the assumptions that allow one to estimate the size of the missing groups, and briefly note extensions to estimate time to disease, and to allow for stratification of disease rates by age, gender, and race.

email: sylvan.wallenstein@mssm.edu

CALIBRATED SPATIAL STRATIFIED ESTIMATOR IN SPATIAL POPULATION

Jong-Seok Byun*, Hanshin University-Korea
Chang-Kyoon Son, Korea Institute for Health and Social Affairs
Jong-Min Kim, University of Minnesota

In general, the sampling design for the spatial population studies needs a model assumption of a dependent relationship, where the interesting parameters can be the population mean, proportion and area. This research also has the three assumptions as follows: (i) a spatial population has a large, close and irregular curve shape in two-dimension space such as the wasted area contaminated by some material, or the degree of distribution of animals or plants, (ii) there exists a dependent relationship between the observational points or sample points, and (iii) the population is distributed with cluster. We know that the study of an interested spatial population which is stratified by a geographical condition or shape, and the degree of distort of an estimation area is much useful. In light of this, if auxiliary information of the target variable such as wasted area contaminated by some material and the degree of distribution of animal or plants is available, then the spatial estimator might be improved through the calibration procedure. In this research, we propose the calibration procedure for the spatial stratified sampling in which we consider the one and two-dimensional auxiliary information.

email: jsbyun@hs.ac.kr

TESTS TO IDENTIFY CASES REQUIRING PROPORTIONAL HAZARDS MODELS WITH ESTIMATED INVERSE-SELECTION-PROBABILITY WEIGHTS

Qing Pan*, University of Michigan
Douglas E. Schaubel, University of Michigan

Often in observational studies of time to an event, the study population is a biased (i.e., unrepresentative) sample from the target population. In the presence of biased samples, it is common to weight subjects by the inverse of their respective selection probabilities. Pan and Schaubel recently proposed inference procedures for an inverse selection probability weighted (ISPW) Cox model, applicable when selection probabilities are not treated as fixed but estimated empirically. The proposed weighted regression parameter estimator is consistent for the target population parameter, while the unweighted estimator converges to a modification of the true value; the modification resulting from the potentially biased sampling mechanism. Similar statements apply to the weighted and unweighted cumulative hazard estimators. Although parameter estimation is more efficient relative to treating the estimated weights as fixed, computation is more intensive. Therefore, we propose a method for evaluating the bias in the unweighted partial likelihood and Breslow-Aalen estimators. Asymptotic properties of the test statistics are derived. The finite-sample significance level and power are evaluated through simulation. The proposed methods are then applied to data from a national organ failure registry to evaluate the bias in a post kidney transplant survival model.

email: qingpan@umich.edu

IMPUTATION IN A MULTIMODE MULTI-INSTRUMENT STUDY OF CANCER CARE

Yulei He*, Harvard Medical School
Alan M. Zaslavsky, Harvard Medical School

The Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium is a multisite, multimode and multiwave study examining the care delivered to population-based cohorts of newly diagnosed patients with lung and colorectal cancer and assessing predictors and outcomes of that care. Missing data are a serious concern for the CanCORS, as for other observational studies. We use weighting and multiple imputation to deal with nonresponse in the CanCORS baseline survey. The baseline survey uses several different instruments applicable to various groups of patients; consequently, the target population must be clearly defined for each analysis so we can calculate appropriate nonresponse weights. For imputation, it would be difficult to formulate a joint model that characterizes the underlying relationships among all the variables, especially since the surveys use multiple response formats including nominal, ordinal, and semicontinuous data, with many structured skip patterns. Instead, we applied the sequential conditional regression imputation approach (Raghunathan et al. 2001), which specifies a collection of models regressing incomplete outcomes on other covariates. We assess the performance of this approach in this complex dataset.

email: he@hcp.med.harvard.edu

ANALYSIS OF EFFECT OF SIDE COLLISION AVOIDANCE SIGNAL ON SIMULATED DRIVING WITH A NAVIGATION SYSTEM

Yi Ye*, University of North Florida
Dongyuan Wang, University of North Florida

Thirty Students from the University of North Florida participated in the study on the effect of side collision avoidance signal (SCAS) on Simulated Driving with a Navigation System. A computer-based STISIM driving simulator was used for this project. Participants were to respond to two signals. One signal was a visually displayed directional signal generated by a simulated navigation system. The other signal was an auditory signal from a simulated SCAS, which is presented shortly after the presence of the visual signal. Participants were instructed that the auditory signal conveyed directional information about impending threats, and that they needed to decide their response as they saw fit. Reaction time to the collision warning signal was the dependent variable. The statistical chanllenges in the analysis include smoothing, change point detection, generalized linear model, and etc.

email: yye@unf.edu

EVALUATING EARLY PROTEINURIA CHANGES AS A SURROGATE ENDPOINT FOR RENAL DISEASE OUTCOMES: A BAYESIAN APPROACH

Qian Shi*, University of Iowa
Mary K. Cowles, University of Iowa

The protective effect of angiotensin-II-receptor blocker irbesartan on end-stage renal disease (ESRD) was established in Irbesartan Diabetic Nephropathy Trial (IDNT). Baseline and follow-up albumin/creatinine ratios (ACR) and renal outcomes (time-to-event true endpoint) were recorded in IDNT. The objective of the present study was to evaluate early changes in proteinuria (continuous marker) as a surrogate for ESRD, using data from IDNT. Trial- and individual-level measures of surrogate validity, relative effect (RE) and adjusted association (AA), were used to assess the surrogacy. Two Bayesian joint models, normal/lognormal (JNL) and normal/Weibull (JNW) models, were developed to produce the posterior mean and 95% credible sets of RE and AA. The latter model captures the individual-level association between marker values and true endpoint through a latent component. Estimated RE's are approximately 0.5, with relatively wide 95% credible intervals. Estimated AA's are close to 0.3 and 0.7 for JNL and JNW model respectively. Sensitivity analyses show that inference is robust to the choice of the prior. Results from both models show that early changes in proteinuria may not be a reliable surrogate, at either the trial level or the individual level, for the renal disease outcomes.

email: qian-shi-1@uiowa.edu

A NEW MULTILEVEL NONLINEAR MIXTURE DIRICHLET MODEL FOR EST DATA

Fang Yu*, University of Connecticut
Ming-Hui Chen, University of Connecticut
Lynn Kuo, University of Connecticut
Wanling Yang, The University of Hong Kong

EST (Expressed Sequence Tags) experiments are employed to study gene expression from 'single pass' cDNA sequences. We use the observed relative frequency count for each unique tag (EST) to quantify the gene expression intensity. Similar to SAGE, the sample size is typically small relatively to the number of unique tags. Consequently, empirical estimators tend to have overestimation and underestimation of the population frequency. To overcome such problems, we propose a multinomial model with new multi-level nonlinear mixture Dirichlet prior distribution on the expression levels for each tag from multiple libraries of each tissue type. A Bayesian model selection criterion is developed to compare several proposed models that are different in either gene specific or library specific assumptions in the prior construction. The advantage of our model is that (1) it resolves the issues caused by over-representation and under-representation, (2) it allows us to borrow information from different libraries of the same tissue type, and (3) it provides direct measures of the gene-level expressions. We also develop novel computational algorithms and a gene selection criterion for detecting genes with different expressions across different tissue types. A real EST data set is analyzed to illustrate the proposed methodology.

email: fangyu@stat.uconn.edu

87. BIOSURVEILLANCE AND ANOMALY DETECTION

BIOSURVEILLANCE AND THE BIOSENSE PROGRAM

Henry R. Rolka*, Centers for Disease Control and Prevention

Empirical biosurveillance systems frequently use temporal and geographic contexts to establish a baseline against which to compare recent data. Numerous varieties of data types are used in varying levels of geographic scope across levels of public health. There are many components all of which must develop and function in concert in order for success in early event detection or situational awareness. A brief background and history of what has been done in this area at CDC will be presented with an update of the BioSense Program.

email: HRolka@CDC.Gov

AUTOMATED TIME SERIES FORECASTING FOR BIOSURVEILLANCE

Galit Shmueli*, University of Maryland
Howard S. Burkom, Johns Hopkins Applied Physics Laboratory
Sean P, Murphy, Johns Hopkins Applied Physics Laboratory

For robust detection performance, classic control chart alerting algorithms for biosurveillance require input data free of trends, day-of-week effects, and other systematic behavior. We describe three forecast methods and compare their predictive accuracy on each of 16 authentic syndromic data streams. The methods are 1) a nonadaptive linear regression model using a long historical baseline, 2) an adaptive regression model with a shorter, sliding baseline, and 3) the Holt-Winters exponential smoothing method.

email: gshmueli@rhsmith.umd.edu

HOW TO LIE WITH ROC CURVES AND RUN AMOC

Howard S. Burkom*, Johns Hopkins Applied Physics Laboratory
Sean P. Murphy, Johns Hopkins Applied Physics Laboratory

Recent research in biosurveillance has dealt with the evaluation of univariate and multivariate alerting algorithms for routine monitoring for disease outbreaks. The detection performance of these methods is often measured using ROC and AMOC curves. For known signal strength and shape, ROC curves plot the detection probability, or sensitivity, of an algorithm as a function of the false alarm rate. AMOC curves similarly give the tradeoff between alerting timeliness and the false alarm rate based on a timeliness score. Both curves are implicit functions of the alerting threshold. This presentation examines the calculation of these functions and presents pitfalls to avoid and adaptations needed for understanding of algorithm utility. For privacy reasons, biosurveillance data are difficult to obtain, and known outbreaks are too rare for reliable sensitivity estimates. Thus, data are often seeded with artificial signals in repeated trials for ROC/AMOC calculations. For practical insight into algorithm performance, these signals should be representative of the presumable footprint of a disease outbreak given the characteristics of the background data. Presented examples will range from data series with rich temporal structure to sparse counts, with discussion of appropriate measures in each situation.

email: Howard.Burkom@jhuapl.edu

EVALUATION OF THE DC DEPARTMENT OF HEALTH'S SYNDROMIC SURVEILLANCE SYSTEM

Michael A. Stoto*, Georgetown University School of Nursing and Health Studies
Beth Ann Griffin, RAND Corporation
Arvind K. Jain, RAND Corporation
John Davies-Cole, Department of Health
Chevelle Glymph, Department of Health
Garret Lum, Department of Health
Gebreyesus Kidane, Department of Health
Sam Washington, Department of Health

In September 2001, the District of Columbia Department of Health (DOH) began a syndromic surveillance program based on emergency room (ER) visits. ER logs from nine hospitals are categorized into mutually exclusive syndromes such as unspecified infection and gastrointestinal illness and analyzed using a variety of statistical detection algorithms. This paper characterizes the performance of these statistical detection algorithms in practical terms, and helps identify the optimal parameters for each algorithm. Analyses were conducted to improve the sensitivity of each algorithm to detecting simulated outbreaks by fine tuning key parameters used in the algorithms. Simulation studies using the data show that over a range of simulated outbreak types, the multivariate CUSUM algorithms performed more effectively than other algorithms. Performance of the algorithms is also examined by applying them to known outbreaks such as flu seasons and a previously undetected series of gastrointestinal illness outbreaks. Our analyses appear to indicate that the DC DOH system may prove to be more valuable in identifying the beginning of the flu season than for bioterrorist attacks. The analysis also indicates that when researchers/analyst apply these algorithms to their own data, fine tuning of parameters is necessary to maximize their performance.

email: stotom@georgetown.edu

88. INNOVATIONS IN SURVIVAL ANALYSIS METHODOLOGY FOR PUBLIC HEALTH PROBLEMS

TIME-VARYING CROSS-RATIO ESTIMATION FOR BIVARIATE SURVIVAL DATA

Bin Nan*, University of Michigan
Xihong Lin, Harvard University
James M. Robins, Harvard University

Cross-ratio is an important measure of local dependence between two survival times. A constant cross-ratio determines the Clayton (Biometrika, 1978) copula model for the joint survival function, and it can be estimated either by the two-stage method of Shih and Louis (Biometrics, 1995), or by the semiparametric likelihood method of Glidden and Self (Scand. J. Statist., 1999). Very often the cross-ratio is a function of times, and its estimation is done by using a specific functional form of the joint survival function. Recently, Nan et al. (JASA, 2006) studied the estimation of a piecewise constant cross-ratio that yields a sequence of Clayton copula models by artificially left truncating and right censoring the survival data. In this talk, we mimic the idea of the Cox's partial likelihood for the relative hazards regression and propose a pseudo partial likelihood approach for the estimation of the log cross-ratio that has a linear parametric form and thus is a smooth function of survival times. The method does not involve joint survival function. Its performance will be evaluated by intensive simulations. Potential extensions of the proposed method will be discussed. An application to marker evaluation for menopausal transition will be illustrated.

email: bnan@umich.edu

ESTIMATING THE EFFECT OF A TIME-DEPENDENT THERAPY IN OBSERVATIONAL STUDIES

Douglas E. Schaubel*, University of Michigan
John D. Kalbfleisch, University of Michigan

Survival analysis is often used to compare experimental and conventional therapies. In observational studies, complexity in the data structure, therapy assignment and the nature of the treatment effects may make it difficult to estimate the effect of therapy on outcomes in an accurate but parsimonious manner. For example, there are several important considerations in comparing survival between transplanted and wait-listed organ failure patients: treatment is time-dependent; the achieved reduction in mortality depends on time until transplant; post-transplant and wait-list mortality hazards are non-proportional; an intermediate event can occur which precludes transplantation (namely, removal from the wait-list). Organ transplantation has a long history in survival analysis and each of these issues can be addressed using both time-dependent covariates and time-dependent effects. However, such traditional methods generally yield parameter estimators with rather restrictive interpretations and which do not address the questions of chief interest. We propose novel semiparametric methods which involve weighting results from a sequence of stratified proportional hazards models. Asymptotic properties of the various effect measures are derived and their finite-sample applicability is assessed through simulation. The proposed methods are applied to data from the Scientific Registry of Transplant Recipients.

email: deschau@umich.edu

GENERALIZED SHARED FRAILTY MODELS

Alex Tsodikov*, University of Michigan
Szu-Ching Tseng, University of California-Davis

A generalization of the concept of frailty and self-consistency was presented in (JRSSB 65, 759-774, 2003) and applied to a family of univariate survival models, the so-called Nonlinear Transformation Models (NTM). A multivariate extension of this approach is presented. A subclass of Archimedian copula models is identified that generalizes shared frailty models and associated EM algorithms. A connection to recently proposed MM algorithms that extend the EM concept without using missing data arguments is established.

email: tsodikov@umich.edu

89. INSTRUMENTAL VARIABLE METHODS FOR CAUSAL INFERENCE

REGRESSION AND WEIGHTING METHODS USING INSTRUMENTAL VARIABLES

Zhiqiang Tan*, Johns Hopkins University

Recent researches in econometrics and statistics have gained considerable insights into the use of instrumental variables (IVs) for causal inference. We build on the modern IV framework including assumptions and identification results, and develop two estimation methods in parallel to regression adjustment and propensity score weighting in the case of treatment selection based on covariates. The IV assumptions are made explicitly conditional on covariates to allow for the fact that instruments can be related to these background variables. The regression method focuses on the relationship between responses (observed outcome and treatment status jointly) and instruments adjusted for covariates. The weighting method focuses on the relationship between instruments and covariates in order to balance different instrument groups with respect to covariates. The approach is flexible enough to host various semiparametric and nonparametric techniques (including model building and checking) that attempt to learn associational relationships from observed data. We illustrate the methods by an application to estimating return to education.

e-mail: ztan@jhsph.edu

STRUCTURAL PROPORTIONAL HAZARDS MODELS FOR CAUSAL INFERENCE IN RANDOMIZED TRIALS

Els J. Goetghebeur*, Ghent University-Belgium
An Vandebosch, Janssen Pharmaceutica-Belgium

We propose Structural Proportional Hazards models as an alternative to Structural Accelerated Failure Time models to evaluate the effect of observed treatment exposure on HIV incidence in a randomized HIV prevention trial conducted in Africa and Asia (Van Damme et al., Lancet 2004). Straightforward implementation of this new methodology restricts flexible model fitting as it involves demanding grid searches for the estimation of more dimensional structural parameter spaces and their precision. To overcome some of these limitations we invoke an auxiliary model for potential exposures and solve score equations averaged over multiple imputations for causal parameters. In the process we have a more symmetrical treatment of missing information on potential experimental exposure for subjects on the placebo arm and potential placebo response for subjects on the treatment arm. It also allows for a natural extension to incorporate interval censored data and recover some information.

e-mail: els.goetghebeur@ugent.be

THRESHOLD CROSSING MODELS AND BOUNDS ON TREATMENT EFFECTS

Edward J. Vytlacil*, Columbia University
Azeem Shaikh, University of Chicago

This paper considers the evaluation of the average treatment effect of a binary endogenous regressor on a binary outcome when one imposes a threshold crossing model on both the endogenous regressor and the outcome variable but without imposing parametric functional form or distributional assumptions. Without parametric restrictions, the average effect of the binary endogenous variable is not generally point identified. This paper constructs sharp bounds on the average effect of the endogenous variable that exploit the structure of the threshold crossing models and any exclusion restrictions. We also develop methods for inference on the resulting bounds.

e-mail: ev2156@columbia.edu

EFFICIENT NONPARAMETRIC ESTIMATION OF CAUSAL EFFECTS IN RANDOMIZED TRIALS WITH NONCOMPLIANCE

Jing Cheng*, University of Florida College of Medicine
Dylan S. Small, University of Pennsylvania
Zhiqiang Tan, Johns Hopkins University
Thomas R. Ten Have, University of Pennsylvania

Causal approaches based on the potential outcome framework provide a useful tool for addressing the noncompliance problems in randomized trials. Various estimators, e.g. the instrumental variable (IV) estimator, have been proposed for causal effects of treatment. In this paper, we propose a new empirical likelihood-based estimator of causal treatment effects in randomized clinical trials with noncompliance. By using the empirical likelihood approach to construct a profile random sieve likelihood and taking into account the mixture structure in outcome distributions, this estimator is robust to parametric distribution assumptions and more efficient than the standard IV estimator. Our method is applied to data from a randomized trial of an encouragement intervention to improve adherence to prescribed depression treatments among depressed elderly patients in primary care practices.

e-mail: jcheng@biostat.ufl.edu

90. DOSE-FINDING IN CLINICAL TRIALS

MODEL-BASED DESIGNS FOR DRUG COMBINATIONS

Valerii V. Fedorov*, GlaxoSmithKline

A parametric model is considered for the joint probability of the efficacy-toxicity response as a function of the two doses of the combined drugs. Based on fixed trial-specific standards of the minimum acceptable efficacy response rate and the maximum tolerated toxicity rate, the therapeutic region is defined as a region on the plane of available doses of the two combined drugs. The optimization problem is formulated as a maximization, within the therapeutic region, of the probability of positive efficacy response with non-toxicity as a function of the two doses. The proposed penalized adaptive design selects the successive dose combinations by maximizing the increment of information per cost unit given the current data.

e-mail: valeri.v.fedorov@gsk.com

TWO-STAGE PHASE I DESIGNS IN CANCER TRIALS

Tze L. Lai*, Stanford University

The primary objective of a Phase I cancer trial is to estimate the maximum tolerated dose (MTD) of a new treatment, and a secondary objective is to study the treatment's efficacy. To avoid subjecting patients to severe toxicity when there is no experience with the new treatment, a conservative dose-escalation design is often used even though it has been documented in the statistical literature to be very inefficient for estimating the MTD. To address these issues, we propose a two-stage design, in which the first stage consists of m dose levels chosen by the conventional dose escalation approach. A parametric quantal response curve is fitted at the end of the first stage and updated sequentially afterwards. The second stage chooses the remaining n-m dose levels via a nonlinear sequential design similar to that of Wu (1985, JASA), but with a penalty term for the excess probability of toxic response, or via its Bayesian counterpart. The two-stage procedure can also incorporate efficacy information at the end of the first stage to come up with a more realistic definition of the MTD, which is often loosely defined and can vary with the treatment's observed efficacy.

e-mail: lait@stat.stanford.edu

MONITORING LATE ONSET TOXICITIES IN PHASE I TRIALS USING PREDICTED RISKS

Neby Bekele, University of Texas-M. D. Anderson Cancer Center
Yuan Ji*, University of Texas-M. D. Anderson Cancer Center
Yu Shen, University of Texas-M. D. Anderson Cancer Center
Peter F. Thall, University of Texas-M. D. Anderson Cancer Center

Late onset toxicities are a serious concern in phase I trials of cytotoxic agents. Since most dose-limiting toxicities occur soon after the start of therapy, conventional dose-finding methods rely on a binary indicator of toxicity occurring within a relatively short initial time period. If an agent causes late onset toxicities, however, such methods may allow an undesirably large number of patients to be treated at toxic doses before any toxicities are observed. A method addressing this problem is the time-to-event continual reassessment method. We propose a Bayesian dose-finding method similar to the TITE-CRM in that doses are chosen based on right-censored toxicity time data. The main new aspect of our method is a set of rules, based on predictive probabilities, that temporarily suspend accrual if the risk of toxicity at prospective doses for future patients is unacceptably high. If additional follow up data later reduce the predicted risk of toxicity to an acceptable level, then accrual is re-started, and this process may be repeated several times during the trial. An extensive simulation study shows that our proposed method provides a greater measure of safety than the TITE-CRM while still reliably choosing the preferred dose, and this advantage increases with accrual rate. The price of this extra safety is that on average the trial takes longer to complete.

e-mail: yuanji@mdanderson.org

PATIENT-SPECIFIC DOSE-FINDING BASED ON BIVARIATE OUTCOMES WITH COVARIATES

Peter F. Thall*. University of Texas-M. D. Anderson Cancer Center
Hoang Nguyen, University of Texas-M. D. Anderson Cancer Center

A Bayesian method for covariate-adjusted, or "individualized" dose-finding based on a bivariate (efficacy, toxicity) outcome is presented. The method extends Thall and Cook (Biometrics 60:684-693, 2004). Implementation requires an informative prior on covariate effects, obtained from historical data or by elicitation. In the underlying probability model, dose and covariate main effects and dose-covariate interactions are included in the linear components of the marginal outcome probabilities. For each of a representative set of covariate vectors, limits on the probabilities of efficacy and toxicity specified by the physician are used to construct bounding functions that are used to determine the acceptability of each dose for each possible covariate vector. The physician also must specify equally desirable target (efficacy, toxicity) probability pairs for a reference patient's covariates to characterize trade-offs between the two outcomes. Each patient's dose is chosen to optimize the efficacy-toxicity trade-off for his/her specific covariates. Because the selected doses are covariate-specific and the method is sequentially outcome-adaptive, different patients may receive different doses at the same interim point in the trial, and some initially eligible patients may have no acceptable dose. The method is illustrated by application to a phase I/II trial in acute leukemia.

e-mail: rex@mdanderson.org

91. NON-STATIONARY TIME SERIES ANALYSIS WITH APPLICATIONS TO BIOMEDICAL DATA

DISCRIMINATION OF BRAIN SIGNALS USING LOCALIZED HIGHER ORDER SPECTRA

Hernando Ombao*, University of Illinois at Urbana-Champaign

We consider a data set that consists of MEG (magnetoencephalogram) signals recorded from healthy controls and schizophrenic patients. Our neuroscience collaborators are interested in identifying features that can separate the two groups with the hope that these physiological measures may be used in conjunction with behavioral measures for patient diagnosis. In this talk, we will develop an automatic procedure for time-frequency spectral feature selection via localized transforms. Moreover, given the high degree of complexity of brain signals, we will consider the time-evolutionary spectrum of non-linear transforms as potential features for classification and discrimination.

