ENAR

Table of Contents

Acknowledgements

SPONSORS

We gratefully acknowledge the support of:

Amgen, Inc.
Biogen Idec
Bristol-Myers Squibb Company
Cytel Inc.
Glaxo Smith Kline
Hoffmann La Roche
ICON Clinical Research
Inspire Pharmaceuticals, Inc.
J & J Pharmaceutical Research & Development
Merck & Company Inc.
Merck Research Laboratories
Novartis Pharmaceuticals, Inc.
PPD, Inc.
Rho, Inc.
Sankyo Pharma Development
sanofi-aventis
SAS Institute
Schering-Plough Research Institute
Statistics Collaborative, Inc.
Wyeth Research

EXHIBITORS

We gratefully acknowledge the support of:

Allergan
Amgen, Inc.
The Cambridge Group, Ltd.
Cambridge University Press
CRC Press - Taylor and Francis Group
Insightful Corporation
JMP Genomics
Kforce Clinical Research Staffing
Minitab Inc.
Oxford University Press
PPD, Inc.
SAS Publishing
SIAM (Society for Industrial and Applied Mathematics)
Smith Hanley Assoc LLC
Springer
John Wiley & Sons, Inc.

Officers and Committees

January – December 2006

EXECUTIVE COMMITTEE – OFFICERS

President	Jane Pendergast
Past President	Peter Imrey
President-Elect	Lisa LaVange
Secretary (2005-2006)	Lance Waller
Treasurer (2006-2007)	Oliver Schabenberger

REGIONAL COMMITTEE (RECOM)

President (Chair), Jane Pendergast
Nine ordinary members (elected to 3-year terms), the Executive Committee, and Scarlett Bellamy (RAB Chair, ex-officio).

2004-2006	2005-2007	2006-2008
Bruce Craig	Gregory Campbell	John Bailer
Amita Manatunga	Naisyin Wang	Stacy Linborg
		Tom TenHave

REGIONAL MEMBERS OF THE COUNCIL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Ron Brookmeyer, Susan Ellenberg, Roderick Little, Louise Ryan, and Janet Wittes

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

Chair: Scarlett Bellamy

2004-2006	2005-2007	2006-2008
Hongshik Ahn	Barbara Bailey	Michael Hardin
Brent Coull	Sudipto Banerjee	Eileen King
Debashis Ghosh	Jason Connor	Carol Lin
Amy Herring	Todd Durham	Keith Muller
Tom Loughin	Kirk Easley	Soomin Park
Jared Lunceford	Abie Ekangaki	Shyamal Peddada
Jeffrey Morris	Deborah Ingram	Jose Pinheiro
Kerrie Nelson	Xuejen Peng	Paul Rathouz
Frank Roesch	James Rosenberger	Jeremy Taylor
Helen Zhang	Maura Stokes	Melanie Wall

Programs

2006 Joint Statistical Meeting
Brent Coull

2006 Spring Meeting – Tampa, FL
Program Chair: Montserrat Fuentes
Program Co-Chair: Jose Pinheiro
Local Arrangements Co-Chairs: TBD

2007 Joint Statistical Meeting
Christopher Coffey

2007 Spring Meetings – Atlanta, GA
Program Chair: Amy Herring
Local Arrangements Chair: Robert Lyles

Biometrics Executive Editor
Marie Davidian

Biometrics Co-Editors
Laurence Freedman, Mike Kenward, 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
Jane Pendergast (President)
Peter Imrey (Past-President)
Lisa LaVange (President-Elect)

ENAR Standing/Continuing Committee Chairs

Nominating (2005)	Marie Davidian
Sponsorship	Frank Shen
Information Technology Oversight (ITOC)	Bonnie LaFleur

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: Jane Pendergast

Workshop for Junior Researchers

Paul Rathouz (Chair)
Brent Coull
Marie Davidian
Michael Epstein
Xihong Lin
Diana Miglioretti

Fostering Diversity Workshop

Co-Chair: Scarlett Bellamy
Co-Chair: Mahlett Tadesse
DuBois Bowman
Marie Davidian
Joel Greenhouse
Jacqueline Hughes-Oliver
Stacy Lindborg
Amita Manatunga
Dionne Price
DeJuran Richardson
Louise Ryan
Kimberly Sellers
Keith A. Soper
Tom Ten Have
Lance Waller

ENAR Student Award Committee

Marie Davidian (Chair)
Karen Bandeen-Roche
Murray Clayton
Brent Coull
Daniel Hall
Philip Dixon
Montserrat Fuentes
Jean Opsomer
Alicia Toledano
Jianwen Cai
David Dunson
Jeff Morris
Josh Tebbs