e-mail: ombao@uiuc.edu

LOCAL SPECTRAL ENVELOPE

David S. Stoffer*, University of Pittsburgh

A statistical concept called the spectral envelope was introduced as a general method to study patterns in long sequences of letters or symbols (such as codes). The most well known application of this methodology is the analysis of DNA sequences. Because many of the applications where the spectral envelope has been an asset are situations where the assumptions of stationarity and homogeneity are questionable, there was a need to develop a local version of the methodology, and that is the focus of this talk.

e-mail: stoffer@pitt.edu

MULTIVARIATE TIME-DEPENDENT SPECTRAL ANALYSIS USING CHOLESKY DECOMPOSITION

Ming Dai*, University of North Carolina-Charlotte
Wensheng Guo, University of Pennsylvania

In this paper, we propose a nonparametric method to estimate the spectrum of a multivariate locally stationary process, which is assumed to be smooth in both time and frequency. In order to ensure that the final estimate of the multivariate spectrum is positive definite while allowing enough flexibility in estimating each of its elements, we propose to smooth the Cholesky decomposition of an initial spectral estimate and the final spectral estimate is reconstructed from the smoothed Cholesky elements. We propose a two-stage estimation procedure. The first stage approximates the locally stationary process by a piecewise stationary time series to obtain an initial spectral estimate. The second stage uses a smoothing spline ANOVA to jointly smooth each Cholesky element in both time and frequency, and reconstructs the final estimate of the time varying multivariate spectrum for any time-frequency point. The final estimate is a smooth function in time and frequency, has a global interpretation, and is consistent and positive definite. We show that the Cholesky decomposition of a time varying spectrum can be used as a transfer function to generate a locally stationary time series with the designed spectrum. This not only provides us much flexibility in simulations, but also allows us to construct bootstrap confidence intervals on the time varying multivariate spectrum.

e-mail: mdai@uncc.edu

TIME-FREQUENCY FUNCTIONAL MODEL

Li Qin*, Fred Hutchinson Cancer Research Center
Wensheng Guo, University of Pennsylvania
Brian Litt, University of Pennsylvania

Unlike traditional time series analysis that focuses on one long time series, in many biomedical experiments, it is common to collect multiple time series and focus on how the covariates impact the patterns of stochastic variation over time. In this article, we propose a time-frequency functional model for a family of time series indexed by a set of covariates. This model can be used to compare groups of time series in terms of the patterns of stochastic variation and to estimate the covariate effects. We focus our development on locally stationary time series and propose the covariate-indexed locally stationary setting, which include stationary processes as special cases. We use smoothing spline ANOVA models for the time-frequency coefficients. A two-stage procedure is introduced for estimation. In order to reduce the computational demand, we develop equivalent state space model to the proposed model with an efficient algorithm. We also propose a new simulation method to generate replicated time series from their design spectra. An epileptic intracranial electroencephalogram (IEEG) data set is analyzed for illustration.

e-mail: lqin@fhcrc.org

92. INTRODUCTORY LECTURE SESSION: SOFTWARE PACKAGES FOR HANDLING MISSING DATA

XMISS IN CYTEL STUDIO 7 FOR REGRESSION MODELS WITH MISSING COVARIATES

Ming-Hui Chen*, University of Connecticut

XMISS is a statistical software package which is designed to handle missing covariates. XMISS implements the ML methodology based on the EM in multiple linear regression for continuous outcomes, logit, probit and complementary log-log regression for binary and binomial outcomes and Poisson regression for count data. The first version of XMISS package was released by Cytel Inc. in November 2005. The current version of the software provides ML estimates of the regression coefficients and their standard errors. The software provides confidence intervals and p-values for the regression coefficients as well as summary statistics for all covariates with missing values. XMISS also provides estimates and standard errors for the parameters of the covariate distribution. The current version of the software can handle up to 50 covariates total, of which 10 of them may have missing values. The software automatically allows for interaction terms between the covariates. Future releases will accommodate both missing discrete categorical and/or response data in GLMs and survival models with right censoring. In this presentation, we will provide an overview of the statistical methods and the computational algorithms used in XMISS. Live demos of the XMISS software using the various missing data models will also be given.

e-mail: mhchen@stat.uconn.edu

SOFTWARE PACKAGES FOR HANDLING MISSING DATA

Nicholas J. Horton*, Smith College
Ken P. Kleinman, Harvard Medical School and Harvard Pilgrim Health Care

Missing data arise in almost all real-world studies, and an extensive statistical literature has developed to address these complications. Software to account for missing values is now widely available, and is increasingly used in published studies. We will describe software implementations available within SAS (PROC MI and IVEware), and summarize the advantages and limitations of these methods for the analysis of incomplete data regression models.

e-mail: nhorton@email.smith.edu

93. ANALYSIS OF LABORATORY EXPERIMENTS

DESIGNS FOR MODERN BIOASSAY: EXPLOITING THE COMBINED STRENGTHS OF LAB ROBOTS AND MULTILEVEL MODELS

David M. Lansky*, Lansky Consulting, LLC

Bioassays use comparisons among similar dose-response curves, under the assumption that two samples contain the same active analyte, to estimate the relative functional activity of the samples. Cell based bioassays are routinely performed in 96 well culture plates with samples and doses assigned to rows or columns, imposing a strip-plot or crossed error design. Implementation of even the simplest of these designs using randomization is rarely practical with hand pipetting. Randomization of these designs by robot is becoming routine. Exploiting the combined capabilities of lab robots and multilevel models offers opportunities for newly practical novel design combinations. Rather than focus on formal optimality criteria, we have instead focused on designs that have good properties on a collection of practical and statistical criteria. Effective designs include familiar design methods such as partially balanced incomplete blocks (PBIB) and uniformity trials, as well as less familiar design methods such as strip-plot and staggered dose designs.

e-mail: david@lanskyconsulting.com

THE ASSESSMENT OF ANTITUMOR ACTIVITY FOR SOLID TUMOR XENOGRAFT STUDIES

Jianrong Wu*, St Jude Children's Research Hospital

Three exponential tumor growth models are developed to characterize tumor growth in the mouse solid tumor xenograft study for the Pediatric Preclinical Testing Program (PPTP). Furthermore, a model-based antitumor activity measurement, the AUC ratio, is proposed. The typical antitumor activity measurements used in preclinical xenograft studies, such as $\log_{10}$ cell kill and tumor growth inhibition T/C could either be inapplicable or mislead the antitumor activity evaluation. The AUC ratio provides a more insightful antitumor activity evaluation throughout the entire treatment period.

e-mail: jianrong.wu@stjude.org

APPLICATION OF A FOUR PARAMETER LOGISTIC MODEL FOR ESTIMATING TITERS OF FUNCTIONAL MULTIPLEXED PNEUMOCOCCAL OPSONOPHAGOCYTIC KILLING ASSAY (MOPA)

Deli Wang*, The Comprehensive Cancer Center-The University of Alabama at Birmingham
Robert L. Burton, The University of Alabama at Birmingham
Moon H. Nahm, The University of Alabama at Birmingham
Seng-jaw Soong, The Comprehensive Cancer Center-The University of Alabama at Birmingham

A multiplexed In vitro opsonization assay (MOPA) is widely accepted to quantitate Streptococcus pneumococcal antibodies to serotype-specific pneumococcal capsular polysaccharide (PS). A titer estimation method is needed for a large scale data generated by MOPA which uses serum more efficiently. In order to improve the reliability of OPA results, we developed a non-linear fitting method using Levenberg-Marquardt algorithm with a choice of robust procedure to estimate titers of OPA data. Performance of the proposed method was evaluated by comparing precision and accuracy of titer estimation with traditional methods used in the literature by analyzing six experimental data sets. Goodness of fit to experimental data for the two model based methods were also assessed. We conclude that the four parameters logistic model is a better choice for titer estimation of OPA data than other methods. The eleven-dilution design provided more information than the eight-dilution design for titer estimation but analyses based on two designs both support our choice of the four parameter logistic model than other methods for titer estimation of OPA data. Different experimental protocols did not affect the choice of titer estimation methods. A computer software using the statistical language R and Microsoft Excel was also developed to implement our calculation algorithm for OPA data.

e-mail: deliwang@uab.edu

BINARY TIME SERIES MODELING WITH APPLICATION TO KINETIC STUDIES IN MICROPIPETTE EXPERIMENTS

Ying Hung*, Georgia Institute of Technology
Chien-Fu Jeff Wu, Georgia Institute of Technology

Micropipette experimentation is a new biotechnological method developed to measure the kinetic rates of cell adhesion interactions which play an important role in tumor metastasis and cancer mutation. Traditional analysis of micropipette experiments assumes that the adhesion test cycles are independent Bernoulli trials. This assumption can often be violated in practice. In this paper, a multiple time series model incorporating random effects is developed to analyze the repeated adhesion tests. A goodness-of-fit statistic is introduced to assess the adequacy of distribution assumptions on the dependent binary data with random effects. The asymptotic distribution of the goodness-of-fitstatistic is derived. Application of the proposed methodology to some real data in an T-cell micropipette experiment reveals some interesting information on the dependency between repeated adhesion tests.

e-mail: yhung@isye.gatech.edu

THE EXCHANGEABLE LOGISTIC REGRESSION IN CORRELATED DATA

Xin Dang*, University of Mississippi
Hanxiang Peng, University of Mississippi

In this talk, we propose the exchangeable logistic regression generalizing the logistic regression. We introduce a new class of link functions based on completely monotone functions. From the Bayesian approach,the new class can be viewed as new prior distributions. We give an optimal estimating function (quasi-score) with the score of the binomial distribution as a special case. The proposed theories are applied to analyize a real data from teratology.

e-mail: xdang@olemiss.edu

RISK ANALYSIS USING GENERALIZED LINEAR MIXED EFFECTS MODELS

Matthew W. Wheeler*, National Institute for Occupational Safety and Health
A. John Bailer, Miami University

Risk researchers are frequently confronted with dose-response data collected from repeated experiments conducted at multiple laboratories. Often, there is significant lab-to-lab variability. In addition, experiments conducted in the same lab may be correlated. When considering a dichotomous response in this situation, the use of generalized linear mixed effects models (GLMM) may be appropriate. Although these models have been used in many settings, including developmental toxicity studies, we are not aware of their use to reflect components of variation when estimating risk. We apply GLMM to dose-response modeling in order to estimate the benchmark dose (BMD). We illustrate the estimation of the BMD from the U.S. EPA Region IX reference toxicity test database, and further study the BMD estimated from generalized linear models when a random effect is present but not included in the analysis, i.e. fitting a non-mixed effects model to true mixed effects data, through a simulation study. Preliminary findings suggest that accounting for these extra components of variation, when appropriate, improves accuracy in BMD estimation. The findings and conclusions in this abstract have not been formally disseminated by the National Institute for Occupational Safety and Health and should not be construed to represent any agency determination or policy.

e-mail: MWheeler@cdc.gov

COMPARISON OF DESIGNS FOR RESPONSE SURFACE MODELS WITH RANDOM BLOCK EFFECTS

Sourish C. Saha*, University of Florida
Andre I. Khuri, University of Florida

The purpose of this article is to compare designs for response surface models with a random block effect. To assess the quality of prediction associated with a given design, the scaled prediction variance is considered as a design criterion. The proposed approach is based on using quantiles of this design criterion on concentric surfaces within the experimental region. The dependence of these quantiles on the unknown value of the ratio of two variance components, namely, the ones for the block effect and the experimental error, is depicted by plotting the so-called quantile dispersion graphs (QDGs). These plots provide a clear assessment of the quality of prediction associated with a given design. A numerical example is presented to illustrate the proposed methodology.

e-mail: ssaha@stat.ufl.edu

94. LONGITUDINAL COUNT DATA

MODELING DELIRIUM PROGRESSION USING A NON-HOMOGENEOUS MARKOV PROCESS

Rebecca A. Hubbard*, University of Washington
Jesse R. Fann, University of Washington
Lurdes YT Inoue, University of Washington

Markov processes are useful in modeling multi-state disease processes especially under panel observation in which disease state is ascertained at irregularly spaced follow-up times. However, limited statistical methods are available for dealing with temporal non-homogeneity in which the rate of transition between disease states varies over the observation period. In this talk we present methods for modeling non-homogeneous Markov processes using panel data. We apply our methods to estimate delirium progression in a cohort of stem cell transplantation patients.

e-mail: rhubb@u.washington.edu

MODELING LONGITUDINAL COUNT DATA WITH THE POSSIBILITY OF DROPOUTS

Mohamed A. Alosh*, Food and Drug Administration

The primary endpoint for many clinical trials is counts, which frequently used to reflect the severity of the disease. Examples of this include the number of lesions in dermatologic indications such as acne, basal-cell carcinoma or actinic keratosis or the number of daily seizers in epilepsy trials. For such trials the primary endpoint is usually evaluated at successive intervals during the course of the trial. One of the features of such trials is heterogeneity among subjects in their baseline counts and/or change in their expected counts during of the course of the trial. Dropouts in such trials might be related to the severity of the disease as expressed in counts. We consider random effect approach for modeling such data along with their dropouts. We investigate the utility of the proposed approach through a simulation experiment and by application to clinical trial data.

e-mail: aloshm@cder.fda.gov

NONPARAMETRIC MODELS FOR MULTIVARIATE PANEL COUNT DATA

Li-Yin Lee*, University of Wisconsion, Madison
KyungMann Kim, University of Wisconsion, Madison

Panel count data consists of event counts that occur at unknown time points, but are recorded at regular intervals. Such data are often seen in clinical, health expenditure, reliability, and demographic studies. While methods of analyzing univariate panel count data have been proposed, the analysis of multivariate panel count data has not been investigated. This analysis is important when the studies include multiple types of recurrent events of interest. We propose nonparametric estimators of mean functions of counting processes for bivariate panel count data based on the maximum pseudo-likelihood estimators (Sun and Kalbfleisch, 1995 and Wellner and Zhang, 2000). The dependence of event processes is modeled by a frailty that takes the subject heterogeneity and the correlation of event types into account. We consider two Poisson models of bivariate counts: one with a shared gamma frailty and the other with a bivariate log normal frailty. The methods are illustrated by an analysis of data from a cancer chemoprevention trial on the effectiveness of difluoromethylornithine (DFMO) in preventing the non-melanoma skin cancer in patients with a history of prior skin cancer. Results from a limited simulation study will be reported in order to compare the two frailty models in terms of their operating characteristics qualitatively.

e-mail: liyinlee@wisc.edu

REGRESSION ANALYSIS OF MULTIVARIATE PANEL COUNT DATA

Xin He*, University of Missouri-Columbia
Xingwei Tong, Beijing Normal University
Jianguo Sun, University of Missouri-Columbia
Richard Cook, University of Waterloo

Panel count data frequently occur in periodic follow-up studies that concern recurrence rates of some recurrent events. Fields that produce such data include epidemiological studies, medical follow-up studies, reliability studies and tumorigenicity experiments. Multivariate panel count data arise if more than one type of recurrent events are of interest and this article discusses regression analysis of multivariate panel count data. For inference, we present a class of marginal mean models which leave the dependence structures for related types of recurrent events completely unspecified. To estimate regression parameters, some estimating equations are developed and the resulting estimates are consistent and asymptotically normal. Simulation studies show that the proposed estimation procedures work well for practical situations and the methodology is applied to a motivated example.

e-mail: xhw4c@mizzou.edu

COVARIANCE STRUCTURES FOR A MIXED EFFECT MARKOV MODEL FOR REPEATED BINARY OUTCOMES

Robert J. Gallop*, West Chester University

In many areas of research, repeated binary measures often represent a two-state stochastic process, where individuals can transition among two states. In a behavioral or physical disability setting, individuals can flow from susceptible or subthreshold state, to an infectious or symptomatic state, and back to a subthreshold state. Quite often the transition among the states happens in continuous time but is observed at discrete, irregularly spaced timepoints which may be unique to each individual. Methods for analyses of such data are typically based on the Markov assumption. Cook (Biometrics, 1999; 55: 915-920) introduced a conditional Markov model that accommodates the subject-to-subject variation in the model parameters with random effects. Pair of random effects were chosen as independent gamma random variables. This choice guaranteed closed form solution to the likelihood function. This research extends the random effect design to accommodate correlation and heterogeneity in variance between the random effect terms. This methodology is illustrated by applications to a data set from a parasitic field infection survey and a data set from a cocaine treatment study. Clinical and statistical interpretations of the random effects and correlation structure will be discussed.

e-mail: rgallop@wcupa.edu

SEMIPARAMETRIC ESTIMATION METHODS FOR PANEL COUNT DATA USING MONTONE POLYNOMIAL SPLINES

Minggen Lu*, University of Iowa
Ying Zhang, University of Iowa
Jian Huang, University of Iowa

We study semiparametric likelihood-based methods for panel count data using monotone polynomial splines with proportional mean model. The generalized Rosen algorithm, proposed by Zhang & Jamshidian (2004), is used to compute the estimators of both the baseline mean function and the regression parameter. We show that the proposed spline likelihood-based estimators of baseline mean function are consistent and their rate of convergence can be faster than cubic root of n. The asymptotic normality of estimators of the regression parameter is also established. Simulation studies with moderate samples show that the spline estimators of baseline mean function are more efficient both statistically and computationally than their alternatives proposed in Wellner & Zhang (2005). A real example from a bladder tumor clinical trial is used to illustrate the methods.

e-mail: minggen-lu@uiowa.edu

MARGINALIZED RANDOM EFFECTS MODELS FOR LONGITUDINAL ORDINAL DATA

Keunbaik Lee*, University of Florida
Michael J. Daniels, University of Florida

Random effects are often used for generalized linear models to explain the serial dependence for longitudinal categorical data. Heagerty (1999) has proposed marginalized random effects models for the analysis of longitudinal binary data to permit likelihood-based estimation of marginal regression parameters. In this paper, we propose a new model to extend Heagerty's work to accommodate longitudinal ordinal data. Maximum marginal likelihood estimation is proposed utilizing Quasi-Newton algorithms with Monte Carlo integration of the random effects. Our approach is applied to analyze quality of life data from a recent colorectal cancer clinical trial.

e-mail: lee@stat.ufl.edu

95. MICROARRAY ANALYSIS III

APPLICATIONS OF SPACINGS IN GENOMICS

Stanley B. Pounds*, St. Jude Children's Research Hospital
Cheng Cheng, St. Jude Children's Research Hospital

Spacings, the distance or time between consecutive occurrences of a specific event, have proven to be a useful tool to predict properties of DNA clone libraries. Therefore, spacings may also be useful for other applications in genomics research. We have developed a statistical method that uses the spacings between consecutive calls of loss-of-heterozygosity (LOH) generated by a SNP microarray to infer regions of LOH for specific samples. The proposed method is a useful data-reduction strategy, simplifies the multiple-testing problem, leads to theoretically conservative FDR estimates, and is robust against the inclusion of poor-quality data. In a recent study of LOH in therapy-related leukemia, the proposed method identified more confirmed regions of LOH than did a widely-used method based on hidden Markov modeling. Conceptually, the method may be generalized to infer regions corresponding to other genomic features as well.

e-mail: stanley.pounds@stjude.org

A METHOD FOR GENE SET ENRICHMENT ANALYSIS OF TOXICOGENOMICS DATA

Rongheng Lin*, National Institute of Environmental Health Science, NIH
Shuangshuang Dai, Alpha-Gamma Technologies, Inc.
Richard D Irwin, National Institute of Environmental Health Science, NIH
Alexandra N. Heinloth, National Institute of Environmental Health Science, NIH
Gary A. Boorman, National Institute of Environmental Health Science, NIH
Bhanu P. Singh, National Institute of Environmental Health Science, NIH
Leping Li, National Institute of Environmental Health Science, NIH

Identifying differentially expressed genes or sets of genes remains an important task for microarray studies. Recently, gene set enrichment analysis (GSEA) has gained recognition as a way to identify biological pathways/processes that are associated with a phenotypic endpoint. In GSEA, a local statistic is used to assess the association between the expression level of a gene and the value of a phenotypic endpoint. Then GSEA combines these gene-specific local statistics to evaluate association for pre-selected sets of genes. Commonly used local statistics include t statistics for binary phenotypes and correlation coefficients that assume a linear or monotone relationship between a continuous phenotype and gene expression level. Methods applicable to continuous non-monotone relationships are needed. Herein we propose to use as the local statistic the square of multiple correlation coefficient from fitting natural cubic spline models to the phenotype-expression relationship. Next, we incorporate this association measure into the GSEA framework to identify significant gene sets. Furthermore, we describe a procedure for inference across multiple GSEA analyses. We illustrate our approach using gene expression and liver injury data from liver and blood samples from rats treated with eight hepatotoxicants under multiple time and dose combinations.

e-mail: linr2@niehs.nih.gov

FEATURE SELECTION WITH A SUPERVISED PARTITIONING SCHEME

Yaomin Xu*, Case Western Reserve University and The Cleveland Clinic Foundation
Jiayang Sun, Case Western Reserve University

High throughput genomic studies, such as those conducted using DNA arrays, now provide millions of data points. To address the problem of data mining of such data, we describe a novel data mining method based on a supervised partitioning scheme. The method identifies subsets of features (e.g. genes) that best agree with one or more properties of the samples. The algorithm is applicable to the studies such as gene selection in gene expression studies, SNP selection in whole genome association study and QTL type analysis, or generally any type of feature selection problem.

e-mail: xuy3@ccf.org

ANALYZING GENE EXPRESSION DATA FROM SOME NON-MICROARRAY TRANSCRIPTION PROFILING TECHNIQUES

Sujay Datta*, Texas A&M University

For more than a decade, global gene expression assays have enabled scientists to measure expression levels or test hypotheses based on entire genomes of organisms. Microarrays (cDNA or oligonucleotide) have been widely used for this purpose and data analysis methods for microarrays are well-documented. However, some alternative transcription profiling techniques have recently been developed that are based on transcript counting (i.e., counting the number of individual mRNA molecules produced from each gene). Two examples are SAGE (Serial Analysis of Gene Expression) and MPSS (Massively Parallel Signature Sequencing). The latter has some advantages over the microarray-based technology in terms of sensitivity, data format and need for prior identification/characterization of genes. But like microarrays, MPSS data are subject to noise from various sources that make a statistical analysis challenging. Here we briefly summarize various approaches to analyzing expression data from these count-based high-throughput techniques.

e-mail: sdatta@stat.tamu.edu

CLUSTERING THRESHOLD GRADIENT DESCENT REGULARIZATION: WITH APPLICATIONS TO MICROARRAY STUDIES

Shuangge Ma*, Yale University
Jian Huang, University of Iowa

An important application of microarray technology is to discover important genes and pathways that are correlated with clinical outcomes such as disease status and survival. While a typical microarray experiment surveys gene expressions on a global scale, there may be only a small number of genes that have significant influence on a clinical outcome of interest. In addition, expression data have cluster structures and the genes within a cluster have coordinated influence on the response, but the effects of individual genes in the same cluster may be different. For microarray studies with smooth objective functions and well defined cluster structure for genes, we propose a clustering threshold gradient descent regularization (CTGDR) method, for simultaneous cluster selection and within cluster gene selection. We apply this method to regression models for binary classification and censored survival data with microarray gene expression data as covariates, assuming known cluster structures of gene expressions. Compared to the standard TGDR and other regularization methods, the CTGDR takes into account the cluster structure and carries out feature selection at both the cluster level and within-cluster gene level.

e-mail: shuangge.ma@yale.edu

A COPULA APPROACH TO MISSING DATA IN GENETIC ASSOCIATION STUDIES

Gina M. D'Angelo*, University of Pittsburgh
Eleanor Feingold, University of Pittsburgh
Lisa A. Weissfeld, University of Pittsburgh

A limited amount of research has focused on association studies with missing genotype data. Typically, missing data is assumed to be missing completely at random (MCAR) and complete case analysis is the method of choice resulting in biased and inefficient results. When fitting models with genotypes at several markers as simultaneous predictors, the complete case approach can lead to significant reduction in sample size. To address the problem where the SNPs are missing at random (MAR), we suggest using a copula approach to improve the coefficient estimates and their standard errors in a logistic regression model. An advantage of employing the copula is that the joint distribution of the outcome and covariate is specified, and since the covariates are missing their distributions must be specified. An added benefit to the copula approach is that it can handle nonmonotonic data. We compare the copula approach to multiple imputation and complete case analysis.

e-mail: gmdst17@pitt.edu

96. QUANTITATIVE TRAIT LOCI

A LIKELIHOOD-BASED METHOD FOR MAPPING QUANTITATIVE TRAIT LOCI THAT CONTROL LONGITUDINAL BINARY RESPONSES

Hongying Li*, University of Florida
Ramon C. Littell, University of Florida
Rongling Wu, University of Florida

Because of the fundamental relevance of function-valued or longitudinal traits for agricultural, biomedical and health sciences, their genetic mapping with molecular markers has emerged as one of the most rigorous and important areas in genetic research. Most current approaches for mapping longitudinal traits assume that these traits vary in a continuous pattern with an underlying normal distribution. However, there is also a group of traits of scientific interest, such as disease status or mortality, which are binary in nature. In this talk, we will present a statistical method for the identification of quantitative trait loci (QTL) that encode longitudinal binary responses. The framework of this model is constructed within the context of the approach, as advocated by Fitzmaurice & Laird (1993), in which the association between longitudinal binary responses is modeled in terms of conditional log odds-ratios. We develop an iterative Newton-Raphson procedure for obtaining the maximum likelihood estimates of the genetic effects of QTL. Results from simulation studies reveal the statistical properties of the model and its usefulness in practical mapping projects.

e-mail: hyli@ufl.edu

NONPARAMETRIC MODELING OF LONGITUDINAL COVARIANCE STRUCTURE IN FUNCTIONAL MAPPING OF QUANTITATIVE TRAIT LOCI

John Stephen F. Yap*, University of Florida
Rongling Wu, University of Florida

Estimation of the covariance structure of longitudinal processes is a fundamental prerequisite for the practical deployment of functional mapping designed to study the genetic regulation and network of quantitative variation in dynamic complex traits. We present a nonparametric approach to global estimation of the longitudinal covariance structure of a quantitative trait measured repeatedly at times. Spline approximation is applied to unconstrained parameterization of the covariance matrix obtained via modified Cholesky decomposition as was done by Huang et al. (2006). This approach is embedded within the framework for functional mapping to genomewide scan for the existence of quantitative trait loci underlying a dynamic trait, leading to enhance the breadth of use of this mapping method while preserving its biological relevance. A live example from a mouse genome project is analyzed to illustrate the methodology. Extensive simulation studies are performed to reveal the statistical properties and advantages of the nonparametric covariance modeling for functional mapping.

e-mail: jyap@stat.ufl.edu

MODELING OF MULTIPLE TRAITS FOR GENOME-WIDE EPISTATIC QTL MAPPING

Samprit Banerjee*, University of Alabama at Birmingham
Nengjun Yi, University of Alabama at Birmingham

Identifying complex epistatic quantitative trait loci (QTL) poses a bewildering challenge to contemporary statistical geneticists. The majority of the existing methods focus their attention on a single trait, even though typically data on more than one phenotype are collected. In this paper, we present statistical models and methods for performing multiple trait analysis on complex traits in experimental crosses from two inbred lines using a composite model space approach to develop Bayesian model selection framework to map genome-wide epistatic QTL . The joint analysis would take into consideration the correlation structure of multiple traits by invoking a Seemingly Unrelated Regression (SUR) model. SUR modeling allows different traits to have different genetic regression models (seemingly unrelated) but strings them all together by providing a correlation structure to the residuals. The joint analysis had several advantages over single trait analysis, including the expected improvement in statistical power to detect QTL and in precision of parameter estimation. It can also provide us with a greater insight in the nature of genetic correlations in certain regions of the genome by testing pleiotropy (one QTL affecting different traits) and pleiotropy vs. close linkage (multiple nearby non-pleiotropic QTL).

e-mail: samban@uab.edu

A SCORE TEST FOR LINKAGE ANALYSIS OF ORDINAL TRAITS

Rui Feng*, University of Alabama at Birmingham
Heping Zhang, Yale University

Statistical methods for linkage analysis are well established for both binary and quantitative traits. However, numerous diseases including cancer and psychiatric disorders are rated on discrete ordinal scales. To analyze pedigree data with ordinal traits, we recently proposed a latent variable model which has higher power to detect linkage using ordinal traits than methods using the dichotomized traits. The challenge with the latent variable model is that the likelihood is usually very complicated and as a result, the computation of the likelihood ratio statistic is too intensive for large pedigrees. In this paper, we derive a computationally efficient score statistic based on the identity by decent (IBD) sharing information between relatives. Using simulation studies, we examined the asymptotic distribution of the test statistic and the power of our proposed test under various genetic models. We compared the computing time as well as power of the score test with other alternatives. We then applied our method for the Collaborative Study on the Genetics of Alcoholism (COGA) and performed a genome scan to map susceptibility genes for alcohol dependence. We found a strong linkage signal on chromosomes 4.

e-mail: rfeng@ms.soph.uab.edu

ANALYZING AND MODELING DICHOTOMOUS TRAITS IN LARGE COMPLEX PEDIGREES

Charalampos Papachristou*, University of Chicago
Carole Ober, University of Chicago
Mark Abney, University of Chicago

Although it is believed that many common complex disorders have a genetic basis, attempts to unravel the transmission mechanism governing such traits have met with limited success. It has been suggested that isolated founder populations with large, known pedigrees may be advantageous for complex trait mapping. However, their utility has been moderated by the extreme computational intensity involved in the analysis of such pedigrees as a whole. We are proposing a likelihood method for modeling the transmission of dichotomous traits that can handle large pedigrees in a fast and efficient way. Using generalized linear mixed models, we extend the method of Abney et al. (2002) for mapping quantitative trait loci (QTLs), to accommodate binary traits. The high dimensionality of the integration involved in the likelihood prohibits exact computations. We show that one can overcome this hurdle and obtain the maximum likelihood estimates of the model parameters through the use of an efficient Monte Carlo expectation maximization (MCEM) algorithm. Analysis of data from a 13-generation pedigree consisting of 1,653 Hutterites, focusing on the diabetes phenotype, reveals evidence for the existence of at least one locus with dominance mode of trait transmission.

e-mail: babis@uchicago.edu

AN APPROXIMATE BAYESIAN APPROACH FOR QUANTITATIVE TRAIT LOCI ESTIMATION

Yu-Ling Chang*, University of North Carolina - Chapel Hill
Fred A. Wright, University of North Carolina - Chapel Hill
Fei Zou, University of North Carolina - Chapel Hill

Bayesian approaches have been widely used in Quantitative Trait locus (QTL) linkage analysis in experimental crosses, and have advantages in interpretability and in constructing parameter probability intervals. Most existing Bayesian linkage methods involve Monte Carlo sampling, which is computationally prohibitive for high-throughput applications such as eQTL analyses. In this paper, we present a Bayesian linkage model that offers highly interpretable posterior densities for linkage, without the need for Bayes factors in model selection. For this model, we develop Laplace approximations for integration over nuisance parameters in backcross and F2 intercross data. Our approach is highly accurate and very fast compared with alternative grid search approach, importance sampling, gibbs sampling and adaptive quadrature, and is suitable for high-throughput applications. Our results also provide insight into the connection between the LOD curve and the posterior probability for linkage.

e-mail: changy@email.unc.edu

QTL DETECTION FOR ZERO-INFLATED POISSON TRAITS WITH RANDOM EFFECTS

Rhonda R. DeCook*, University of Iowa
Dan Nettleton, Iowa State University

QTL detection methods were initally introduced for traits that were normally distributed. Methods for some non-normal traits, such as binary and ordinal traits, have also been developed in recent years. In this talk, we present a method for detecting QTL for zero-inflated Poisson (ZIP) traits when random effects are present. Though normality-based methods can be reasonably applied to some ZIP distributions, others that are more non-normal require a different approach. The method we present can be applied to any ZIP distribution and accounts for random effects arising due to the study design. Using simulation, we compare our method to two existing applicable approaches. The method is illustrated using QTL data collected on two ecotypes of the Arabidopsis thaliana plant where the trait of interest was shoot count, and counts were taken on different days and in different groups.

e-mail: rdecook@iastate.edu

97. NON- AND SEMI-PARAMETRICS

ESTIMATING LINEAR FUNCTIONALS OF INDIRECTLY OBSERVED INPUT FUNCTIONS

Eun-Joo Lee*, Millikin University

We consider the usual estimator of a linear functional of the unknown input function in indirect nonparametric regression models. The unknown regression function which is the parameter of interest, is infinite dimensional. Since the function in a separable Hilbert space has a Fourier expansion in an orthonormal basis, the Fourier coefficients will be estimated. It is surprising to see that the traditional estimator of the Fourier coefficients is not asymptotically efficient according to Hajek-LeCam convolution theorem. Since this estimator, however, is sqrt (n) – consistent, it can be improved in an asymptotic sense. The possible improvement of this estimator will be discussed. We will also compare the improved estimator with the traditional estimator through simulation studies.

e-mail: elee@millikin.edu

THE MATCHED PAIRS SIGN TEST USING BIVARIATE RANKED SET SAMPLING

Hani M. Samawi*, Georgia Southern University
Mohammad F. Al-Saleh, Yarmouk University
Obaid A. Al-Saidy, Sultan Qaboos University

The matched pairs sign test using bivariate ranked set sampling (BVRSS) is introduced and investigated. We show that this test is asymptotically more efficient than its counterpart sign test based on a bivariate simple random sample (BVSRS). The asymptotic null distribution and the efficiency of the test are derived. The Pitman asymptotic relative efficiency is used to compare the asymptotic performance of the matched pairs sign test using BVRSS versus using BVSRS. For small sample sizes, the bootstrap method is used to estimate P-values. Numerical comparisons and real data are used to gain insight about the efficiency of the BVRSS sign test compared to the BVSRS sign test. Our numerical and theoretical results indicate that using BVRSS for the matched pairs sign test is substantially more efficient than using BVSRS.

e-mail: hsamawi@georgiasouthern.edu

ESTIMATION IN NONPARAMETRIC MODELS WITH MISSING OUTCOME AT RANDOM

Lu Wang*, Harvard University School of Public Health
Xihong Lin, Harvard University School of Public Health
Andrea Rotnitzky, Harvard University School of Public Health

In this paper, we considered estimation in univariate nonparametric mean models with missing outcome at random for cross-sectional studies. We modified the traditional kernel generalized estimating equations (GEE) to accommodate the missing data. Three methods will be proposed in this paper, including naive kernel GEE, inverse probability weighted (IPW) kernel GEE and augmented inverse probability weighted (AIPW) kernel GEE. We will present the asymptotic properties of the corresponding estimators at convergence and compare them in terms of consistency and asymptotic efficiency. The double robustness of AIPW kernel GEE estimator will be discussed as well as potential ways to enhance its efficiency. Simulations were conducted to describe the finite sample performance of the proposed methods and we illustrated the methods through an application example.

e-mail: luwang@hsph.harvard.edu

ON ROBUSTNESS OF CLASSIFICATION BASED ON DEPTH TRANSVARIATION

Nedret Billor*, Auburn University
Ash Abebe, Auburn University
Asuman S. Turkmen, Auburn University
Nudurapati Sai, Auburn University

Depth transvariation based classification is a new nonparametric classification technique based on classifying a multivariate data point through maximizing the estimated transvariation probability of statistical depths. In this paper we will study the effect of outliers on classification based on depth transvariation and compare the performance of this new classification method based on depth transvariation with the classical classification techniques in the presence of outliers. Various simulations and real examples are given to reveal the robustness of classification based on depth transvariation.

e-mail: billone@auburn.edu

TESTING THE EQUALITY OF MEAN FUNCTIONS FOR CONTINUOUS TIME STOCHASTIC PROCESSES

Yolanda Munoz Maldonado*, University of Texas-Houston, School of Public Health

A test statistic is developed to determine whether two or more groups of functions, derived from continuous time stochastic processes, belong to the same population. An approximate chi-square test is proposed for comparing n group means. The test statistic considered in this paper uses a quadrature approximation to the L_2 norm. The proposed test statistic is shown to be a U-statistic and its asymptotic properties are formulated. A small Monte Carlo study is conducted and we illustrate its application with a real data set.

e-mail: Yolanda.M.Munoz@uth.tmc.edu

SEMIPARAMETRIC VARYING COEFFICIENT PERIODIC REGRESSION MODEL

Yingxue Cathy Liu*, Texas A&M University
Naisyin Wang, Texas A&M University

Circadian time keeping mechanisms play a key role in how living organisms adopt to daily fluctuations in light and temperature. Scientists are interested in identifying endogenous oscillator genes and evaluating their functions. This is often achieved by studying circadian rhythmic measurements, such as gene expression levels, from subjects of different genetic strains under different treatment conditions. Free-running periods (FRPs), the corresponding phase and the best-fit curves of gene expression levels were most commonly estimated by fast Fourier transform-non-linear least squares method (FFT-NLLS; Plautz et al. 1997). However, the outcomes are not always satisfactory. The variation in the estimated phase could lead to an inconclusive finding. Alternatively, we propose to tackle this challenging issue by fitting the observations with a semiparametric varying coefficient periodic regression model. Our method can accommodate measurements from one or multiple subjects. We studied and compared several alternative procedures and established their asymptotic properties. Our simulation outcomes show that substantial gains over the use of FFT-NLLS can be achieved by applying our methods. The findings were further verified in a study of a cyanobacterial circadian data set.

e-mail: yliu@stat.tamu.edu

EFFICIENT ESTIMATION OF POPULATION QUANTILES IN GENERAL SEMIPARAMETRIC REGRESSION MODELS

Arnab Maity*, Texas A&M University

This paper considers a large family of semiparametric regression models, in which the main interest is on estimating population quantiles. Special cases of our framework include generalized partially linear models, generalized partially linear single index models, structural measurement error models and many others. We derive plug-in kernel-based estimators of the population quantiles of random variables and derive their asymptotic distribution. We also establish the semiparametric efficiency of these estimators under mild assumptions. We apply our methodology in an example in nutritional epidemiology where estimation of quantiles of usual intake of various dietary components are of primary interest. The generalization to the important case that responses are missing at random is also addressed.

e-mail: amaity@stat.tamu.edu

98. KINETIC AND OTHER MODELING, INCLUDING PK/PD

VIRAL KINETIC MODELING OF HBV DNA

Larry F. Leon*, Bristol-Myers Squibb Pharmaceutical Research Institute
Anne Cross, Bristol-Myers Squibb Pharmaceutical Research Institute

Viral kinetic modeling has been widely applied to studying the predictive value of early response to HCV therapy on sustained viral response. The first decline in HCV RNA levels is generally steep and is presumed to result from blockade of virion production and release. The second elimination phase is slower and is presumed to reflect immune-mediated clearance of HCV-infected cells. This biphasic profile is modeled by an exponential model with a functional form in time that reflects the biphasic pattern. The model was subsequently used to describe the early effects of antiviral therapy on HBV DNA. However, the model does not accommodate patients whose early HBV DNA response does not rapidly decline and this can lead to problems in estimation. While the exponential model is useful for describing viral kinetic parameters (e.g., efficacy), simple regression models can be used when describing the viral pattern over time is the primary goal. We propose a spline regression model which represents the biphasic pattern by two lines that are joined together at a single knot (transition point between phases). The spline model is easy to implement, does not suffer from estimation problems, and lends itself easily to treatment comparisons. We apply the spline model to the analysis of HBV clinical trial data recently presented at AASLD (2006, poster 014) and assess the performance of the models in simulation studies.

e-mail: larry.leon@bms.com

A NEW STEADY STATE ASSESSMENT METHOD BASED ON NONLINEAR MIXED EFFECT MODELING

Quan Hong*, Eli Lilly and Company
Eyas Abu-Raddad, Eli Lilly and Company
Didier Renard, Eli Lilly and Company

Assessing steady state attainment of drug plasma concentration is an important task of drug development. Current methods in practice rely on ANOVA-based difference test or equivalence test which aim to show that mean drug plasma concentration have no significant difference between days at steady-state. A new method is proposed to assess steady state attainment, which is built on nonlinear mixed effect modeling of pharmacokinetic drug accumulation under multiple dosing administration. A hypothesis testing procedure is established to determine whether steady state is achieved at interested time points. In addition, the actual time to reach steady state can be estimated at both the mean and individual level. Extensive simulation studies are conducted to demonstrate the reliable performance of the proposed method and its advantages compared to existing methods. The robustness of the method is also examined under various pharmacokinetic assumptions to prove the method's wide applicability in drug development.

e-mail: quanhong@gmail.com

BAYESIAN SEMI-PARAMETRIC ANALYSIS OF PK/PD MECHANISTIC MODELS

Michele Guindani*, University of Texas-M.D. Anderson Cancer Center
Peter Müller, University of Texas-M.D. Anderson Cancer Center
Gary Rosner, University of Texas-M.D. Anderson Cancer Center

Joint PK/PD mechanistic models allow a description of the entire course of the individual physiological response after treatment with anticancer drugs while using the time course of plasma concentration as input into the model. The use of flexible hierarchical non-parametric bayesian models has proved helpful in identifying within a large groups of patients those characteristics that significantly alter a drug's disposition and a patient's response to the drug. In particular, Dirichlet Process Mixture models have been successfully used to characterize patients specific profiles (Muller and Rosner, 1997) and combine information across related studies (Muller et al., 2002). We extend those studies, considering explicit PK/PD mechanistic models to describe patient profiles. Joint inference for the parameters involved in the PK/PD dynamics can be obtained. Patients characteristics can be incorporated and these characteristics can help in identifying therapeutic subgroups and improve patient predictions.

e-mail: mguindani@gmail.com

COVARIATE MODEL FOR STUDYING THE PHARMACOGENETIC ARCHITECTURE OF DRUG RESPONSE BY ALLOMETRIC SCALING

Wei Hou*, University of Florida
Rongling Wu, University of Florida

The identification of specific genetic variants that contribute to interpersonal variability in response to drugs is a first step for the design of personalized medications for the control and prevention of a disease based on a patient's genetic makeup. However, identifying and mapping genes for differentiation in drug response is challenged in statistics because the effects of a drug varying across different dose levels present a longitudinal issue. Also, the efficiency and toxicity of a drug are complicated by a number of non-genetic factors, such as the patient's age, sex, nutritional status, renal or liver function and metabolic rate. This talk will focus on the development of a unifying model for the identification of DNA sequence variants that determine longitudinal drug response by implementing the effects of different non-genetic factors as covariate within the context of a mixture model. In particular, a universal biological principle that describes the scaling relationship between pharmacodynamic metabolism and body size will be integrated into the model for mapping drug response. This integrative model allows for the estimation and test of DNA variants that pleiotropically affect the efficacy of a drug and its biological response to a patient's body size. A real example from the pharmacogenomic study of heart disease was used to validate the utilization of the model. The statistical behavior of the model is investigated through Monte Carlo simulation studies.

e-mail: whou@biostat.ufl.edu

A BAYESIAN APPROACH FOR DIFFERENTIAL EQUATION MODELS WITH APPLICATION TO HIV DYNAMIC STUDIES

Tao Lu*, University of South Florida
Yangxin Huang, University of South Florida

A virologic marker, viral load, is currently used to evaluate antiviral therapies in AIDS clinical trials. This marker can be used to assess the antiviral potency of therapies, but is easily affected by drug exposures, drug resistance and other factors during the long-term treatment process. HIV dynamic studies have significantly contributed to the understanding of HIV pathogenesis. However, the models of these studies are used to quantify short-term HIV dynamics, and are not applicable to describe long-term virologic response to ARV treatment. Pharmacokinetics, drug resistance and imperfect adherence to prescribed antiviral drugs are important factors explaining the resurgence of virus. To better understanding of the factors responsible for the virologic failure, this paper develops a mechanism-based nonlinear differential equation models for characterizing long-term viral dynamics. A Bayesian nonlinear mixed-effects modeling approach is investigated to estimate dynamic parameters. The correlations of baseline factors and virologic/immunologic responses with estimated dynamic parameters are explored and some biologically meaningful correlation results are presented. These results suggest that viral dynamic parameters may play an important role in understanding HIV pathogenesis, designing new treatment strategies for AIDS patients.

e-mail: tlu1@hsc.usf.edu

A PARAMETRIC MODEL FOR A CHOICE RT EXPERIMENT

Jennifer L. Asimit*, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto
Willard J. Braun, University of Western Ontario
William A. Simpson, Simulation and Modelling Section, Defence Research and Development Canada

A parametric model for data from a point process version of a choice reaction time experiment is introduced, and used to make inferences on the eye-brain-hand system. Nonparametric intensity estimates are used with the intensity expressions under the model, to obtain estimates of the model parameters. The model is fitted to real choice reaction time experiments run at nine different stimulus rates.

e-mail: jprokop@stats.uwo.ca

CLUSTERED DATA MODELS FOR THE INFARCTED HEART

Raymond G. Hoffmann*, Medical College of Wisconsin
Nicholas Pajewski, Medical College of Wisconsin
Xiaoguang Zhu, Medical College of Wisconsin
Ming Zhao, Medical College of Wisconsin
Raymond Migrino, Medical College of Wisconsin

2D strain echocardiography was used to dynamically subdivide rat hearts into 6 correlated segments after a rat ischemia-reperfusion study of myocardial infaction. Both the correlation among the segments and the non-normal distribution of the per cent infarction data needed to be modeled to determine the ability of echocardiographic strain measurements to predict a threshold area of myocardial infarction. Several approaches to this problem are compared: bootstrapped variance estimates using an approximate exponential family GEE model, use of the simplex distribution which has an exponential dispersion model similar to an exponential family and a hierarchical approach that allows the parameters to vary between animals. The methods are tested with simulated data with known correlation structure and mean structure as well as with the rat data.

e-mail: hoffmann@mcw.edu

99. ERROR PROBABILITIES, SAMPLE SIZE, AND MULTIPLICITY: LIKELIHOOD METHODS

FOLLOW THE LIKELIHOOD PRINCIPLE AND OBSERVE MISLEADING EVIDENCE LESS OFTEN: IMPLICATIONS FOR STUDIES WITH MULTIPLE ENDPOINTS