Student Award Winners

Van Ryzin Award Winner
Edoardo Airoldi, Carnegie Mellon

Award Winners
Daniel Almirall, University of Michigan
Meng Chen, University of Wisconsin-Madison
Xiuyu Cong, Rice University
David Engler, Harvard University
Hae-Young Kim, University of North Carolina - Chapel Hill
Robert T. Krafty, University of Pennsylvania
Joo Yeon Lee, Brown University
Keunbiak Lee, University of Florida
Benjamin Leiby, University of Pennsylvania
Tao Liu, University of Pennsylvania
Fan Lu, University of Wisconsin-Madison
Haijun Ma, University of Minnesota
Arnab Maity, Texas A&M University
Natasa Rajicic, Harvard School of Public Health
Philip Reiss, Columbia University
Rui Song, University of Wisconsin-Madison
Li Su, Brown University
Junfeng Sun, University of Nebraska Medical Center
Xiaodan Wei, University of Wisconsin-Madison
Yan Zheng, University of Minnesota - Twin Cities

2006 ENAR SPECIAL THANKS

2006 ENAR Program Committee
Montserrat Fuentes (Chair), North Carolina State University
José Pinheiro (Co-Chair), Novartis Pharmaceuticals
Chuck Anello, U.S. Food and Drug Administration
Bradley Carlin, University of Minnesota
Paul Rathouz, University of Chicago

ASA Section Representatives
B. Christine Clark, ICON Clinical Research
Parthasarathi Lahiri, University of Maryland
Robert Lyles, Emory University
Jeffrey Morris, University of Texas MD Anderson Cancer Center
Todd Nick, Cincinnati Children's Hospital Medical Center
Alistair James O'Malley, Harvard Medical School
Ingo Ruczinski, Johns Hopkins University
Lara Schmidt, RAND Corporation
Jun Zhu, University of Wisconsin – Madison

IMS Program Co-Chairs
Jason Fine, University of Wisconsin - Madison
Michael Kosorok, University of Wisconsin - Madison

ENAR Education Advisory Committee
Jane Pendergast (Chair), University of Iowa
Marie Davidian, North Carolina State University
Montserrat Fuentes, North Carolina State University
Timothy Gregoire, Yale University
Peter Imrey, Cleveland Clinic Foundation
José Pinheiro, Novartis Pharmaceuticals
Louise Ryan, Harvard School of Public Health
Tom TenHave, University of Pennsylvania

ENAR Student Awards
Marie Davidian, North Carolina State University

ENAR Diversity Workshop
Scarlett Bellamy, University of Pennsylvania
Mahlet Tadesse, University of Pennsylvania

NCI-Sponsored Junior Researcher's Workshop
Paul Rathouz (Chair), University of Chicago
Brent Coull, Harvard School of Public Health
Marie Davidian, North Carolina State University
Michael Epstein, Emory University
Xihong Lin, Harvard School of Public Health
Diana Miglioretti, Group Health Cooperative

ENAR Presidential Invited Speaker

Scott L. Zeger, PhD

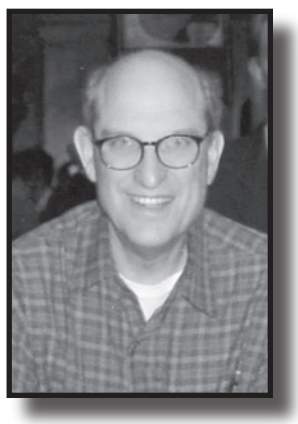
Frank Hurley and Catharine Dorrier Professor in Biostatistics
Chair, Department of Biostatistics, Johns Hopkins University

Scott L. Zeger is professor and chair of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. He received his BA in biology at the University of Pennsylvania in 1974 and his PhD in statistics from Princeton University in 1982. His research is on statistical methods for time series and longitudinal studies.

Dr. Zeger is a Fellow of the American Statistical Association, Fellow of the American Association for the Advancement of Science and co-editor of the Oxford Press journal Biostatistics. He was awarded the 1987 Snedecor Award (with Kung- Yee Liang) for best paper in biometry, in which they introduced the Generalized Estimating Equations (GEE) method. The American Public Health Association recognized Dr. Zeger in 1991 with the Spiegelman Award for contributions to health statistics. In 1987 and 2002, the Johns Hopkins Bloomberg School of Public Health Student Assembly honored Dr. Zeger with the Golden Apple Award for excellence in teaching. Recently, Science Watch identified Dr. Zeger as one of the top 25 most-cited mathematical scientists of the past decade.

Dr. Zeger's research focuses on the design and analysis of data from public health and biomedical studies. His specialty is on drawing inferences from data collected over time on cohorts of individuals. He has made substantive contributions to studies of smoking and health, mental health and environmental health.

IMS Medallion Lecturer



Lawrence D. Brown

Miers Busch Professor in the Department of Statistics
Wharton School of the University of Pennsylvania

Lawrence D. Brown is the Miers Busch Professor in the Department of Statistics at the Wharton School of the University of Pennsylvania. Previously he has held faculty positions at Cornell University, Rutgers University and the University of California, Berkeley. He has a B.S. from the California Institute of Technology (1961) and a Ph.D. from Cornell University (1964). His is a fellow of both the IMS and the ASA, a past-president of the IMS, a member of the US National Academy of Sciences, a former IMS Wald lecturer and a recipient of the ASA's Samuel Wilks Award.