Jeffrey D. Blume*, Brown University

I have argued elsewhere, on philosophical grounds, that likelihood ratios are a better tool for measuring the strength of statistical evidence than p-values. In this talk, I'll take a different approach and show that they also have better frequentist properties – even in cases with multiple endpoints. This is true despite the fact that no adjustments are used to control the probability of observing misleading evidence. We'll see that the probability of making an 'error' (either Type I or Type II) is minimized when using likelihood ratios as a measure of the strength of evidence and that this result holds even when the Type I error of hypothesis testing is adjusted to avoid inflation. Hence, likelihood ratios are less likely to be misleading than p-values, even in trials with multiple endpoints. A key component of this result is that the probability of observing misleading evidence (i.e., the likelihood version of the Type I error) converges to zero as the sample size increases. Now even the overall chance of being misled (which increases with each additional endpoint) can be driven to zero by simply increasing the sample size. The obvious advantage here is that clumsy adjustments of error probabilities can be avoided without sacrificing frequentist performance.

e-mail: jblume@stat.brown.edu

EVIDENTIAL LIKELIHOOD APPROACH TO MULTIPLE TEST ADJUSTMENTS: APPLICATION TO LINKAGE ANALYSIS

Lisa J. Strug*, Columbia University
Susan E. Hodge, Columbia University, and New York State Psychiatric Institute

The "multiple testing problem" currently bedevils the field of genetic epidemiology. The conventional solution to this problem relies on the classical Neyman-Pearson statistical paradigm and involves adjusting one's error probabilities. This adjustment is, however, problematic because in the process of doing that, one is also adjusting one's measure of evidence. Investigators have actually become wary of looking at their data, for fear of having to adjust the strength of the evidence they observed at a given locus on the genome every time they conduct an additional test. Elsewhere, we have presented the "evidential paradigm," to be used when planning and evaluating linkage studies. The evidential paradigm uses the lod score (as opposed to a p-value) as the measure of evidence, and provides new, alternatively defined error probabilities (analogous to Type I and Type II error rates), i.e., probabilities of being misled. We showed how this paradigm separates or decouples the two concepts of error probabilities and strength of the evidence. Here we apply the evidential paradigm to the multiple testing problem - specifically, in the context of linkage analysis. We provide a rigorous argument outlining how replication samples themselves provide appropriate adjustments for testing multiple hypotheses on a data set.

e-mail: ljs2109@columbia.edu

LIKELIHOOD-BASED INFERENCE FOR EARLY STOPPING IN SINGLE ARM CLINICAL TRIALS WITH TIME-TO-EVENT OUTCOMES

Elizabeth Garrett-Mayer*, Comprehensive Cancer Center-Johns Hopkins University
Sidney Kimmel Comprehensive Cancer Center-Johns Hopkins University

With the increase of cytostatic therapies in cancer research, clinical response rate (defined as tumor shrinkage) is no longer an adequate measure of treatment efficacy in many phase II oncology clinical trials. More often the outcome of interest is time to disease progression or death, yet investigators still want the ability to stop a trial early for futility. Commonly used early stopping designs, such as Simon's two-stage design, are not appropriate with time to event outcomes. Likelihood theory provides a framework for which sensible early stopping rules can be determined. Likelihood-based methods allow multiple looks at the data without the frequentist type I error penalty. Accumulating evidence can be evaluated as it is collected to determine if strong evidence for futility exists. Likelihood approaches require a parametric model, leading to a potential concern about model misspecification. However, with careful design considerations, it is shown that early stopping determinations are robust. Keywords: survival analysis; early stopping rules; likelihood inference; Phase II studies

e-mail: esg@jhu.edu

100. INTRODUCTORY LECTURE SESSION: GROUP RANDOMIZED TRIALS

POWER IN GROUP-RANDOMIZED TRIALS

David M. Murray*, The Ohio State University

Group-randomized trials face two penalties not found in the usual randomized clinical trial: extra variation and limited degrees of freedom. Unless these penalties are considered during the design of the trial, the study can be severely underpowered. This talk will discuss the two penalties, review the most common design and analysis plans used in group-randomized trials, and summarize methods for power analysis for those studies. Particular attention will be given to methods to reduce the extra variation and increase the degrees of freedom.

e-mail: dmurray@sph.osu.edu

DYNAMIC BLOCK-RANDOMIZATION IN GROUP-RANDOMIZED TRIALS WHEN THE COMPOSITION OF BLOCKING FACTORS IS NOT KNOWN IN ADVANCE

Scarlett L. Bellamy*, University of Pennsylvania

We present a practical method for randomizing with blocking factors in group-randomized trials when the composition of blocking factors is not known in advance. For example, suppose the desired goal of an intervention study is to randomize units to one of two interventions, blocking on a dichotomous factor (e.g., gender), but the total number of units, and therefore number or composition, of males and females among those units assembled for randomization cannot be determined in advance. Since it is often the case in such settings that study personnel do not know which of the scheduled subjects will or will not show up, a dynamic randomization scheme is required to accommodate the unknown composition of the blocking factor once a group is assembled for randomization. These settings are further complicated when there are multiple such blocking factors. This talk highlights a practical randomization strategy that ensures the integrity of the randomization process in these settings.

e-mail: bellamys@mail.med.upenn.edu

THE USE OF GROUP SEQUENTIAL DESIGNS IN GROUP RANDOMIZED TRIALS

Ziding Feng*, Fred Huchinson Cancer Research Center
Charles E. McCulloch, Univeristy of California at San Francisco

Group Sequential Design (GSD) is a popular design in therapeutic trials due to ethic considerations as well as cost. However, GSD has not bee used in Group randomized trials (GRTs). There are situations in GRTs that require the option of early termination for ethic reason. GRTs are usually very expensive and therefore early termination when justified could also save precious resources. GRTs are usually used in the settings that require relatively long period of intervention, recruiting groups instead of individuals into the study, and design parameters include both the number of groups and the number of individual per group. These factors made the use of GSDs in GRTs challenging. We will discuss when a GSD should be used in GRTs, how to perform data analysis, and statistical design considerations for such studies. Examples are drawn from real GRTs. Key Words: Group Randomized Trials, Group Sequential Design.

e-mail: zfeng@fhcrc.org

THE MERITS OF BREAKING THE MATCHES; A CAUTIONARY TALE

Allan Donner*, University of Western Ontario

Matched-pair cluster randomization trials are frequently adopted as the design of choice for evaluating an intervention offered at the community level. However, previous research has demonstrated that a strategy of breaking the matches and performing an unmatched analysis may be more efficient than performing a matched analysis on the resulting data, particularly when the total number of communities is small and the matching is judged as relatively ineffective. The research concerning this question has naturally focused on testing the effect of intervention. However, a secondary objective of many community intervention trials is to investigate the effect of individual-level risk factors on one or more outcome variables. Focusing on the case of a continuous outcome variable, we show that the practice of performing an unmatched analysis on data arising from a matched-pair design can lead to bias in the estimated regression coefficient, and a corresponding test of significance which is overly liberal. However, for large-scale community intervention trials, which typically recruit a relatively small number of large clusters, such an analysis will generally be both valid and efficient.

e-mail: donner@schulich.uwo.ca

101. MARGINALIZED MODELS

MARGINALIZED MODELS FOR LONGITUDINAL ORDINAL DATA

Michael J. Daniels*, University of Florida
Keunbaik Lee, University of Florida

We construct marginalized models for longitudinal ordinal data using first and second order Markov dependence structures that exploit the ordering of the ordinal responses. Simulations are conducted to assess the impact of missing (at random) data on inferences in the presence of mis-specification of the dependence structure. Methods are illustrated on quality of life data from a recent colorectal cancer clinical trial. In this trial, there was considerable dropout due to both death and progression. We will discuss an approach to estimate the causal effect of the treatments in the presence of death/progression.

e-mail: mdaniels@stat.ufl.edu

MIXTURES OF MARGINALIZED MODELS FOR BINARY PROCESS DATA WITH CONTINUOUS-TIME DROPOUT

Li Su, Brown University
Joseph W. Hogan*, Brown University

We are interested in drawing inference about the effects of covariates on the marginal mean of a binary process that may be censored by dropout. Measurement times in a binary process are irregular across units or individuals, dropout occurs in continuous time, and dropout may be `informative' in the sense that it is dependent on missing values conditional on observed data. We propose a mixture of marginalized models for inference about the full-data covariate effects. A marginalized model is specified conditional on dropout time, with covariate effects and serial correlation parameters related to dropout time through smooth but unspecified functions. The full-data model is obtained by mixing over the dropout distribution, which also can be left unspecified. Inference is fully Bayesian and can be implemented in WinBUGS. The model is illustrated using longitudinal data on depression from an HIV cohort study.

e-mail: jwh@brown.edu

MARGINALIZED MODEL ESTIMATION UNDER A BIASED SAMPLING STUDY DESIGN: A CONDITIONAL LIKELIHOOD APPROACH

Jonathan S. Schildcrout*, Vanderbilt University
Patrick J. Heagerty, University of Washington

Longitudinal regression analysis of cohort study data can be used to examine the relationship between time-varying exposures and a binary response vector. However, over time, new hypotheses emerge that cannot be addressed with the information collected on the original cohort. For example, electronic medical and pharmacy records can be exploited to capture the effect of a daily, medication dose on the occurrence of daily adverse event symptoms among hospitalized patients; however, if interest lies in a dose by genotype interaction, additional (genotypic) information is required. In such a circumstance, analysis of the original cohort may not not feasible (due to costs associated with genotyping). In this talk, we discuss a study design in which individuals are selected based on values in their response vectors. We show that with targeted, outcome-dependent sampling, acquisition of subject-specific genotypic information is necessary only on a subset of individuals. We consider marginalized model parameter estimation, and a conditional maximum likelihood approach. In many realistic scenarios, conditional maximum likelihood estimates based on the subsample are highly efficient relative to maximum likelihood estimates based on the original cohort. Thus, with little to no loss of information, we may conserve study resources considerably.

e-mail: jonathan.schildcrout@vanderbilt.edu

102. NEW METHODS USING STATISTICAL GRAPHICS AS TOOLS IN ADDRESSING RESEARCH QUESTIONS

VISUALIZING GEOSPATIAL TIME SERIES USING MICROMAPS AND CONDITIONED COMPARISONS

Daniel B. Carr*, George Mason University
Chunling Zhang, George Mason University

This talk presents graphics templates and software for visualizing times series for regions such as states. The templates address tasks that include action prioritization, pattern discovery and hypothesis generation. The templates and examples address issues related to comparability, change blindness and cognitive complexity. The featured dynamic software, call temporal change maps (TCmaps) supports a variety of alternative views. One such view includes three time-ordered sequences of micromaps shown left to right in horizontal scrollable panels. The three sequences are binary micromaps highlighting regions with high, middle, and low values respectively. A dynamic three-class slider converts continuous values into high, middle and low classes. One or more additional series shows explicit differences between temporally adjacent binary micromaps. This addresses the issue of change blindness. The talk will presents live examples showing additional software features, and discuss merits of alternative approaches to addressing the tasks and issues.

e-mail: dcarr@gmu.edu

GRAPH-THEORETIC SCAGNOSTICS FOR PROJECTION PURSUIT

Heike Hofmann*, Iowa State University
Dianne H. Cook, Iowa State University
Hadley Wickham, Iowa State University

Finding ‘interesting’ projections in high dimensional space has a long tradition, yielding in methodological solutions such as the grand tour (Asimov 1958) or projection pursuit methods. While the grand tour walks through the high-dimensional space on a path that covers all possible lower dimensional projections, this path is optimized in projection pursuit methods for one specific optimality index. John and Paul Tukey (1985) suggested ‘scagnostics’ as a way to describe diagnostic properties of a scatterplot. Wilkinson et al (2005) recently extended this concept to a graph-theoretic approach. Using a set of nine indices, they used graph theoretic scagnostics for re-ordering scatterplots in a scatterplot matrix according to ‘skinniness’, ‘clumpiness’, number of outliers, etc. We propose an application of graph-theoretic indices as optimization criterion in projection pursuit: instead of a single criterion, a combination of these indices allows us to look for projections that are e.g. 80% ‘skinny’, 15% ‘clumpy’ with 0% outliers, etc. By allowing to change the index setting interactively, the analyst can guide the projection pursuit in more ways than the usual parameters allow.

e-mail: hofmann@iastate.edu

GRAPHICAL DISPLAYS FOR THE ANALYSIS AND PRESENTATION OF COMPLEX STATISTICAL DATA

Linda W. Pickle*, StatNet Consulting, LLC

As statistical data become more complex, in terms of numbers of both observations and variables being collected, the traditional forms of graphical displays become less useful. Overplotting can obscure any patterns in the data and bivariate scatter plots can fail to display important multivariate relationships. Geospatial data, in particular, is problematic, with one map representing sometimes millions of data points. Design of effective graphics depends at least in part on the technical skills of the expected audience as well as the medium of display (paper, static computer monitor, or interactive web display). In this talk, we review newer graphical methods proposed for complex data, focusing on, but not limited to, geospatial data. We will show how some of these methods are being combined to form new analytic tools for cancer surveillance research and to communicate statistical results to the public.

e-mail: lpickle@statnetconsulting.com

103. PANEL DISCUSSION: ROLE OF BIOSTATISTICIANS IN POLICY ISSUES

PANEL DISCUSSION

David Banks, Duke University
Barry Graubard, National Cancer Institute
Tom Louis, Johns Hopkins University
Sally C. Morton, RTI International

104. ENVIRONMENTAL STATISTICS

SPACE-TIME BAYESIAN SURVIVAL MODELING OF CHRONIC WASTING DISEASE IN DEER

Hae Ryoung Song*, University of South Carolina
Andrew Lawson, University of South Carolina
Dennis Heisey, University of Wisconsin-Madison
Erik Osnas, University of Wisconsin-Madison
Damien Joly, Fish and Wildlife Division, Alberta Sustainable Resource Development
Julie Langenberg, Wisconsin Department of Natural Resources

We propose a Bayesian hierarchical model to investigate the pattern of spatial and temporal variation in disease prevalence of chronic wasting disease (CWD) in white-tailed deer in Wisconsin. The aims of this study are to describe the geographical distribution of CWD and to assess the effect of demographic factors such as age and sex on prevalence of CWD. For the framework in our research, we adapt a Bayesian hierarchical survival model in which an unobservable latent process and an observation process are modeled at different hierarchical stages incorporating uncertainties in data and prior information. Since we do not observe the exact infection time, we consider our data censored and develop survival models based on the harvest time. Our model explains the variation of CWD using individual covariates, while accounting for spatial, temporal dependence and space-time interactions. It also allows the investigation of space-time spread of the disease by incorporating spatial, temporal, and space-time interaction infection rates. We apply our model to white-tailed deer data collected in Wisconsin over the years 2001-2006.

e-mail: hrsong@gwm.sc.edu

PROJECTED MULTIVARIATE LINEAR MIXED-EFFECTS MODELS FOR CLUSTERED ANGULAR DATA

Daniel B. Hall*, University of Georgia
Lewis Jordan, University of Georgia
Jinae Lee, University of Georgia
Jing Shen, IBM Thomas J. Watson Research Center

In this talk we extend the projected multivariate linear model of Presnell et al. (1998) to the clustered data case via the inclusion of cluster-specific random effects. For low dimensional random effects, we describe EM and Monte Carlo EM algorithms to accomplish maximum likelihood estimation in these models. For higher dimensional random effects, an approximate likelihood estimation approach is proposed. Finite sample properties of these estimation methods are investigated via simulation study. An example involving the microbril angle of loblolly pine, a property related to wood strength, is used to motivate and illustrate this class of models.

e-mail: dhall@stat.uga.edu

A DISTANCE-BASED CLASSIFIER WITH APPLICATION TO MICROBIAL SOURCE TRACKING

Jayson D. Wilbur*, Worcester Polytechnic Institute

Most classification rules can be expressed in terms of distances from the point to be classified to each of the candidate classes. For example, linear discriminant analysis classifies points into the class for which the sample Mahalanobis distance is smallest. However, dependence among these point-to-group distance measures is generally ignored. In this talk, a general classification rule will be defined which uses information about this dependence structure to improve classification. This work was initially motivated by the problem of microbial source tracking which aims to identify sources of fecal contamination in water resources based on genotypic and phenotypic variation in public health indicator organisms such as E. coli. An application of the proposed methodology to microbial source tracking will be presented.

e-mail: jwilbur@wpi.edu

A COPULA MODEL FOR SPATIAL PROCESS WITH AIR POLLUTION APPLICATION

Engin A. Sungur, University of Minnesota
Yoonsung Jung*, Kansas State University
Jong-Min Kim, University of Minnesota

By Sklar theorem in 1957, a multivariate distribution can be represented in terms of its underlying margins by binding them together using a copula function. In this paper, we will discuss depednece mechanisms of time space stochastic processes by using copulas and incorporate an archimedian copula into spatial statistics by introducing the concept of dependence covariates. The spatial copula model will be applied to fit the air pollution data.

e-mail: ysjung72@ksu.edu

CLUSTERING WITH LOGISTIC MIXTURE MODEL

Yongsung Joo*, University of Florida
Keunbaik Lee, University of Florida
Joonghyuk Kim, University of Florida
Sungtaek Yun, Korea University

We study the impacts of two representative agricultural activities, fertilizers and lime application, on water quality. Because of heavy usage of nitrogen fertilizers, nitrate (NO&#8722;3) concentration in water is considered as one of the best indicators for agricultural activity. The mixture of normal distributions has been widely applied to cluster waters based on log(NO&#8722;3) concentration. However, this method fails to yield satisfying results because it cannot distinguish low level fertilizer impact and natural background noise. To improve performance of cluster analysis, we introduce the logistic mixture of multivariate regressions model. In this approach, waters are clustered based on the relationships between major element concentrations and physicochemical variables, which are different in agrochemical-impacted waters and natural background waters. Such modelling approaches are often called model-based clustering methods.

e-mail: yjoo@phhp.ufl.edu

ON COMPARISON OF ESTIMATION METHODS IN CAPTURE-RECAPTURE STUDIES

Chang Xuan Mao, University of California-Riverside
Na You*, University of California-Riverside

Mixture model is a common method used to model individual heterogeneity in capture-recapture studies. Pledger (2000, 2005) advised using two-point mixture model, but Dorazio and Royle (2003) showed the advantages of beta-binomial mixture model for certain situations. A controversy arises due to the nonidentifiability of binomial mixture model. In order to make the controversy more clearly, a bias decomposition and the boundary problem are studied. Some numerical results are presented at last.

e-mail: nayou@ucr.edu

COMPARISONS OF SETS OF MULTIVARIATE TIME SERIES

Jaydip Mukhopadhyay*, University of Connecticut
Nalini Ravishanker, University of Connecticut

We discuss the problem of comparsion of several multivariate time series via their spectral properties. For two independent multivariate gaussian stationary time series, such a comparsion is made via a likelihood ratio test based on the estimated cross-spectra of the series. This is an extension of the maximum periodogram ordinate test developed in the literature to compare two independent univariate stationary time series. A simulation based critical value enables effective comparison of several such multivariate time series, and is useful in applications to biomedical time series or marketing or manufacturing time series.Extension of the spectral approach for comparison and clustering of nonlinear time series as well as nonstationary and categorical time series is discussed.

e-mail: jaystat@yahoo.com

105. EPIDEMIOLOGY USING BAYESIAN AND EMPIRICAL BAYES METHODS

SEMIPARAMETRIC BAYESIAN MODELS FOR SHORT-TERM PREDICTION OF CANCER MORTALITY SERIES

Kaushik Ghosh*, New Jersey Institute of Technology
Ram C. Tiwari, National Cancer Institute

We present two models for short-term prediction of the number of deaths that arise from common cancers in the United States. The first is a local linear model, in which the slope of the segment joining the number of deaths for any two consecutive time periods is assumed to be random with a nonparametric distribution, which has a Dirichlet process prior. For slightly longer prediction periods, we present a local quadratic model. This extension of the local linear model includes an additional `acceleration' term that allows it to quickly adjust to sudden changes in the time series. The proposed models can be used to obtain the predictive distributions of the future number of deaths, as well their means and variances through Markov chain Monte Carlo techniques. We illustrate our methods by runs on data from selected cancer sites.

e-mail: ghosh@njit.edu

BAYESIAN ANALYSIS OF THE 1918 PANDEMIC INFLUENZA IN BALTIMORE, MD AND NEWARK, NJ

Yue Yin*, Johns Hopkins University
Donald Burke, University of Pittsburgh
Derek Cummings, Johns Hopkins University and University of Pittsburgh
Thomas A. Louis, Johns Hopkins University

Around 30 million people were killed in the influenza pandemic of 1918. Though this pandemic has been thoroughly studied, several features remain unexplained. The change of disease transmissibility is suspected to have caused the multiple temporal waves in incidence, which has important implications for pandemic preparedness. With an emphasis in modeling the disease transmissibility, we analyze the population level time series data of influenza in Baltimore using Bayesian method. We found a time-varying disease transmissibility which peaks in the middle of the epidemic.

e-mail: yyin@jhsph.edu

CONSTRAINED BAYES PREDICTION OF LEFT-CENSORED HIV RNA LEVELS AT A MEANINGFUL TIME POINT

Reneé H. Moore*, University of Pennsylvania School of Medicine
Robert H. Lyles, Rollins School of Public Health-Emory University
Amita K. Manatunga, Rollins School of Public Health-Emory University
Kirk A. Easley, Rollins School of Public Health-Emory University

In previous work, we have presented methodology to compute constrained Bayes (CB) predictors of random effects for repeated measures data subject to a limit of detection (LOD). Along with reducing the shrinkage of the Bayes predictor, these CB predictors maintain favorable properties such as minimal bias, reasonable mean square error of prediction, and matching the first two moments of the random effects of interest. In some biomedical examples, it is helpful to predict random effects for repeated measures data subject to a LOD at a clinically meaningful time point. In this talk, we present CB predictions of HIV RNA levels at the time infants who may be at increased risk of rapidly progressing to the AIDS stage reach Class A HIV Status, as defined by the Centers for Disease Control and Prevention. We compare the Bayes and CB estimates of the HIV RNA levels that adjust for the fact that each infant may or may not have non-detectable measurements and the fact that each infant usually reaches Class A HIV Status at a unique time point. We evaluate the performance of the CB predictors through simulation studies.

e-mail: rhmoore@mail.med.upenn.edu

THE USE OF HIERARCHICAL MODELS TO STUDY GENETIC RISK FACTORS

Marinela Capanu*, Memorial Sloan-Kettering Cancer Center
Colin Begg, Memorial Sloan-Kettering Cancer Center

Recent technological progress has led to rapid identification and sequencing of large numbers of genetic variants. Studying the associations between these variants and a particular disease is of great importance to epidemiologists in their quest to decipher the etiology of the disease. Hierarchical modeling is a technique which has been shown to provide more accurate and stable estimates of genetic variants, by incorporating exchange of information through the higher levels of the multilevel model. This talk presents recent research on the application and implementation of hierarchical modeling regression to handle these issues using pseudo-likelihood and Gibbs sampling methods.

e-mail: capanum@mskcc.org

VALIDATING RISK PREDICTION MODELS USING FAMILY REGISTRIES

Wenyi Wang*, Johns Hopkins Bloomberg School of Public Health
Alison P. Klein, Johns Hopkins School of Medicine-Johns Hopkins Bloomberg School of Public Health
Brian Caffo, Johns Hopkins Bloomberg School of Public Health
Giovanni Parmigiani, Johns Hopkins Bloomberg School of Public Health-Johns Hopkins School of Medicine

This paper reports a novel design and analysis framework for utilizing family data to validate individualized risk assessment models. It is motivated by applications to genetic diseases, and illustrated in the context of pancreatic cancer risk prediction. A subset of risk prediction models for cancer emphasize on predicting the risk of developing cancer with a known genetic effect or familial aggregation. Validation in independent data sets is an essential component for the selection, improvement and application of newly developed prediction models. Family registry is used to study familial cancer syndromes. We describe a design strategy for defining a prospective cohort from an existing familial cancer registry, and performing the associated validation analysis. To account for the correlation among individuals from the same family, we used Markov chain Monte Carlo methods to derive the joint risk status of correlated individuals based on the underlying prediction model and evaluated the results using the observed versus expected ratio, the receiver operating characteristic curve, the concordance index and the Brier's score, both at the family level and the individual level. We illustrate our method using PanCPRO, a Mendelian risk prediction model of pancreatic cancer, based on data from the National Familial Pancreatic Tumor Registry.

e-mail: wwang2@jhsph.edu

ON AN EMPIRICAL BAYES ESTIMATOR FOR THE BLUE OF HIV POPULATION BASED ON CD-4 CELL COUNT

Suddhendu Biswas *, University of North Texas Health Science Center
Sejong Bae, University of North Texas Health Science Center
Karan P. Singh, University of North Texas Health Science Center

Commenges and Etcheverry (1993) employed Empirical Bayes approach for obtaining prior distribution of HIV on the basis of the transmission of HIV infection from carrier to susceptible at varying Poisson rates over time. However, the authors did not consider the double decrement pattern of HIV positive individual, viz., the mortality of HIV positive individuals in the state of HIV and the transition of HIV affected individuals to the state of AIDS in the course of growth of HIV population. An extension of the procedure used by Commenges and Etcheverry (1993) is, thus, warranted by considering survival of HIV infected individuals and elimination of the number of AIDS cases formed during the growth of HIV population by considering a suitable incubation period distribution based on reduction of CD-4 cells. By applying the Gauss Aitken generalized least squares method on the linear model given the AIDS vector and the transition matrix, BLUE of fresh incidences of HIV over time were re-estimated. Also, an attempt is made to estimate the proportion of HIV positive population not reaching the state of AIDS by using Martingales Stopping rule.

e-mail: sbae@hsc.unt.edu

THE DIRICHLET PROCESS PRIOR FOR CHOOSING THE NUMBER OF LATENT CLASSES OF DISABILITY AND BIOLOGICAL TOPICS

Tanzy M. Love*, Carnegie Mellon University
Cyrille Joutard, GREMAQ, University Toulouse 1
Edoardo Airoldi, Carnegie Mellon University
Stephen Fienberg, Carnegie Mellon University

The Dirichlet process prior can be used as a prior distribution on the class assignment of a set of objects. This can be naturally implemented in hierarchical Bayesian mixed-membership models (HBMMM) and these encompass a wide variety of models with latent structure for clustering and classification. As in most clustering methods, a principal aspect of implementing HBMMMs is the choice of the number of classes. Strategies for inference on the number of classes (such as RJMCMC methods (Green, 1995)) can be difficult to implement without expertise. The Dirichlet process prior for class assignment can reduce the computational problem to a Gibbs Sampler with book-keeping complications. We produce novel analyses of the following two data sets: (1) a corpus of scientific publications from the Proceedings of the National Academy of Sciences examined earlier in Erosheva, Fienberg, and Lafferty (2004) and Griffiths and Steyvers (2004); (2) data on American seniors from the National Long Term Care Survey examined earlier in Erosheva (2002) and Stallard (2005). Here, our aim is to compare models and specifications by their treatment of these two data sets. Our specifications generalize those used in earlier studies. For example, we make use of both text and references to inform the choice of the number of latent topics in our publications data. We compare our analyses with the earlier ones, for both data sets.

e-mail: tanzy@cmu.edu

106. METHODS FOR HIGH DIMENSIONAL DATA

ROBUST PARTIAL LEAST SQUARES REGRESSION

Asuman S. Turkmen*, Auburn University
Nedret Billor, Auburn University

PLS is a class of methods for modeling relations between sets of observed variables by means of latent variables. The main idea in PLS is to summarize high dimensional predictor variables into a smaller set of uncorrelated, so called latent variables, that have maximal covariance to the response variables. PLS has received a great amount of attention in the field of chemometrics. The success of PLS in chemometrics resulted in a lot of applications in other scientific areas including bioinformatics, food research, medicine, pharmacology, finance, astronomy, and social sciences. Outliers in high dimensions are difficult to detect, but they generally affect the estimation. Robust techniques for ordinary least squares (OLS) regression have been widely developed, but the estimators are still affected by multicollinearity. Although PLS handle collinearity problem, it fails to deal with data containing outliers. Therefore in this study, a new robust PLS regression method, which is resistant to masking and swamping problems, is developed. This new robust method is based on iteratively reweighted technique. Simulated and real data sets are used to assess the performance of the newly proposed robust PLS regression method.

e-mail: turkmas@auburn.edu

CANONICAL PARALLEL DIRECTION FOR PAIRED HIGH DIMENSIONAL LOW SAMPLE SIZE DATA

Xuxin Liu*, University of North Carolina at Chapel Hill
Steve Marron, University of North Carolina at Chapel Hill

High Dimensional, Low Sample Size (HDLSS) data are emerging in a number of areas of science. A lot of them deal with the comparison between paired data sets, i.e. two data sets have the same structure. In this talk, we propose a canonical parallel view for the difference between paired data sets in a novel and useful way. Since the canonical parallel direction characterizes this difference, it can be used for linear adjustment of the difference. The view and adjustment for the NCI60 microarray data set illustrate the good performance of this method. The mathematical statistics of this are usefully studied through the linear shifted model. I found the conditions for both the consistency and the strong inconsistency of the canonical parallel direction in the HDLSS asymptotic context.

e-mail: sprucepku@gmail.com

ROBUST TESTS FOR DETECTING A SIGNAL IN A HIGH DIMENSIONAL SPARSE NORMAL VECTOR

Eitan Greenshtein, SAMSI
Junyong Park*, University of Maryland

Let Z_i, $i=1,...,n$, be independent random variables, $E Z_i=\mu_i$, and $Var(Z_i)=1$. We consider the problem of testing $H_0: \mu_i=0, i=1,...,n$. The setup is when n is large, and the vector $(\mu_1,...,\mu_n)$ is `sparse’, e.g., $\sum_{i=1}^n \mu_i^2=o(\sqrt{n})$. We suggest a test which is not sensitive to the exact tail behavior implied under normality assumptions. In particular, if the ‘moderate deviation’ tail of the distribution of Z_i, may be represented as the product of a tail of a standard normal and a ‘slowly changing’ function, our suggested test is robust. Such a tail behavior, and a need for such a robust test, is expected when the Z_i are of the form $Z_i=\sum_{j=1}^m Y_{ij}/ \sqrt{m}$, for large m, $m<<n$, and independent Y_{ij}.

e-mail: junpark@math.umbc.edu

SUFFICIENT DIMENSION REDUCTION WITH MISSING PREDICTORS

Lexin Li*, North Carolina State University
Wenbin Lu, North Carolina State University

In high-dimensional data analysis, sufficient dimension reduction (SDR) methods provide an effective tool in reducing the dimension of the predictors, while retaining full regression information and imposing no traditional parametric models. However, it is common in high-dimensional data that a subset of predictors may have missing observations. Existing SDR methods resort to the complete-case analysis by removing all the subjects with missingness in any of the predictors under inquiry. Such an approach does not make effective use of all the data, and is valid only when missingness is independent of the observed and unobserved quantities. In this article, we propose a new class of SDR estimators under a more general missing data mechanism, which allows missingness to depend on the observed data. We focus on a widely used SDR method, sliced inverse regression, and propose an augmented inverse probability weighted sliced inverse regression estimator (AIPW-SIR). We show that AIPW-SIR is doubly robust and asymptotically consistent, and demonstrate that AIPW-SIR is more effective than the complete-case analysis through both simulations and real data analysis. We also outline the extension of the AIPW strategy to other SDR methods including sliced average variance estimation and principal Hessian directions.

e-mail: li@stat.ncsu.edu

IMPROVING OPTIMAL SUFFICIENT DIMENSION REDUCTION WITH SYMMETRIC PREDICTORS IN MULTIVARIATE REGRESSION

Jae Keun Yoo*, University of Louisville

Recently, Yoo and Cook (2006) developed an optimal version of Cook and Setodji (2003). Assuming symmetric predictors, we improve Yoo and Cook (2006) by iteratively estimating the inner product matrix used in their method without changing their asymptotic results. A comparison of the three methods is provided via simulation and data analysis.

e-mail: peter.yoo@louisville.edu

CANCER OUTLIER DIFFERENTIAL GENE EXPRESSION DETECTION

Baolin Wu*, University of Minnesota

We study statistical methods to detect cancer genes that are over- or down-expressed in some but not all samples in a disease group. This has proven useful in cancer studies where oncogenes are activated only in a small subset of samples. We propose the outlier robust t-statistic, which is intuitively motivated from the t-statistic, the most commonly used differential gene expression detection method. Using real and simulation studies, we compare the outlier robust t-statistic to the recently proposed COPA and the outlier sum statistic.

e-mail: baolin@biostat.umn.edu

GENERALIZED LATENT VARIABLE MODELS FOR SPATIAL CORRELATED BINARY DATA WITH APPLICATIONS TO DENTAL OUTCOMES

Yanwei Zhang*, Michigan State University
David Todem, Michigan State University
KyungMann Kim, University of Wisconsin-Madison

Analysis of dental caries is traditionally based on DMFS and DMFT scores, which are a summary of a caries experience for each individual. Although this approach has aided our understanding of the pattern of dental caries, there are still some fundamental questions that remain unanswered. As an example, it is well believed among dentists that there are spatial symmetries in the mouth with respect to caries, but this has never been shown in a statistical sense. An answer to this question requires the analysis to be performed at the tooth level. This then necessitates the use of methods for correlated data. In this paper, we propose a generalized latent variable model for the incidence of dental caries at each location while taking into account the unique spatial co-dependence of teeth within a given mouth. We develop a test statistic for assessing the symmetry of dental caries in a mouth using a parametric bootstrap procedure. Data from a cross-sectional survey are used to illustrate the method.

e-mail: zhangy@stt.msu.edu

107. MULTIVARIATE AND CORRELATED DATA

MULTIVARIATE CLUSTERED NON-NORMAL DATA MODELING: WITH APPLICATIONS TO PERIODONTAL RESEARCH

Bin Cheng*, Columbia University

Multivariate clustered data arise naturally in periodontal research since both gingival index and attachment level are periodontal outcomes. The modeling and analysis of the dental data are challenging because of their specific features. First, the dental measures are non-normal and are naturally multilevel: site within tooth, tooth within quadrant, and quadrant within mouth (patient). Second, it is known that periodontal disease as measured by attachment level tends to develop symmetrically between the left and right quadrants. Third, in a longitudinal study design, the spatial correlation within a mouth and the temporal correlation across different visits may not be of equal clinical importance, hence addressing both aspects may require different types of modeling efforts. We propose multivariate multilevel models to address the above issues. Specifically, we suggest several multivariate distributions to jointly modeling gingival index and attachment level. And at each time point, a mixed effects model is formed by assuming a between-tooth spatial correlation, together with quadrant-level and subject-level random effects. Across different time points, either a mixed effects model approach or a marginal modeling approach is adopted to account for repeated measures at different occasions.

e-mail: bc2159@columbia.edu

A GENERAL METHOD OF CONSTRUCTING A TEST OF MULTIVARIATE NORMALITY WITH APPLICATION TO LONGITUDINAL DATA ANALYSIS

Tejas A. Desai*, The Indian Institute of Management-Ahmedabad, India

We present a general method of constructing a test of multivariate normality using any given test of univariate normality of complete or randomly incomplete data. A simulation study considers multivariate tests constructed using the univariate versions of the Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-Von-Mises, and Anderson-Darling tests. Finally, we present an example wherein we apply our methodology to analysis of real longitudinal data.

e-mail: tdesai@iimahd.ernet.in

A GENERALIZATION OF Z TEST TO MULTIVARIATE NORMALITY

Haiyan Su*, University of Rochester
Changyong Feng, University of Rochester
Hongyue Wang, University of Rochester
Xin Tu, University of Rochester
Wan Tang, University of Rochester

Normal distribution underlines the premise for inference for many popular statistical models, especially for longitudinal data analysis such as the mixed-effects model. Validating the normality assumption is crucially for applications of such normal-based models to yield correct inferences. However, most of the current methods for testing the assumption of normality are focused on univariate analysis. In this paper, we propose a new approach to test the normality assumption for multivariate and longitudinal data analyses. In addition, we consider the issue of missing data, a common problem in longitudinal data analysis, and develop corresponding methods to address this important issue under MCAR and MAR, the most common missing data mechanisms in practical applications. Simulation studies are conducted to study the performance of the proposed methodology.

e-mail: haiyan_su@urmc.rochester.edu

CORRELATED BIVARIATE CONTINUOUS AND BINARY OUTCOMES: ISSUES AND APPLICATIONS

Armando Teixeira-Pinto*, Harvard University, Faculdade de Medicina da Universidade do Porto
Sharon-Lise Normand, Harvard Medical School-Harvard School of Public Health

Increasingly, multiple outcomes are collected in order to characterize treatment effectiveness or evaluate the impact of large policy initiatives. Often the multiple outcomes are measured on different scales, such as continuous and binary responses. The common approach to this type of data is to model each outcome separately ignoring the potential correlation between the responses. We describe and contrast several full likelihood and quasi-likelihood multivariate methods for mixed type of outcomes. We present a new multivariate model to analyze binary and continuous correlated outcomes using a latent variable. We study the efficiency gains of the multivariate methods relative to the univariate approach. For complete data, all the models gave consistent estimates for the parameters. When the mean structure of all outcome have the same set of covariates, the gains in efficiency by taking a multivariate approach are negligible. If each outcome has a different set of associated covariates then some of the model estimates can have a dramatic decrease in their variances if a multivariate approach is used. Three real examples illustrates the different approaches.

e-mail: apinto@fas.harvard.edu

A NORMAL-MIXTURE MODEL WITH RANDOM-EFFECTS METHOD FOR ANALYZING RR-INTERVAL DATA

Jessica M. Ketchum*, Virginia Commonwealth University

A methodology for fitting random-effects models with normal-mixture residuals to Heart Rate Variability (HRV) data is presented. The estimation of the parameters is carried out via an EM algorithm. By appropriately defining weights for the incomplete data in the E-step of the EM algorithm, the suggested method uses the current random-effects model methodology. This simplifies the estimation under the normal-mixture considerably. The applications of this model to the HRV data are presented. Results from a simulation study are discussed.

e-mail: mckinneyjl@vcu.edu

AN R-SQUARE STATISTIC FOR FIXED EFFECTS IN THE GAUSSIAN LINEAR MODEL WITH STRUCTURED COVARIANCE

Lloyd J. Edwards*, University of North Carolina at Chapel Hill
Keith E. Muller, University of Florida
Russell D. Wolfinger, SAS Institute
Bahjat F. Qaqish, University of North Carolina at Chapel Hill
Oliver Schabenberger, SAS Institute

The Gaussian linear model with structured covariance stands as one of the most widely used statistical tools for the analysis of correlated outcomes. The R-square statistic and generalizations assess the proportion of variation explained by the model in a wide variety of univariate and multivariate regression and ANOVA settings. A univariate model R-square corresponds to the comparison of two models: 1. a full model that consists of independent predictors and an intercept; 2. a null model that has only the intercept. Model R-square corresponds to a test of the null hypothesis that regression coefficients all equal zero. As a 1–1 function solely of R-square and the degrees of freedom, the F statistic corresponding to the null hypothesis for the test of overall regression links the equivalent formulations. The same principles motivate the proposed R-square statistic for the Gaussian linear model with structured covariance. The R-square statistic arises as a 1–1 function of an appropriate F statistic for comparing a full model for fixed effects to a null model containing only the intercept while having the same covariance structure. The proposed R-square statistic can be interpreted as the proportionate reduction in residual variation achieved by introducing a set of fixed effect predictors.

e-mail: Lloyd_Edwards@unc.edu

AN IMPROVED GENETIC ALGORITHM USING A DERIVATIVE-FREE DIRECTIONAL SEARCH

Wen Wan*, Virginia Polytechnic Institute and State University
Jeffrey B. Birch, Virginia Polytechnic Institute and State University

The genetic algorithm (GA), an important tool used in optimization, has been applied in various fields including Biometrics. However, the general GA is usually computationally intensive, often having to perform a large number of evaluations of an objective function. This paper presents a computationally efficient genetic algorithm by applying the method of steepest ascent/descent from response surface methodology to improve the general GA. Several objective functions, such as low-dimensional versus high-dimensional cases, and smooth response surface versus bumpy response surface cases, are employed to illustrate the improvement of the proposed method, through a Monte Carlo simulation study using a split-plot design. A real problem (Donato et al.,2006) relevant to pharmaceutical and biomedical application is also used to illustrate the improvement of the proposed method over the traditional genetic algorithm and over the method implemented in the Nemrod software used by Donato et al. (2006).

e-mail: wenw@vt.edu

108. DIAGNOSTICS II

A PROBIT LATENT CLASS MODEL WITH GENERAL CORRELATION STRUCTURES FOR EVALUATING ACCURACY OF DIAGNOSTIC TESTS

Huiping Xu*, Purdue University
Bruce A. Craig, Purdue University

Traditional latent class modeling has been widely applied to assess the accuracy of diagnostic tests with dichotomous results. These models, however, commonly assume that the tests are independent conditional on the true disease status, which is rarely valid in practice. While alternative models using probit analysis have been proposed that incorporate dependence among multiple tests, these models commonly assume a restricted correlation structure. In this paper, we propose a probit latent class model that allows general correlation structures. Our model is a generalization of several useful probit latent class models and provides a flexible framework for applications. For maximum likelihood estimation, a parameter-expanded Monte Carlo EM algorithm is implemented and a simple simulation approach is developed for evaluating model fit and comparing models. Our method is illustrated by a simulation study and applied to two published medical data sets.

e-mail: xu20@stat.purdue.edu

ROC ANALYSIS FOR LONGITUDINAL DISEASE DIAGNOSTIC DATA WITHOUT A GOLD STANDARD TEST

Chong Wang*, Cornell University
Bruce W. Turnbull, Cornell University
Yrjo T. Grohn, Cornell University
Soren S. Nielsen, The Royal Veterinary and Agricultural University-Denmark

We develop a Bayesian methodology based on a latent change-point model to estimate the ROC curve of a diagnostic test for longitudinal data. We consider the situation where there is no perfect reference test, i.e. no ‘gold standard’. A change-point process with a Weibull-like survival hazard function is used to model the progression of the hidden disease status. Our model adjusts for the effects of covariate variables, which may be correlated with the disease process or with the diagnostic testing procedure, or both. Markov chain Monte Carlo methods are used to compute the posterior estimates of the model parameters that provide the basis for inference concerning the accuracy of the diagnostic procedure. Based on our Bayesian approach, the posterior probability distribution of the change-point onset time can be obtained and used as a new criterion for disease diagnosis. We discuss an application to an analysis of ELISA scores in the diagnostic testing of paratuberculosis (Johne’s Disease) for a longitudinal study with 1997 dairy cows.

e-mail: cw245@cornell.edu

BAYESIAN SAMPLE SIZE DETERMINATION WITH TWO POSSIBLY CORRELATED IMPERFECT DIAGNOSTIC TESTS

Dunlei Cheng*, Baylor University

This paper investigates the Bayesian sample-size problem for one and two prevalence(s) with misclassification when two diagnostic tests are dependent with each other. We consider two criteria for sample-size determination: the average coverage criterion and the average length criterion. Simulations demonstrate that the posterior average length decreases as sample size increases for both one population and two-population cases. However, average posterior coverage does not always increase with sample size under both one-and two-proportion scenarios. Comparing average posterior length between the dependence and independence models indicates when two error-prone tests are dependent with each other, the posterior length is a little wider than when two infallible tests are independent across all sample sizes. Meanwhile, for both one-proportion and two-proportion cases, average posterior coverage for the dependence model is smaller than that for the independence model.

e-mail: dunlei_cheng@baylor.edu

COMPARING MULTIPLE SENSITIVITIES AND SPECIFICITIES WITH DIFFERENT DIAGNOSTIC CRITERIA: APPLICATIONS TO SEXUAL ABUSE RESEARCH AND STUDIES OF HIGH-RISK SEXUAL BEHAVIOR

Qin Yu*, University of Rochester
Wan Tang, University of Rochester
Xin Tu, University of Rochester

When comparing sensitivities and specificities from multiple diagnostic tests particularly in biomedical research, different test kits have the same diagnostic criterion and are applied to independent groups of subjects with the same disease condition within each group. Although this process gives rise to clustered or correlated test outcomes, the associated inference issues have well been discussed in the literature. In mental health and psychosocial research, sensitivity and specificity are also widely used to characterize the reliability of instruments for diagnosis of mental health disorders, assessment of psychological well-being and evaluation behavioral patterns. However, unlike biomedical applications, different tests are often obtained under different diagnostic criteria, making inference difficult to perform and incorrect when analyzed using methods for comparing multiple test kits in biomedical research. In this paper, we develop a non-parametric approach to address the inference problems. We also address missing data under ignorable nonresponse, the most common type of missingness in such applications. The approach is illustrated with data from two studies examining sexual abuse and high-risk sexual behavior. Key words: Bivariate monotone missing data pattern, Diagnostic test, Missing data, Missing at random, Psychosocial research.

e-mail: qin_yu@urmc.rochester.edu

DIRECT ESTIMATION OF THE AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE IN THE PRESENCE OF VERIFICATION BIAS

Hua He*, University of Rochester
Michael P. McDermott, University of Rochester

The area under a receiver operating characteristic (ROC) curve (AUC) is a commonly used index for summarizing the ability of a continuous diagnostic test to discriminate between healthy and diseased subjects. If all subjects have their true disease status verified, one can directly estimate the AUC nonparametrically using the Wilcoxon statistic. In some studies, verification of the true disease status is performed only for a subset of subjects. Estimators of the AUC based only on verified subjects are typically biased; this is known as verification bias. In the presence of verification bias, one way to estimate the AUC is to compute the empirical AUC from a bias-corrected ROC curve; however, there is no closed-form expression available for either this estimator or its variance. Inference requires the use of resampling methods such as the bootstrap. In this paper, we develop a new method for directly estimating the AUC in the setting of verification bias based on U-statistics and inverse probability weighing. Closed-form expressions for the estimator and its variance are derived. We also show that the new estimator is equivalent to the empirical AUC derived from the bias-corrected ROC curve arising from the inverse probability weighting approach.

e-mail: huahe@bst.rochester.edu

APPLICATION OF LATENT CLASS MODELS TO DIAGNOSTIC TEST DATA

Afisi S. Ismaila*, McMaster University,-Canada

In clinical epidemiology, it is common to classify individuals according to some characteristics of interest. These classifications are usually done through a variety of tests or observations, which cannot be guaranteed to be completely error-free. The ideal method of determining error rates in a new diagnostic test is to compare its results to the "gold standard test' that ascertains true disease status. This method is valid and reliable only if the "gold standard test' is error-free; a situation that is very rare in diagnostic testing. Some statistical methods have been proposed to meet the challenges of estimating test error rates, disease prevalence and relative risk from misclassified data. One of the methods is latent class models developed from latent structure analysis modified for use in problems involving diagnostic tests in the absence of gold standard. This talk examines the theory and application of latent class modelling techniques to the estimation of test error rates of new diagnostic tests in the presence of imperfect reference standard. A comparison of a Bayesian and classical (maximum likelihood based method) estimation approaches was also carried out using three examples. It also discusses some of the criticisms and recent advances in this area.

e-mail: ismailas@mcmaster.ca

EVALUATION OF AGREEMENT BETWEEN OBSERVERS MAKING REPLICATED BINARY ASSESSMENTS

Michael Haber, Rollins School of Public Health-Emory University
Jingjing Gao* Rollins School of Public Health-Emory University
Huiman Barnhart, Duke Clinical Research Institute-Duke University

Agreement between observers classifying subjects according to a binary trait is usually assessed via Cohen's kappa coefficient, which was found to attain erratic values in some cases. To overcome this, we propose a new approach to evaluating agreement between two observers making replicated binary observation on a set of subjects. This approach compares the probability of disagreement (or discordance) between the observers to the probability of disagreement between replicated observations made by the same observer on the same subject. We consider two situations: (1) a symmetric assessment of agreement between two observers, and (2) an assessment of the agreement of a new observer with an imperfect 'gold standard'. We develop a nonparametric method for estimation of the new agreement coefficients when observers make replicated readings on each subject. The reliability of the estimation method is examined via a simulation study. Data from a study aimed at determining the validity of diagnosis of breast cancer based on mammograms is used to illustrate the new concepts and methods.

e-mail: jgao@emory.edu

109. STATISTICAL MODELS FOR GENETIC DATA

WHICH MISSING VALUE IMPUTATION METHOD TO USE IN EXPRESSION PROFILES: A COMPARATIVE STUDY AND TWO SELECTION SCHEMES

Guy N. Brock*, University of Louisville
John R. Shaffer, University of Pittsburgh
Richard E. Blakesley-Ball, University of Pittsburgh
Meredith J. Lotz, University of Pittsburgh
George C. Tseng, University of Pittsburgh

Motivation: Gene expression data frequently contain missing values, however, most down-stream analyses for microarray experiments require complete data. In the literature many methods have been proposed to estimate missing values via information of the correlation patterns within the gene expression matrix. However, the specific conditions for which each method is preferred remains largely unclear. In this report we describe an extensive evaluation of current imputation methods on multiple types of microarray experiments including time series, multiple exposures, and multiple exposures × time series data. We then introduce two complementary selection schemes for determining the most appropriate imputation method for any given data. We developed an entropy measure to quantify the complexity of expression matrixes and found that, by incorporating this information, the entropy-based selection (EBS) scheme performed better than any single method. In addition, a self-training selection (STS) scheme, which determines the optimal imputation method via simulation, was proposed which selected the optimal method with an overall accuracy of 96%, although with increased computation cost.

e-mail: guy.brock@louisville.edu

HIGH-DIMENSIONAL MODEL FOR UNDERSTANDING THE GENETIC NETWORK OF ONTOGENETIC ALLOMETRY

Chenguang Wang*, University of Florida
Qin Li, University of Florida
Rongling Wu, University of Florida

Ontogenetic changes in biological shape and its associated allometry have been studied for over a century, but essentially nothing is known about their underlying genetic and developmental mechanisms. In this talk, we will present a statistical model for detecting and mapping quantitative trait loci that regulate allometric changes of different parts of a body in development. This model is founded on a high-dimensional systems approach by assuming that an organism, say plant, is an intricate system in which main stem, branch, leaf and root are viewed as its elements of construction. The sizes of these elements and their scaling relationships with the overall size of body can be explained by global allocation rules for patterns of biomass partitioning. Such rules mathematically described by allometric power equations have been incorporated into a model for genetic mapping of quantitative trait loci (QTL). By estimating the parameters that define power equations, the model can be used to test the genetic control of QTL over the allometric shape of the organism in development. In jointly modeling multiple biological elements for the system, structural antedependence models are used to approximate the structure of covariance matrix, increasing the model's robustness and stability. We used this new high-dimensional model to analyze an example for a soybean genetic study in which a recombinant inbred line-constructed mapping population was measured for different parts of a plant, including stem, branch, leaf and root biomass and genotyped for over 450 molecular markers. Results from simulation studies that mimic this example suggest that our high-dimensional model displays favorable statistical properties and can be effectively applied to any similar QTL mapping project.

e-mail: cgwang@cog.ufl.edu

TIME SQUARED: REPEATED MEASURES ON PHYLOGENY

Hua Guo*, UCLA
Robert E. Weiss, UCLA
Marc A. Suchard, UCLA

Studies of gene expression profiles in response to external perturbation generate repeated measures data that generally follow non-linear curves. To explore the evolution of such profiles across a gene family, we introduce phylogenetic repeated measures (PR) models. These models draw strength from two forms of correlation in the data. Through gene duplication, the family's evolutionary relatedness induces the first form. The second is the correlation across time-points within taxonic units, individual genes in this example. We borrow a Brownian diffusion process along a given phylogenetic tree to account for the relatedness and co-opt a repeated measures framework to model the latter. Through simulation studies, we demonstrate that repeated measures models outperform the previously available approaches that consider the longitudinal observations or their differences as independent and identically distributed by using deviance information criteria as Bayesian model selection tools; PR models that borrow phylogenetic information also perform better than non-phylogenetic repeated measures models when appropriate. We then analyze the evolution of gene expression in the yeast kinase family across three perturbation experiments. Again, the PR models outperform previous approaches and afford the prediction of ancestral expression profiles.

e-mail: guohua@ucla.edu

IMPROVING IDENTIFICATION OF REGULATORY ELEMENTS BY USING CONTEXT DEPENDENT MARKOV BACKGROUND MODELS

Nak-Kyeong Kim*, NCBI, NLM, NIH
Kannan Tharakaraman, NCBI, NLM, NIH
John L. Spouge, NCBI, NLM, NIH

Many computational methods for identifying regulatory elements use a likelihood ratio between motif and background models. Often, the methods use a background model of independent bases. At least two different Markov background models have been proposed with the aim of increasing the accuracy of predicting regulatory elements. Both Markov background models suffer theoretical drawbacks. Here, we developed a third, context-dependent Markov background model from fundamental statistical principles. Datasets containing known regulatory elements in eukaryotes provided a basis for comparing the predictive accuracies of the different background models. Nonparametric statistical tests indicated that Markov models of order 3 constituted a statistically significant improvement over the background model of independent bases. In addition, our model performed better than the previous Markov background models.

e-mail: kimnak@ncbi.nlm.nih.gov

MODELLING AND ESTIMATING DIFFERENCES IN ALLELE FREQUENCIES USING MULTIPLE SNPS

Nicholas J. I. Lewin-Koh*, Eli Lilly and Company
Lang Li, Indiana University School of Medicine

Estimating and comparing allele frequencies between populations is a challenging task. We approach this problem from an empirical Bayes perspective using a hierarchical model similar to the Gamma model used by Newton et al. (2001) for assessing differential expression. If we consider the problem of rare alleles, if an allele is rare in a sample, than the precision of the estimate for a small samples will be very low. By borrowing strength, we can increase precision of the estimates of the rare allele frequencies as well as test all markers jointly for difference within a gene. We present the model and an example demonstrating the technique.

e-mail: nikko@lilly.com

RANDOM FORESTS AND MULTIPLE IMPUTATION FOR UNCOVERING HAPLOTYPE ASSOCIATIONS

B. Aletta S. Nonyane*, University of Massachusetts School of Public Health and Biostatistics
Andrea S. Foulkes, University of Massachusetts School of Public Health and Biostatistics

Understanding the genetic underpinnings to complex diseases requires consideration of analytical methods designed to uncover associations across multiple predictor variables. At the same time, knowledge of allelic phase, that is whether single nucleotide polymorphisms within a gene are on the same (in cis) or on different (in trans) chromosomal copies, may provide crucial information about measures of disease progression. In association studies of unrelated individuals, allelic phase is generally unobservable, generating an additional analytical challenge. We address these challenges through a novel combination of two existing analytical techniques, random forests and multiple imputation. In addition, we develop a resampling based testing procedure to control appropriately for family wise error in this setting. Our method is applied to data arising from a cohort of N=626 HIV-1 infected individuals at risk for anti-retroviral therapy associated dyslipidemia. This approach reveals differential effects of haplotypes on lipids within and across racial/ethnic groups after controlling for potential confounders. A simulation study is provided and demonstrates reasonable power and control of family-wise error. KEY WORDS: Haplotype, genotype, high-dimensional, phenotype, random forest, CART, multiple imputation, HIV-1, lipids.

e-mail: aletta@schoolph.umass.edu

MULTIVARIATE APPROACHES TO ANALYZING GENE EXPRESSION DATA ENHANCED WITH THE DOMAIN KNOWLEDGE

Daniel C. Parks*, GlaxoSmithKline Pharma R&D
Xiwu Lin, GlaxoSmithKline Pharma R&D
Kwan R. Lee, GlaxoSmithKline Pharma R&D

Research in analysis of gene expression data has focused on identifying lists of individual genes that show expression changes associated with a particular response. The commonly used univariate methods, such as overrepresentation analysis (ORA), functional class scoring (FCS), and distribution analysis (DA), were introduced to analyze gene expression data enhanced with domain knowledge, e.g. clustered data using the Gene Ontology. The correlations can carry valuable information that may shed greater light on the biological mechanisms underlying the observed expression changes, which in turn can provide better interpretability to the biologist. However, these methods treat gene classes individually, ignoring the correlations among the gene classes. To address these correlations among gene classes, we propose multivariate approaches to analyze gene expression data. The first method is based on Group LASSO (least absolute shrinkage and selection operator), and the second is entitled the shrinkage partial correlation and randomization test (SPCR). We apply the Group LASSO and SPCR methods and compare these to other methods using a simulation study. We also present analysis of a real dataset.

e-mail: daniel_c_park@gsk.com

110. LOG-RANK OR OTHER COMPARISONS OF SURVIVAL CURVES IN INDEPENDENT OR MATCHED SAMPLES

A WEIGHTED LOG-RANK TEST TO DETECT EARLY DIFFERENCE IN CENSORED SURVIVAL DISTRIBUTIONS

Qing Xu*, University of Pittsburgh
Jong-Hyeon Jeong, University of Pittsburgh

We revisit the weighted log-rank test where the weight function was derived by assuming the inverse Gaussian distribution for an omitted exponentiated covariate that induces a nonproportionality under the proportional hazards model (Oakes and Jeong, 1998). The method was based on the score test statistic for a group comparison from the proportional hazards model. In this paper, we perform a simulation study to compare the test statistic based on the inverse Gaussian distribution with ones using other popular weight functions including members of the Harrington-Fleming's G-rho family (1982). The nonproportional hazards data are generated by changing the hazard ratios over time under the proportional hazards model. The results indicate that the inverse Gaussian-based test tends to have higher power than some of the members that belong to the G-rho family in detecting a difference between two survival distributions when populations become homogeneous as time progresses. One of the datasets from phase III clinical trials on breast cancer will be illustrated as a real example.

e-mail: qix2@pitt.edu

IMPROVING THE EFFICIENCY OF THE LOGRANK TEST USING AUXILIARY COVARIATES

Xiaomin Lu*, North Carolina State University
Anastasios Tsiatis, North Carolina State University

The logrank test is widely used in many clinical trials for comparing the survival distribution between two treatments with censored survival data. Under the assumption of proportional hazards, it is optimal for testing the null hypothesis of H0 :BETA= 0, where BETA denotes the logarithm of the hazard ratio. In practice, additional auxiliary covariates are collected together with the survival times and treatment assignment. If the covariates correlate with survival times, making use of their information will increase the efficiency of the logrank test. In this paper, we apply the theory of semiparametrics to characterize a class of regular and asymptotic linear (RAL) estimators for BETA when auxiliary covariates are incorporated into the model, and derive estimators that are more efficient. The Wald tests induced by these estimators are shown to be more powerful than the logrank test. Simulation studies are used to illustrate the gains in efficiency.

e-mail: xlu2@ncsu.edu

A SUPREMUM LOG-RANK TEST FOR ADAPTIVE TWO-STAGE TREATMENT STRATEGIES AND CORRESPONDING SAMPLE SIZE FORMULA

Wentao Feng*, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

In two-stage adaptive treatment strategies, patients receive one of the induction treatments followed by a maintenance therapy given that the patient responded to their induction therapy. To test for a difference in the effect of different induction and maintenance treatment combinations, a modified supremum weighted log-rank test is proposed. The test is applied to data from a two-stage randomized trial and compared to the results obtained using a standard weighted log rank test. A sample size formula is proposed based on the limiting distribution of the supremum weighted log-rank statistic. The sample size formula reduces to the Eng and Kosorok's sample-size formula for two-sample supremum log-rank test when there is no second randomization. Monte Carlo simulation studies show that the proposed test provides sample sizes which are close to those obtained by standard weighted log-rank test under a proportional hazard alternative. However, the proposed test is more powerful than the standard weighted log-rank test under non-proportional hazard alternatives. KEYWORDS: Adaptive treatment strategies; Brownian motion; Censoring distribution; Counting process; Proportional hazards; Two-stage designs; Sample size formula; Supremum log rank statistics; Survival Function

e-mail: wef5@pitt.edu

EXACT, DISTRIBUTION FREE CONFIDENCE INTERVALS FOR LATE EFFECTS IN CENSORED MATCHED PAIRS

Shoshana R. Daniel*, University of Pennsylvania
Paul R. Rosenbaum, University of Pennsylvania

In a study of provider specialty in the treatment of ovarian cancer, a late divergence in the Kaplan-Meier survival curves hinted at superior survival among patients of gynecological oncologists when compared to patients of medical oncologists; we ask whether this late divergence should be taken seriously. Specifically, we develop exact, permutation tests, and exact confidence intervals formed by inverting the tests, for late effects in matched pairs subject to random but heterogeneous censoring. Unlike other exact confidence intervals with censored data, the proposed intervals do not require knowledge of censoring times for patients who die. Exact distributions are consequences of three results about signs, signed ranks, and their conditional independence properties. One test, the late effects sign test, has the binomial distribution; the other, the late effects signed rank test, uses nonstandard ranks but nonetheless has the same exact distribution as Wilcoxon's signed rank test. A simulation shows that the late effects signed rank test has substantially more power to detect late effects than do conventional tests. The confidence statement provides information about both the timing and magnitude of late effects.

e-mail: skrieger@mail.med.upenn.edu

CHECKING THE CENSORED TWO-SAMPLE ACCELERATED LIFE MODEL USING INTEGRATED CUMULATIVE HAZARD DIFFERENCE

Seung-Hwan Lee*, Illinois Wesleyan University

New statistical tests for the censored two-sample accelerated life model are discussed. From the estimating functions using integrated cumulative hazard difference, stochastic processes are constructed. They can be described by martingale residuals, and, given the data, conditional distributions can be approximated by zero mean Gaussian processes. The new methods, based on these processes, provide asymptotically consistent tests against a general departure from the model. Some graphical methods are also discussed in terms of simulations. In various numerical studies, the new tests performed better than the existing method, especially when the hazard curves cross. The proposed procedures are illustrated with two real data sets.

e-mail: slee2@iwu.edu

INFERENCE FOR SURVIVAL CURVES WITH INFORMATIVELY COARSENED DISCRETE EVENT-TIME DATA: APPLICATION TO ALIVE

Michelle D. Shardell*, University of Maryland School of Medicine
Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health
David Vlahov, Center for Urban Epidemiologic Studies-New York Academy of Medicine
Noya Galai, Johns Hopkins Bloomberg School of Public Health
Samuel R. Friedman, National Development and Research Institutes

In many prospective studies, including AIDS Link to the Intravenous Experience (ALIVE), researchers are interested in comparing event-time distributions (e.g., for human immunodeficiency virus seroconversion) between a small number of groups (e.g., risk behavior categories). However, these comparisons are complicated by participants missing visits or attending visits off schedule and seroconverting during this absence. Such data are interval-censored, or more generally, coarsened. Most analysis procedures rely on the assumption of non-informative censoring, a special case of coarsening at random that may produce biased results if not valid. Our goal is to perform inference for estimated survival functions across a small number of goups in the presence of informative coarsening. To do so, we propose methods for frequentist hypothesis testing and Bayesian inference of the ALIVE data utilizing information elicited from ALIVE scientists and an AIDS epidemiology expert about the visit compliance process.

e-mail: mshardel@epi.umaryland.edu

ESTIMATING CUMULATIVE TREATMENT EFFECTS IN THE PRESENCE OF NON-PROPORTIONAL HAZARDS

Guanghui Wei*, University of Michigan
Douglas E. Schaubel, University of Michigan

Often in medical studies of time to an event, the treatment effect is not constant over time. In the context of Cox regression modeling, the most frequent solution is to apply a model which assumes the treatment effect is either piece-wise constant or varies smoothly over time; i.e., the Cox non-proportional hazards model. This approach has at least two major limitations. First, it is generally difficult to assess whether the parametric form chosen for the treatment effect is correct. Second, in the presence of non-proportional hazards, investigators are usually more interested in the cumulative than the instantaneous treatment effect (e.g., determining if and when the survival functions cross). Therefore, we propose an estimator for the aggregate treatment effect in the presence of non-proportional hazards. Our estimator is based on the treatment-specific cumulative hazards estimated under a stratified Cox model. No functional form for the non-proportionality need be assumed. Asymptotic properties of the proposed estimators are derived, and the finite-sample properties are assessed in simulation studies. Point-wise and simultaneous confidence bands of the estimator can be computed. The proposed method is applied to data from a national organ failure registry.

e-mail: ghwei@umich.edu

Index of Participants

Abebe, Ash 97
Abecasis, Gonçalo R 12, 18
Abney, Mark 96
Abu-Raddad, Eyas 98
Achy-Brou, Aristide CE .... Poster
Adekpedjou, Akim 56
Aerts, Marc 21
Ahn, Jeonyoun 70
Ahn, Hongshik 46
Ahnaou, Abdellah 85
Airoldi, Edoardo 105
Airoldi, Edoardo M 53
Akondy, Rama 59
Alberts, David 84
Alexandridis, Roxana A 11
Allison, David B .... Poster
Alonso, Ariel 82
Alonzo, Todd A 48
Alosh, Mohamed A 94
Al-Saidy, Obaid A 97
Al-Saleh, Mohammad F 97
An, Di 22
An, Ming-Wen 9
Andraca-Carrera, Eugenio 37
Andrei, Adin-Cristian 24
Apanasovich, Tatiyana V 54
Arani, Ramin B 29
Arbogast, Patrick G 35, 58
Archer, Kellie J .... Poster
Armstrong, Derek 47
Asimit, Jennifer L 98
Assam, Pryseley N 82
Auh, Sungyoung 35
Azuero, Andres 12
Baccini, Michela 27
Bach, Peter B 52
Bae, Sejong 105
Baek, Songjoon 46
Bai, Ping 47
Bailer, A. John 58, 93
Balasubramani, G.K. .... Poster
Baldwin, Clinton T. 12
Bandeen-Roche, Karen 76, 77
Bandos, Andriy 48
Banerjee, Samprit 96
Banerjee, Sibabrata 81
Banerjee, Sudipto 6, 61, 75
Bao, Jieqiong .... Poster
Barlas, Suna 10
Barnhart, Huiman X 36, 108
Barry, William T .... Poster
Basagana, Xavier 22
Bassett, Susan Spear .... Poster
Basu, Sanjib 22
Bate, Andrew 43
Bayarri, Susie 5
Bean, Lora J.H. 59
Beane, Glen L .... Poster
Beckett, Laurel A. 47
Beer, David 73
Beerenwinkel, Niko 16
Begg, Colin 105
Bekele, B. Nebiyou 66, 90
Bellamy, Scarlett L 100
Benaglia, Tatiana A. .... Poster
Bengtsson, Henrik 34
Berger, James 5
Berry, Scott M 66
Berry, Donald A. 66
Betensky, Rebecca A. 71
Beunckens, Caroline 10
Bhore, Rafia 29
Bijnens, Luc 85
Billor, Nedret 97, 106
Binongo, Jose N.G .... Poster
Birch, Jeffrey B. 107
Biswas, Suddhendu 105
Biswas, Swati 11, 36
Blakesley-Ball, Richard E. 109
Blume, Jeffrey D 99
Bodian, Carol A 86
Boehnke, Michael 12, 18
Boorman, Gary A 95
Boos, Dennis 33
Bott, Marjorie 58
Bowman, F. DuBois 26, 47
Branscum, Adam 6, 57
Braun, Thomas M 44, 48
Braun, Willard J. 98
Bretz, Frank 8
Brock, Guy N 109
Broemeling, Lyle D 57
Brookhart, M. Alan 38
Brookmeyer, Ronald .... 9, 19, 33, Poster
Brown, C.H. 76
Brownie, Cavell 33
Brumback, Babette A 58
Burke, Donald 105
Burkom, Howard S 87
Burton, Robert L 93
Byun, Jong-Seok ....
Caffo, Brian S. .... 26, 40, 98, 105, Poster

Index of Participants

Cai, Bo 64, 74
Cai, Jianwen 56
Cai, Tianxi 49
Cai, Tony 60
Cai, Weixing Poster
Cai, Xueya Poster
Campbell, Gregory Roundtable, 45
Cao, Jing 25
Capanu, Marinela 105
Carey, Vincent J 55
Carlin, Bradley P Tutorial, 61, 75
Carr, Daniel B 102
Carrasco, Josep L 36
Carroll, Raymond J 40, 54
Carter, Randy L 20, 82
Carter, Rickey E 60
Carvalho, Benilton S 34
Carvalho, Carlos 31
Casella, George 31, 58
Celedon, Juan C Poster
Chakraborty, Bibhas 39, 67
Chan, Wenyaw 10
Chang, Chung 47
Chang, Suming W 65
Chang, Yu-Ling 96, Poster
Chappell, Rick 8, 66
Charnigo, Richard 71
Charvat, Blake F 33
Chatterjee, Nilanjan 83
Chen, Chia-Cheng 36
Chen, Huichao 21
Chen, James J 46
Chen, Liang 69
Chen, Man-Hua 24
Chen, Ming-Hui 83, 86, 92
Chen, Ming Hui Poster
Chen, Shuo 46
Chen, Sining 19, 26
Chen, Yi Poster
Chen, Yong 19, 23
Cheng, Bin 107
Cheng, Cheng 95
Cheng, Dunlei 108
Cheng, Jing 89
Cheng, Yi 66
Cheng, Yu 80
Cheung, Ken YK 8, 66, 90, Poster
Chi, Yueh-Yun 57
Chinchilli, Vernon M 36
Chinnaiyan, Arul M Poster
Chipman, Hugh 16
Chiswell, Karen E 93, Poster
Cho, Hyunsoon 83
Choi, Hyungwon Poster
Choi, Jimin 34
Choi, Nam Hee 11
Choudhary, Pankaj K 36
Christman, Mary C 68
Chu, Haitao 9, 19, 23
Chun, Hyonho 49
Chung, Yeonseung 85
Churchill, Gary A Poster
Clark, Andrew G 18
Coffey, Christopher S 23, Poster
Cole, Stephen 23
Collins, Allan J Poster
Collins, Linda M. 39
Cook, Dianne H 102
Cook, Richard J 51, 94
Cooper, Kristine L 24
Cope, Leslie 34
Correll, Melanie J Poster
Cortiñas, José 85
Cosman, Joshua D Poster
Costantino, Joseph P 13
Coull, Brent A 71
Cowles, Mary K 86
Craig, Bruce A 7, 108, Poster
Crainiceanu, Ciprian M 64, 82, 83
Crawford, Sara B 22, Poster
Crawford, Carol A.G. 61
Crombleholme, Timothy Poster
Cross, Anne 98
Cui, Lu 44
Cui, Yue 74
Cummings, Derek 105
Cutler, Adele 85
D'Agostino, Ralph 29, 52
Dahl, David B 60
Dai, Hongying 71
Dai, Luyan 81
Dai, Ming 91
Dai, Shuangshuang 95
Dallas, Michael J 33
Dang, Xin 93
D'Angelo, Gina M 95
Daniel, Shoshana R 110
Daniels, Michael J 94, 101
Das, Barnali 84, Poster
Dastrup, Elizabeth Poster
Datta, Somnath 13, 41, Poster
Datta, Sujay 95

Index of Participants

Datta, Susmita 46, Poster
David, Rossell 34
Davidian, Marie 35, 67, 84, 103
Davies-Cole, John 87
Davis, Robert L 43
Dawson, Jeffrey D Poster
de Andrade, Mariza 12
DeCook, Rhonda R 96
DeGruttola, Victor G 16, 56
Delongchamp, Robert R 37
Demirhan, Eren 85
Deng, Yihao 35
Derado, Gordana 47
Desai, Tejas A 107
Desai, Varsha G 37
DeSantis, Stacia M 71
Dettling, Marcel 34
Dey, Dipak K 83
Dhar, Sunil 81
Diao, Guoqing 80
Ding, Mingzhou 2
Ding, Hua 58
DiRienzo, A. Gregory 16
Dmitrienko, Alex 60
Dobbin, Kevin K 46
Dodd, Lori E 65
Dominici, Francesca 61
Donner, Allan 9, 100
Drews, Kimberly L Poster
Drinkenburg, Wilhelmus H.I.M 85
Du, Pang 83
Duan, Fenghai 69
Duan, Liangliang 44
Duclos, Chris 61
DuMouchel, William 43
Duncan, Kristin A 4
Duncan, Paul 58
Dunson, David B 42, 49, 53, 74, 77, 81, 85
E, Lidong 25
Easley, Kirk A 105
Edwards, Lloyd J 76, 107
Edwards, Michael C 4
Edwards, Sharon Poster
Egleston, Brian L 20
Eisen, Susan 4
El ghormli, Laure Poster
Elliott, Michael Poster
Ellis, Steven P 35
Ely, E. Wesley 35
Eschrich, Steven 57
Eskridge, Kent M Poster
Esserman, Denise A 83
Etta, Pisano D 57
Etzel, Carol J 10
Etzioni, Ruth 11
Faerber, Jennifer 39
Fan, Jianqing 54
Fang, Hongbin 85
Fann, Jesse R 94
Fardo, David W Poster
Fay, Michael 27
Fedorov, Valerii V 90
Feingold, Eleanor 59, 95
Feinstein, Steven B 57
Feng, Aimin 43
Feng, Changyong 107
Feng, Chunyao 58
Feng, Dai 47
Feng, Rui 96
Feng, Wentao 110
Feng, Ziding 100
Ferris, Dustin 49
Feuer, Eric Roundtable
Field, William R 84
Fienberg, Stephen E 53, 105
Filardo, Giovanni Poster
Finch, Stephen J 72
Fine, Jason P 28, 80
Finley, Andrew O 75
Fitzmaurice, Garrett Short Course
Follmann, Dean A 27
Foulkes, Andrea S 16, 109
Frangakis, Constantine E 9, 27, Poster
Frankel, Michael R Poster
Franklin, Christine A 30
Freedman, Andrew 52
Fridley, Brooke L 12
Friedman, Samuel R 110
Fuentes, Montserrat 75
Gail, Mitchell 52
Gaile, Daniel P 59, 69
Gajewski, Byron J 20, 58
Galai, Noya 110
Galfalvy, Hanga 35, 73
Gallas, Brandon D 48, 65
Gallop, Robert J 94
Gamalo, Mark A Poster
Gangnon, Ronald E 61, Poster
Ganju, Jitendra 44
Gao, Guimin Poster

Index of Participants

Gao, Haitao ..... Poster
Gao, Jingjing 108
Gao, Tianming ..... Poster
Gao, Xiaoli 59
García-Zattera, María José 10
Gardiner, Joseph C ..... Poster
Gardner, Ian 57
Gardner, Sandra L 29
Garrett-Mayer, Elizabeth ..... 99, Poster
Garrow, Donald 57
Gatsonis, Constantine 65
Gebregziabher, Mulugeta G 9
Gelfand, Alan E ..... 61, 74, 75, Poster
Gelfond, Jonathan A 34
George, Edward 16
Gerstein, Mark 2
Ghosal, Subhashis 48
Ghosh, Debashis ..... 3, 40, 73, 82, 88, Poster
Ghosh, Kaushik 105
Ghosh, Malay ..... 9, Poster
Ghosh, Sujit K 13
Gilbert, Peter B 41
Gilbertson, David T ..... 38, Poster
Gleave, Martin 11
Glickman, Mark E 4
Glueck, Deborah H 57
Glymph, Chevelle 87
Goetgeluk, Sylvie 20
Goetghebeur, Els J 20, 89
Gold, David L 69
Gönen, Mithat ..... Poster
Gordon, Derek 72
Gray, Gerry 45
Gray, Simone ..... Poster
Grayburn, Paul ..... Poster
Greenhouse, Joel ..... Roundtable, 20
Greenshtein, Eitan 106
Gregoire, Timothy ..... Roundtable
Gribbin, Matthew J 23, 29
Griffin, Beth Ann 87
Griswold, Michael ..... Poster
Grohn, Yrjo T 108
Gross, Kevin 7, 68
Gu, Jiezhun 48
Guan, Shanhong 8
Guan, Weihua 12
Guimaraes, Paulo 9
Guindani, Michele 98
Gunter, Lacey 67
Guo, Hua 109
Guo, Qiang 3
Guo, Wensheng 91
Gupta, Mayetri ..... 6, Poster
Gur, David 48
Gurka, Matthew J 44, 76
Gustafson, Paul 19
Haber, Michael 36, 108
Habtzghi, Desale H 13
Haddad, Nick M 68
Hall, Allyson 58
Hall, Daniel B 104
Halloran, M. Elizabeth ..... Tutorial, 59
Hamilton, Cody ..... Poster
Hamman, Baron ..... Poster
Hammerstrom, Thomas 29
Hanfelt, John J 22
Hanford, Kathy 3
Hanson, Timothy E 6, 57, 61
Harezlak, Jaroslaw 77
Hartmann, Katherine E 71
Harvey, Danielle J 47
Haslett, John 76
Haslett, Steve J 76
Hayter, Anthony J ..... Poster
He, Hua 108
He, Qinying 37
He, Wenqing 59
He, Xin 21, 94
He, Xuming 54, 69
He, Yulei 86
He, Zhuoqiong 58
Heagerty, Patrick J 38, 101
Hebbel, Robert 59
Hedayat, Sam 36
Hedden, Sarra L ..... Poster
Hedeker, Donald 39
Heinloth, Alexandra N 95
Heisey, Dennis 104
Heitjan, Daniel F 23
Henderson, David 25
Hernan, Miguel A ..... Tutorial, 20
Herrin, Amy E 60
Herring, Amy H ..... 31, 71, 74, Poster
Herson, Jay 33, 99
Hertzberg, Vicki S ..... Poster
Hesterberg, Tim 25
Hettmansperger, Thomas ..... Poster
Higdon, David M 47, 75
Hillis, Stephen L ..... Poster
Hin, Lin Y 55

Index of Participants

Ho, Yen-Yi 34
Hodge, Susan E. 99
Hodges, James S 74
Hoffmann, Raymond G 98
Hofmann, Heike 102
Hogan, Joseph W 39, 101
Hong, Quan 98
Hong, Soonphil 34
Hooker, Giles 70
Hooten, Mevin B 68
Hoppin, Jane A 74
Horton, Nicholas J 92
Horvath, Steve 3, Poster
Hotilova, Lejla 48
Hou, Wei 98
Houck, Patricia R Poster
Houseman, E. Andres 71
Hsu, Chiu-Hsieh 84
Hsu, Li 80
Hu, Feifang 44
Hu, Haiyan 2
Hu, Jianhua 60
Hu, Peter 14
Huang, Bevan E 18, 72
Huang, Bo 8
Huang, Jian 59, 73, 94, 95, Poster
Huang, Jianhua 64
Huang, Lan 84, Poster
Huang, Peng 82
Huang, Tao 54
Huang, Xianzheng Poster
Huang, Yangxin 98
Huang, Yijian 28
Huang, Yu 2
Hubbard, Rebecca A. 94
Hudgens, Brian R 68
Hudgens, Michael G 16, 25, 41
Hudson, Suzanne S 25
Hughes-Oliver, Jacqueline M 15, Poster
Hung, H.M. James Short Course, 14
Hung, Ying 93
Huo, Xiaoming Poster
Huque, Mohammad 29
Hutson, Alan Poster
Hwang, J.T. G. 60
Hyun, Seunggeun 73
Ibrahim, Joseph G Short Course, 6, 34, 81,
83, Poster
Inoue, Lurdes Y.T 11, 69, 94
Irizarry, Rafael A 34
Irony, Telba 45
Irwin, Richard D 95
Ishwaran, Hemant 31
Islam, Shahidul 59
Ismaila, Afisi S 33, 108
Iyer, Hari K 25
Jacqmin-Gadda, Helene Poster
Jain, Arvind K 87
Jaki, Thomas 81
Jan, Hannig 25
Jang, Woncheol 70
Jara, Alejandro 10
Jennison, Christopher 50
Jeong, Jong-Hyeon 13, 110
Ji, Sungchul 11
Ji, Yuan 90
Jiang, Aixiang 59, 69
Jiang, Guoyong 8
Jiang, Shan 77
Jin, Seoweon 70
Jin, Zhezhen 28
Joffe, Marshall M 38, 39
Johnson, Brent 104
Johnson, Elizabeth A Poster
Johnson, Jacqueline L 23
Johnson, Timothy D 26, 47, Poster
Johnson, Wesley 6
Joly, Damien 104
Jones, Dean P 46
Jonkman, Jeffrey N Poster
Joo, Yongsung 104
Jordan, Lewis 104
Jornsten, Rebecka J 11, 49
Joutard, Cyrille 105
Jung, Sin-Ho 13, 49
Jung, Yoonsung 104
Kader, Gary D 30
Kairalla, John A Poster
Kaizar, Eloise E. 20
Kalbfleisch, Jack D 51
Kalbfleisch, John D 88
Kang, Minho 34
Kang, Shinsoo 35
Kaufmann, Joerg 65
Kearney, Greg 61
Keith, Scott W Poster
Kelemen, Arpad G 72
Keles, Sunduz 11, 49, 69
Kendziorski, Christina 3
Kennedy, Richard E Poster

Index of Participants

Ketchum, Jessica M ... 107
Khuri, Andre I ... 8, 93
Kidane, Gebreyesus ... 87
Kim, Carin J ... 23
Kim, Choongrak ... 34, 46
Kim, Daeyoung ... 25
Kim, Hae-Young ... 25, 37
Kim, Inyoung ... 11
Kim, Jinmee ... 34
Kim, Jonghyeon ... 35
Kim, Jong-Min ... 86, 104
Kim, Joonghyuk ... 104
Kim, KyungMann ... 94, 106
Kim, Mi-Ok ... 81
Kim, Nak-Kyeong ... 109
Kim, Seoung Bum ... 46, 109
Kim, Sinae ... 3, 64, Poster
Kim, Sungduk ... 83
King, Tonya S ... 36
Kinney, Satkartar K ... 49
Klar, Neil ... 9
Klein, Alison P ... 105
Kleinman, Ken P ... 92
Klesczewski, Kenneth ... 44
Klingenberg, Bernhard ... 8
Knapp, Shannon M ... Poster
Koch, Armin ... 50
Koch, Gary ... 78
Kodell, Ralph L ... 46
Koehler, Kenneth J ... 35
Kolm, Paul ... Poster
Kosinski, Andrzej S ... 36
Kosorok, Michael ... Roundtable
Krompf, Kerrie ... Poster
Kunselman, Allen R ... 33
Kuo, Lynn ... 86
Kupper, Lawrence L ... 76
Kwon, Deukwoo ... Poster
Lagakos, Stephen W ... 24
Lai, Tze L ... 90
Lamb, Molly M ... 57
Lan, KuangKuo G ... 14
Landsittel, Doug ... 49
Lange, Christoph ... Poster
Langenberg, Julia A ... 104
Langtimm, Catherine A ... 7
Lansky, David M ... 93
Larsen, Michael D ... 22, 35
Larson, Steven M ... Poster
Laud, Prakash ... 5
LaValley, Michael P ... 37
Lawless, Jerry F ... 51
Lawson , Andrew B ... 71, 104
Lee, Eun-Joo ... 97
Lee, Hye-Seung ... 72
Lee, Jack J ... 25
Lee, Ji-Hyun ... 57, 100
Lee, Jinae ... 104
Lee, Joseph H ... 72
Lee, Keunbaik ... 94, 101, 104
Lee, Kwan R ... 109
Lee, Li-Yin ... 94
Lee, Richard ... Poster
Lee, Robert ... 58
Lee, Seung-Hwan ... 110
Lee, Taewon ... 37
Lei, Yang ... Poster
Leiby, Benjamin E ... 71
Leiva, Ricardo ... 21
Leon, Larry F ... 98
Lesaffre, Emmanuel ... 10
Lessler, Justin ... 19
Lewin, John M ... 57
Lewin-Koh, Nicholas J ... 71, 109
Lewin-Koh, Sock-Cheng ... 71
Li, Ai ... Poster
Li, Di ... 54
Li, Fan ... 27
Li, Haihong ... 20
Li, Hongying ... 96
Li, Hongzhe ... 59, 80
Li, Hua ... Poster
Li, Jialiang ... 21, 48
Li, Lang ... 109
Li, Leping ... 95
Li, Lexin ... 106
Li, Qin ... 109
Li, Ran ... Poster
Li, Runze ... 54
Li, Xiaobo ... Poster
Li, Youjuan ... Poster
Li, Yun ... Poster
Liang, Liming ... 12
Liang, Shengde ... 61
Liang, Yulan Liang ... 72
Lim, Hyun J ... Poster
Lim, Johan ... Poster
Lin, Chen-ju ... Poster
Lin, Danyu ... Tutorial, 13, 18, 72, 80
Lin, Hung-Mo ... Poster

Index of Participants

Lin, Lawrence 36
Lin, Min 12
Lin, Rongheng 95
Lin, Shili 11
Lin, Xiaoyan 74
Lin, Xihong 13, 49, 54, 77, 88, 97
Lin, Xiwu 109
Lin, Yan 59
Lindsay, Bruce G 25
Lipsitz, Stuart 6
Litt, Brian 91
Littell, Ramon C 96
Little, Roderick J.A. 22, 47, Poster
Liu, Changying A
Liu, Fang 34
Liu, Jiajun Poster
Liu, Jinze Poster
Liu, Junmei Poster
Liu, Lei 58, 83
Liu, Lian 82
Liu, Linxu 64
Liu, Peng 60
Liu, Qingzhong 72
Liu, Shufang 13
Liu, Xiaoni 33
Liu, Xuefeng 37, 110
Liu, Xueli Poster
Liu, Xuxin 106
Liu, Yingxue C 97
Liu, Zhenqiu 85
Loh, Ji Meng 84
Lokhnygina, Yuliya 36
Long, Qi 70, 84
Looney, Stephen W Poster
Lotz, Meredith J 109
Louis, Thomas A 61, 83, 103, 105
Love, Tanzy M 105
Lu, Chengxing 9
Lu, Min 58
Lu, Minggen 94
Lu, Tao 98
Lu, Wenbin 12, 73, 106
Lu, Xiaomin 110
Lu, Yun 61
Lucas, Joe 31
Lum, Garret 87
Luo, Sheng 83
Luo, Zhehui Poster
Lyles, Robert H 9, 21, 41, 105
Lynch, Kevin G 10, 71
Ma, Bo 71
Ma, Shuangge 95
Ma, Yan 22
Maathuis, Marloes H 41
Maca, Jeff D 50
MacEachern, Steven N 4
MacLehose, Richard F 53, 74
Madigan, David 43
Maiti, Tapabrata 9
Maity, Arnab 97
Majumdar, Antara 82
Makuc, Diane M Poster
Malcolm, Robert J Poster
Mallick, Bani K 6, 17, 69
Manatunga, Amita K 21, 41, 105, Poster
Mandrekar, Jay 37, 83
Mann, John J 35, 73
Mao, Chang Xuan Poster, 104
Marcus, Michele 21, 41
Marron, Steve 106
Martin, Stacey W 27
Martinez, Fernando J 24
Mason, Christopher J 79
Mazumdar, Sati Poster
McCann, Melinda H 12
McCulloch, Charles E 100
McCulloch, Robert 16
McDermott, Michael P 108
McLain, Alexander C 56
McLeod, Ian 59
Mealli, Fabrizia 27
Mehr, David R 58
Mehrotra, Devan V 27
Mehta, Cyrus R 44
Mei, Yajun 46
Melzer-Lange, Marlene Poster
Menius, Alan J Poster
Mermelstein, Robin 39
Methun, Bengt Short Course
Meyer, Mary C 13
Meyer, Peter M 57
Mi, Xiaojuan Poster
Miecznikowski, Jeffrey C 59
Migrino, Raymond 98
Milbauer, Liming 59
Miller, Richard 57
Miller, Scott W 57
Minard, Charles G 10
Miranda, Marie L 61, Poster
Mitchell, Matthew 15

Index of Participants

Mo, Qianxing 60
Molenberghs, Geert .... Roundtable, 10, 21, 76, ... 82, 85
Monahan, John F .... Poster
Monks, Stephanie A 3, 12
Moon, Hojin 46
Moore, Reneé H 41, 105
Morris, Jeffrey S 4, 79
Müller, Hans-Georg 17
Muller, Keith .... Roundtable
Müller, Peter 5, 34, 74, 98
Mukherjee, Bhramar 9
Mukhopadhyay, Jaydip 104
Mukhopadhyay, Siuli .... 8, Poster
Muller, Keith E .... 23, 57, 76, 107, Poster
Mun, Jungwon 10
Munoz Maldonado, Yolanda 97
Murphy, Sean P 87
Murphy, Susan A 39, 67
Murray, David M 100
Murray, Susan 24
Musani, Solomon K 72
Myers, John A .... Poster
Nagaraja, H.N. 37
Nahm, Moon H 93
Nair, Vijayan N. 39
Nan, Bin 73, 88
Nason, Martha C 14
Neiger, Jeffrey 57
Neira, Mauricio 11
Nelson, Colleen 11
Nelson, David R. .... Poster
Nettleton, Dan 72, 96
Ng, Hon KT .... Poster
Nguyen, Hoang 90
Ni, Xiao 49
Ni, Xuelei (Sherry) .... Poster
Nichols, Thomas E 26
Nie, Lei 23, 41
Nielsen, Soren S 108
Ning, Jing 20, 77
Noble, Robert B 58
Nolan, Vikki 12
Nonyane, B. Aletta S 109
Normand, Sharon-Lise 107
Norris, Emilea 15
Nowak, Norma J 59, 69
Oakes, David 24
Obenchain, Robert L 5
Ober, Carole 96
Oden, Neal 35
Ogden, Robert Todd 47
Oh, Minkyung 34
Oluyede, Broderick O 35, 73
Ombao, Hernando 91
Oquendo, Maria A 35, 73
Osnas, Erik 104
Owzar, Kouros 49
Page, Grier P 69
Paik, Myunghee Cho 22, 72
Pajewski, Nicholas 98
Pak, Youngju 20
Palta, Mari 21
Pan, Qing 86
Pan, Wei 11, 59
Pan, Wenqin 22
Pang, Herbert 11
Papachristou, Charalampos 96
Paquerault , Sophie 65
Pararai, Mavis .... Poster
Parfionovas, Andrejus 85
Park, Cheolwoo 70
Park, Ju-Hyun 74, 81
Park, Junyong 106
Parker, Jennifer D .... Poster
Parks, Daniel C 109
Parmigiani, Giovanni 34, 40, 105
Pashkevich, Maksim A 84
Patel, Rajan S. .... Poster
Patil, Sujata M 21, 94
Patki, Amit 72
Pearson, Caroline A. .... Poster
Peña, Edsel A 51, 56
Pencina, Michael 52
Pendergast, Jane .... Roundtable
Peng, Hanxiang 93
Peng, Limin 21, 28, 71
Pennell, Michael L 53
Pennello, Gene A 45, 65
Pepe, Margaret S .... 48, Poster
Perl, Trish M 19
Peter, Gary F .... Poster
Petrick, Nicholas 47, 65
Pevsner, Jonathan 40
Pfeiffer, Ruth M .... 52, 72, Poster
Piamjariyakul, Ubolrat 58
Pickle, Linda W .... 84, 102, Poster
Pihur, Vasyl .... Poster
Pinheiro, Jose C 8, 50
Plevritis, Sylvia K 84

Index of Participants

Plikaytis, Brian D 27
Poisson, Laila M Poster
Pounds, Stanley B 95
Preisser, John 37, 55, 101
Price, Megan E Poster
Proschan, Michael A 14, 27
Proust-Lima, Cecile Poster
Prucka, William R Poster
Qaqish, Bahjat F 37, 55, 101, 107
Qin, Gengsheng 48, 57
Qin, Li 70, 91
Qu, Annie 55
Qu, Pingping Poster
Qu, Yongming 23, 48
Quan, Hui Poster
Quiton, Jonathan T 56
Rabbee, Nusrat 33, 108
Raby, Benjamin A Poster
Raghunathan, Trivellore E 21
Rahardja, Dewi 23, 48
Ravichandran, Caitlin 56
Ravishanker, Nalini 104
Ray, Shubhankar 17
Ray, Wayne A 58
Recknor, Justin 72
Redden, David T 12
Reecy, James M 72
Reiss, Philip T 70
Renard, Didier 98
Reynolds, John E 7
Rice, Kenneth M 5, 60
Rizzo, Matthew Poster
Robins, James M 20, 88
Rockette, Howard E 48
Rockhold, Frank Invited Address
Rodriguez, Abel 49, 74
Rohde, Charles 23
Rolka, Henry R 87
Romagnuolo, Joseph 57
Rose, Charles E Jr. 27
Rosen, Ori 17
Rosenbaum, Paul R 110
Rosner, Gary L 57, 98
Ross, Zev Poster
Roths, Scott A Poster
Rotnitzky, Andrea 97
Roy, Ananya Poster
Roy, Anuradha 21, 84
Roy, Jason 25, 39
Royle, Jeffrey A 68
Rubin, Donald B 9, 27, 92
Ruczinski, Ingo 40
Runge, Michael C 7
Ryan, Louise M 53, 71
Ryu, Duchwan 6
Ryu, Euijung 37
Saha, Sourish C 93
Sai, Nudurapati 97
Samawi, Hani M 97
Sammel, Mary D 71
Sampson, Allan R 8
Samuelson, Frank W 65
Sanchez, Brisa N 71
Sang, Huiyang 75
Santra, Sourav 22
Saville, Benjamin R Poster, 107
Savitz, David A Poster
Sawrie, David F 23
Scarpa, Bruno 42
Schabenberger, Oliver 107
Scharfstein, Daniel O 19, 20, 67, 110
Scharpf, Robert B 40
Schaubel, Douglas E 51, 56, 86, 88, 110
Scheike, Thomas H 35
Schildcrout, Jonathan S 101
Schipper, Matthew J 70
Schwartz, Brian 26
Scott, Laura 18
Seal, Pradipta 4
Seaman, John 58
Sebastiani, Paola 2, 12
Seltman, Howard 20
Serroyen, Jan 21
Sha, Naijun 64, Poster
Shaffer, John R 109
Shaffer, Michele L 33
Shaikh, Azeem 89
Shamsi, Kohkan 65
Shardell, Michelle D 110
Sharef, Emmanuel 51
Shen, Haipeng 47
Shen, Jing 104
Shen, Ronglai 82, Poster
Shen, Yu 66, 90
Shepherd, Lori 59
Sheppard, Jeannie-Marie 77
Shera, David M 85, Poster
Sherman, Stephanie L 59
Shi, Qian 86
Shih, Weichung J Poster

Index of Participants

Shim, Heejung 69
Shin, Hyejin .... Poster
Shintani, Ayumi K 35
Shirley, Kenneth E 10
Shiu, Shang Ying 65
Shmueli, Galit 87
Shope, Jean T 21
Short, Margaret B 47
Shults, Justine 21, 55
Shyr, Yu 46
Sidik, Kurex .... Poster
Sigal, Bronislava M 84
Sigeti, David E 47
Sima, Choa 69
Simmons, Susan J .... 15, Poster
Simon, Richard .... Short Course
Simonsen, Katy .... Poster
Simpson, William A 98
Singh, Bhanu P 95
Singh, Karan P 105
Sinha, Debajyoti 6, 57, 83
Sinha, Karabi 82
Sinha, Samiran 9
Sivaganesan, Siva 5
Skol, Andrew 18
Slate, Elizabeth H 57
Slattery, Marty 52
Slaughter, James C 71
Small, Dylan S .... 10, 89, Poster
Smith, Brian J 84
Sohn, Kyung-Ah 42
Son, Chang-Kyoon 86
Song, Hae Ryoung 104
Song, Xiao 48
Soong, Seng-Jaw 29, 93
Soper, Keith .... Roundtable
Sotto, Cristina L 10
Soulakova, Julia N 8
Sparks, JonDavid .... Poster
Speckman, Paul L 61
Speed, Terence P 34
Spiegelman, Donna 23
Spouge, John L 109
Stamey, James 58
Stanford, Joseph B 74
Staniswalis, Joan G 70
Steinberg, Martin H 12
Stellman, Jeanne M 86
Stoffer, David S 91
Stokes, Lynne .... Poster
Stoto, Michael A 87
Strawderman, Robert L 51
Strecher, Victor J 39
Strug, Lisa J 99
Su, Haiyan 107
Su, Li 101
Su, Shuchih .... Poster
Suchard, Marc A 109
Sullivan, Lisa M 52
Sun, Dongchu 61, 74, 81
Sun, Jianguo 21, 24, 94
Sun, Jiayang 95
Sun, Kang .... Poster
Sun, Ning 40
Sun , Qiankun .... Poster
Sun, Wenguang 55, 60
Sun, Yanqing 35, 73
Sundaram, Rajeshwari 56, 73
Sungur, Engin A 104
Swamy, Geeta .... Poster
Swan, Christopher 68
Szatkiewicz, Jin P .... Poster
Tadesse, Mahlet G .... 42, 64, Poster
Taljaard, Monica 9
Tamhane, Ajit C 60
Tan, Huaming .... Poster
Tan, Ming 85
Tan, Zhiqiang 13, 89
Tang, Liansheng 48
Tang, Lihua 84
Tang, Niansheng 81
Tang, Wan 107, 108
Tassone, Eric C .... 61, Poster
Taunton, Roma Lee 58
Taylor, Andrew .... Poster
Taylor, Jeremy MG .... 70, 82, 84, Poster
Taylor, Laura L 56
Tebbs, Joshua M .... Poster
Teixeira-Pinto, Armando 107
Telesca, Donatello 69
Ten Have, Thomas R .... Roundtable, 39, 55,
.... 71, 89
Thall, Peter F 44, 90
Tharakaraman, Kannan 109
Therneau, Terry M 79
Thompson, Patricia 84
Thompson, Wesley K .... 17, Poster
Tian, Guoliang 85
Tibshirani, Robert .... IMS Medallion Lecture
Tilahun, Abel 82

Index of Participants

Tintle, Nathan L 72
Tiwari, Hemant K 72
Tiwari, Ram C 105
Todem, David 106
Tong, Tiejun 69
Tong, Xingwei 24, 94
Troxel, Andrea B 83
Truong, Young 47
Tseng, George C 59, 109
Tseng, Szu-Ching 88
Tsiatis, Anastasios A Short Course, 67, 110
Tsodikov, Alexander 88
Tsui, Kwok Leung 46
Tu, Xin 22, 55, 107, 108
Turkmen, Asuman S 97, 106
Turnbull, Bruce W 108
van der Laan, Mark 41
Vandebosch, An 89
VanderWeele, Tyler J 20
Vannucci, Marina 42, 60, 64, Poster
Vansteelandt, Stijn 20
Vasan, Ramachandrun 52
Vaughan, Diane 47
Velasco, Cruz 37
Verbeke, Geert 21, 76
Vexler, Albert 25
Victor, Timothy W Poster
Vlahov, David 110
Vos, Paul W 25
Vytlacil, Edward J 89
Wager, Carrie G 85
Wahed, Abdus S 110
Wall, Melanie M 77, Poster
Wallenstein, Sylvan 86
Waller, Lance A Poster
Wan, Wen 107
Wan, Ying 33
Wang, Bin 73, 81
Wang, Chenguang 109
Wang, Chen-Pin 76
Wang, Chi 13
Wang, Chong 108
Wang, Deli 93
Wang, Dongyuan 86
Wang, Hongwei Poster
Wang, Hongyue 107
Wang, Huixia 69
Wang, Kai Poster
Wang, Lan 55
Wang, Lianming 24, 97
Wang, Lifeng 59
Wang, Lin 44
Wang, Ling 12
Wang, Lu 97
Wang, Mei-Cheng 20, 28
Wang, Naisyin 97
Wang, Nancy 33
Wang, Shuai 19
Wang, Sijian 73
Wang, Sue-Jane Short Course
Wang, Wenyi 105
Wang, Xiaohui S 17
Wang, Xiaoyan 56
Wang, Xin 60
Wang, Xing 60
Wang, Xinlei Poster
Wang, You-Gan 55
Wang, Yuemei Poster
Wannemeuhler, Kathleen A 41
Washington, Sam 87
Wathen, J. Kyle 44
Wei, Guanghui 110
Wei, Peng 11
Wei, Zhi 80
Weinberg, Matthew 57
Weinhandl, Eric D 38, Poster
Weintraub, William S Poster
Weiss, Robert E 109
Weiss, Scott T Poster
Weissfeld, Lisa A 95
Weng, Qian 47
West, Mike 31
West, R. Webster 46
Wetter, David 10
Wheeler, Matthew W 93
White, Richard E Poster
Wickham, Hadley 102
Wiens, Brian 60
Wikle, Christopher K 68
Wilbur, Jayson D 104
Wilding, Gregory E Poster
Williams, Brian J 75
Williams, Kerry Poster
Williamson, John M Poster
Winner, Larry H 58
Wisniewski, Stephen R Poster
Wolfinger, Russell D 107
Wong, Richard Poster
Wood, Phillip 21
Woolson, Robert F Poster
Wouters, Kristien 85
Wright, Fred A. 60, 96, Poster

Index of Participants

Wu, Baolin 34, 106
Wu, Chengqing 25
Wu, Chien-Fu Jeff 93
Wu, Dongfeng 57
Wu, Haiyan 59
Wu, Jianrong 93
Wu, Michael C 49
Wu, Rongling 12, 72, 96, 98, 109
Wu, Song 72
Wu, Wenting 36
Wu, Yuping 46
Xi, Yingmei 83
Xia, Yingcun 21
Xiang, Qinfang 10
Xiao, Rui 12
Xie, Huiliang 73
Xie, Jichun 21
Xing, Biao
Xing, Eric
Xing, Eric P 2, 42, 53
Xu, Huiping 108
Xu, Qing 110
Xu, Yaomin 95
Yan, Jun 28
Yang, Hongmei 85
Yang, Jie 61, 72
Yang, Mingan 74
Yang, Wanling 86
Yap, John Stephen F 96
Ye, Yi 86
Ye, Yining 51
Yeatts, Sharon D 60
Yi, Nengjun 96
Yiannoutsos, Constantin T 9
Yin, Guosheng 8
Yin, Yue 105
Yoo, Jae Keun 96, 106
Yoo, Wonsuk 11
Yoon, Eunyeong 34
Yoshimura, Kenichi Poster
You, Na 104
You, Xiaojun 9
Young, Linda J Roundtable, 30, 61, 102
Yu, Chang Poster
Yu, Fang 86
Yu, Kai F 25
Yu, Qin 108
Yu, Zhangsheng 13, 24
Yuan, Ming 59, 70, Poster
Yuan, Ying 8
Yuan, Zheng 73
Yue, Yu 61
Yun, Sungtaek 104
Zambarano, Robert 43
Zaslavsky, Alan M 86
Zeger, Scott L 61
Zell, Elizabeth R 27
Zelterman, Daniel Poster
Zeng, Donglin 13, 18, 22
Zhang, Chunling 102
Zhang, Daowen 49, 85
Zhang, Hao Helen 49, 85
Zhang, Heping 81, 96
Zhang, Hui 22
Zhang, Hui 26
Zhang, Ji Poster
Zhang, Jing 58
Zhang, Li 9
Zhang, Lixun 84
Zhang, Min 35
Zhang, Peng 24, 81
Zhang, Shiju 69
Zhang, Song 74
Zhang, Suhong 24
Zhang, Xiaoxi 47
Zhang, Yanwei 106
Zhang, Ying J 24, 94
Zhang, Zhiwei 33
Zhao, Hongwei 24
Zhao, Hongyu 40, 69, Poster
Zhao, Liping Poster
Zhao, Luping 61
Zhao, Ming 98
Zhao, Yan D 23, 48
Zhao, Yichuan 58
Zhou, Jasmine 34
Zhou, Qin Poster
Zhou, Xianghong J 2
Zhou, Xiao-Hua 57
Zhou, Xiaohua A 48
Zhou, Yijie 61
Zhu, Dongxiao 95, Poster
Zhu, Hongtu 81, 83
Zhu, Ji 11, 73
Zhu, Liang 21
Zhu, Xiaoguang 98
Zhu, Zhengyuan 84
Ziegler-Graham, Kathryn Poster
Zorych, Ivan 43
Zou, Fei 34, 60, 96, Poster

NOTES

ENAR WEBSITE

WWW.ENAR.ORG

PDL BioPharma™

Cogs are for wheels. Not for people.

PDL BioPharma is a leader in the discovery, development, and marketing of innovative therapies acute-care environments. Our current product pipeline includes six breakthrough investigatio compounds for the treatment of cardiovascular disorders, cancer, and other serious illnesses. Pioneer science, strong commercial products, and a rich development pipeline define PDL BioPharma. And our succ is based on one key factor—our people. Here, every employee directly contributes to our goal-orien and entrepreneurial culture where teamwork, commitment, and initiative are valued and rewarded

Be part of the future of biotechnology. Consider the following positions:

Senior Biostatistician/Principal – 4118BE-R

In this position, you'll be working with complex project teams, providing technical direction statisticians involved in drug development programs. You'll establish and continue to improve standard analysis and reporting conventions for drug development programs, and improve exist methodology for statistical applications to clinical trials

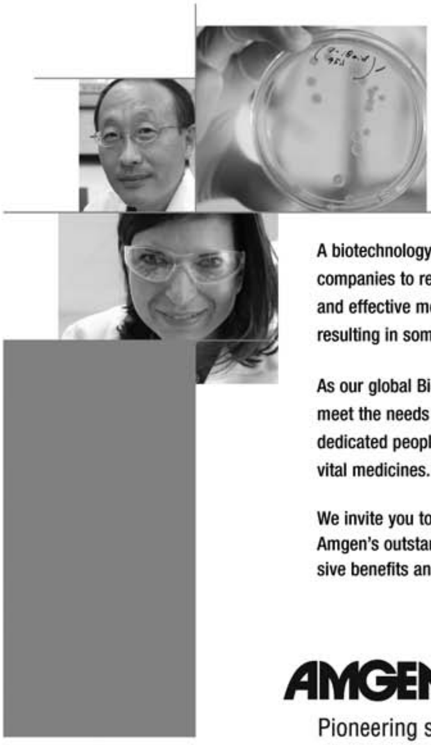
Requirements: An M.A. or Ph.D. in (Bio)statistics with at least 8-15 years' experience in pharmaceutical or biotechnology industry. Experience with NDA filings (preferably 3+). Must have gc written and verbal communication skills. SAS and/or Splus skills required. Recent publication statistical journals or periodicals is strongly preferred. Oncology experience is a plus.

Sr. Bioanalyst – 5018-A

Within this position, you'll be responsible for developing software for the checking, presentation, and anal of clinical study data for submission to regulatory agencies. The software must be developed accordance to departmental SOPs and guidelines, as well as validation criteria specified by regulat agencies.

Requirements: A Bachelors Degree or higher with 8 years SAS programming experience and a minim of 6 years experience in supporting clinical trial data analysis, with minimum of 3 years w responsibility at project level. Skilled programming with the SAS DATA stop and macro facility preferred. Experience with SAS/Graph, SAS/AF and SCL highly desirable. Knowledge of statistics, perl, HTI or web hosting is a plus.

Please apply online at pdl.com/careers



Blazing Trails, Touching Lives.

Join Amgen's Team of Pioneering Scientists.

A biotechnology pioneer since 1980, Amgen was one of the first companies to realize the new science's promise by bringing safe and effective medicines from lab to manufacturing plant to patient – resulting in some of the industry's first blockbuster medicines.

As our global Biostatistics and Epidemiology department grows to meet the needs of Amgen's expanding pipeline, we need talented, dedicated people to join our team and influence the development of vital medicines.

We invite you to become a vital part of a truly extraordinary process. Amgen's outstanding compensation package features comprehensive benefits and relocation assistance.

AMGEN®

Pioneering science delivers vital medicines™

The following opportunities are available at our locations in Thousand Oaks and South San Francisco, California.

Biostatistical Programming Director
Biostatistical Programming Senior Manager
Biostatistical Programming Manager
Biostatistical Programming Senior Associate

Biostatistics Director
Biostatistics Senior Manager
Biostatistics Manager

Epidemiology Manager

Sr. Associate, Clinical Data Management

Please visit Amgen in the ENAR Exhibit Hall. For more information on these opportunities, contact Suzanne Fazalare, Staffing Consultant, at **fazalare@amgen.com** or visit us online at **www.amgen.com/careers.**

www.amgen.com/careers

IBS ENAR

Special Offer to **ENAR Attendees!**

AMERICAN STATISTICAL ASSOCIATION

Get a free American Statistical Association (ASA) publication by joining the ASA during this special promotion!

Enjoy a year of ASA membership and your choice of:

***Journal of the American Statistical Association* (*JASA*)**—*JASA* is the preeminent source of statistical knowledge in the economic, social, physical, engineering, and health sciences. The peer-reviewed journal emphasizes real-world applications. Readers benefit from new applications, revealing case studies, new/useful datasets, and practical indicators. **OR...**

***The American Statistician* (*TAS*)**—Statistical educators, practitioners, researchers, and others interested in the impact of statistics will enjoy *TAS.* Subscribers benefit from challenging discussions about applications and problems in the field, and from resources and reviews for teachers at every level.

Join today and enhance your statistical knowledge right away!

www.amstat.org

ASA Members enjoy:

- *Amstat News*, the monthly membership magazine of the ASA, and *ASA Member News*, our monthly electronic newsletter
- "Members Only" discounts on all ASA publications, meetings, products, continuing education courses, and services
- Access to an invaluable network of professional contacts throughout active regional Chapters and specialty Sections
- Career-enhancing opportunities through the JSM Career Placement Service, *Amstat News*, and online JobWeb postings
- Free web subscription to the *Current Index to Statistics (CIS)*

YES! I would like to join the ASA □ $110 Regular Membership* *JASA* or *TAS* *(circle one)*

* Free online access to *JASA, JBES,* and *TAS*

Name Organization

Address

City State/Province ZIP/Postal Code Country

Phone Email

□ Check/money order payable to the American Statistical Association (in U.S. dollars drawn on a U.S. bank)

Credit Card: □ VISA □ MasterCard □ American Express

Card Number Exp. Date

Name of Cardholder

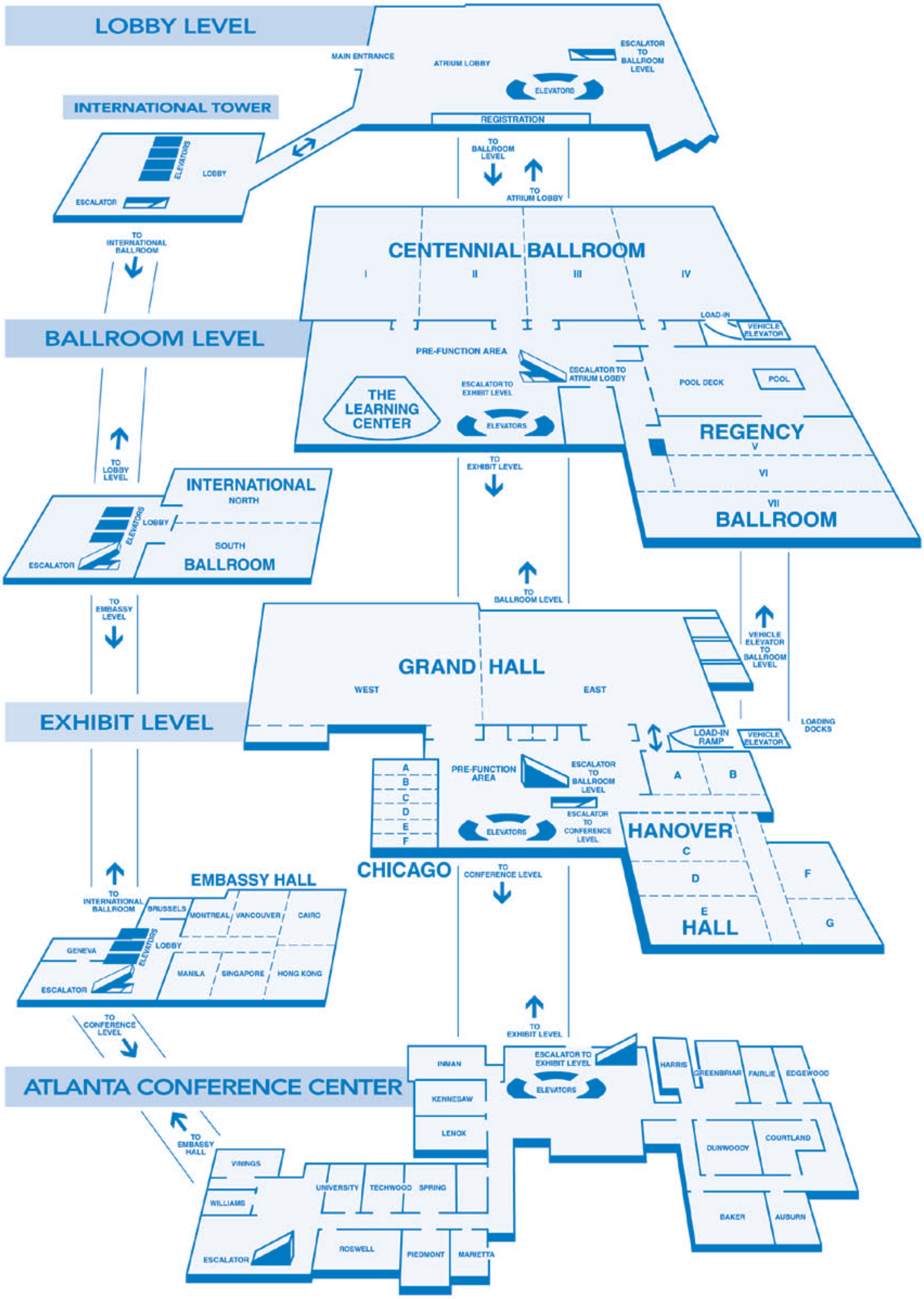
Authorizing Signature

ENAR07

MAIL: American Statistical Association, Dept. 79081, Baltimore, MD 21279-0081

NOTES

Hyatt Regency Atlanta


LOBBY LEVEL
MAIN ENTRANCE
ATRIUM LOBBY
ESCALATOR TO BALLROOM LEVEL
ELEVATORS
REGISTRATION
INTERNATIONAL TOWER
ELEVATORS
LOBBY
ESCALATOR
TO BALLROOM LEVEL
TO ATRIUM LOBBY
TO INTERNATIONAL BALLROOM
CENTENNIAL BALLROOM
I
II
III
IV
LOAD-IN
VEHICLE ELEVATOR
BALLROOM LEVEL
PRE-FUNCTION AREA
ESCALATOR TO ATRIUM LOBBY
ESCALATOR TO EXHIBIT LEVEL
THE LEARNING CENTER
ELEVATORS
POOL DECK
POOL
REGENCY
V
VI
VII
BALLROOM
TO LOBBY LEVEL
INTERNATIONAL
NORTH
LOBBY
ELEVATORS
SOUTH
BALLROOM
ESCALATOR
TO EXHIBIT LEVEL
TO BALLROOM LEVEL
TO EMBASSY LEVEL
GRAND HALL
WEST
EAST
VEHICLE ELEVATOR TO BALLROOM LEVEL
EXHIBIT LEVEL
LOAD-IN RAMP
VEHICLE ELEVATOR
LOADING DOCKS
A
B
C
D
E
F
PRE-FUNCTION AREA
ESCALATOR TO BALLROOM LEVEL
ESCALATOR TO CONFERENCE LEVEL
ELEVATORS
CHICAGO
A
B
HANOVER
C
D
E
HALL
F
G
TO CONFERENCE LEVEL
TO INTERNATIONAL BALLROOM
EMBASSY HALL
BRUSSELS
MONTREAL
VANCOUVER
CAIRO
GENEVA
LOBBY
ELEVATORS
MANILA
SINGAPORE
HONG KONG
ESCALATOR
TO CONFERENCE LEVEL
TO EXHIBIT LEVEL
ATLANTA CONFERENCE CENTER
INMAN
ESCALATOR TO EXHIBIT LEVEL
ELEVATORS
HARRIS
GREENBRIAR
FAIRLIE
EDGEWOOD
KENNESAW
LENOX
TO EMBASSY HALL
VININGS
WILLIAMS
UNIVERSITY
TECHWOOD
SPRING
DUNWOODY
COURTLAND
BAKER
AUBURN
ESCALATOR
ROSWELL
PIEDMONT
MARIETTA