Although Professor Brown is most noted for his research in the theory of admissibility and its relation to properties of random walks and elliptic boundary-value problems, he has published over 100 research papers on a wide variety of theoretical and methodological statistical topics including statistical decision theory and foundations of statistics, sequential analysis, nonparametric function estimation, and data analysis issues related to the U.S. Census and to operation of telephone call-centers. He is the author of an IMS monograph about statistical exponential families and the lead editor of a recent NRC report about the measurement of research and development expenditures in the U.S.

ENAR Short Courses

SC1: Continuous, Discrete, and Incomplete Longitudinal Data
(Full Day: 8:30 a.m. – 5:00 p.m.)
Regency 2

Instructors: Geert Verbeke (Katholieke Universiteit Leuven, Belgium), Geert Molenberghs (Universiteit Hasselt, Belgium)

Description: Starting from an introduction on the linear mixed model for continuous longitudinal data (Verbeke and Molenberghs 2000), extensions will be formulated to model outcomes of a categorical nature, including counts and binary data. Based on Molenberghs and Verbeke (2005), several families of models will be discussed and compared, from an interpretational as well as computational point of view.

First, models will be discussed for the full marginal distribution of the outcome vector. This allows model fitting to be based on maximum likelihood principles, immediately implying inferential tools for all parameters in the models. The main disadvantage of such models is that they require complete specification of all higher-order interactions, which is often based on unrealistic assumptions, and often lead to computational problems, especially in examples with many repeated measurements per subject.

Therefore, alternatives have been formulated in the statistical literature. First, following the reasoning in the linear mixed models, a full marginal model can be obtained from a random-effects approach, where association between repeated measurements within the same subject is believed to be generated by underlying unobserved random effects. Alternatively, semi-parametric methods can be used which no longer require full specification of the likelihood, only of the first moments or of the first and second moments. This leads to the so-called generalized estimating equations (GEE). For both approaches, estimation and inference will be discussed and illustrated in full detail, and it will be extensively argued that the two approaches yield parameters with completely different interpretations. Advantages and disadvantages of both will be discussed in full detail.

Finally, when analyzing longitudinal data, one is often confronted with missing observations, i.e., scheduled measurements have not been made, due to a variety of (known or unknown) reasons. It will be shown that, if no appropriate measures are taken, missing data can cause seriously biased results and interpretational difficulties. Methods to properly analyze incomplete data, under flexible assumptions, are presented. Key concepts of sensitivity analysis are introduced.

DATE: Sunday, March 26, 2006

Full Day Fee	
Members	$210 ($235 after 2/20)
Nonmembers	$260 ($280 after 2/20)
Half Day Fee	
Members	$135 ($160 after 2/20)
Nonmembers	$175 ($200 after 2/20)

Short Course Registration	
Saturday, March 25	3:00 - 5:00 p.m.
Sunday, March 26	7:00 - 8:30 a.m.

(lunch on your own)

Without putting too much emphasis on software, some examples will be given on how the different approaches can be implemented within the SAS software package.

Prerequisites: Throughout the course, it will be assumed that the participants are familiar with basic statistical modeling, including linear models (regression and analysis of variance), as well as generalized linear models (logistic and Poisson regression). Moreover, pre-requisite knowledge should also include general estimation and testing theory (maximum likelihood, likelihood ratio).

As a result of the course, participants should be able to perform a basic analysis for a particular longitudinal data set at hand. Based on a selection of exploratory tools, the nature of the data, and the research questions to be answered in the analyses, they should be able to construct an appropriate statistical model, to fit the model within the SAS framework, and to interpret the obtained results. Further, participants should be aware not only of the possibilities and strengths of a particular selected approach, but also of its drawbacks in comparison to other methods.

The course will be explanatory rather than mathematically rigorous. Emphasis is on giving sufficient detail in order for participants to have a general overview of frequently used approaches, with their advantages and disadvantages, while giving reference to other sources where more detailed information is available. Also, it will be explained in detail how the different approaches can be implemented in the SAS package, and how the resulting outputs should be interpreted.

ENAR Short Courses

SC2: An Introduction to Bayesian Approaches for Data Analysis
(Full Day: 8:30 a.m. – 5:00 p.m.)
Regency 6

Instructor: Alicia Carriquiry (Iowa State University)

The use of Bayesian methods in many areas of application including the biological sciences has mushroomed during the past few years. While the availability of free and easy to use software such as WinBUGS has certainly provided an incentive "to go Bayesian", Bayes methods are widely implemented today because they have proven themselves useful in addressing many complex problems and because they are often simple to implement and intuitively appealing.

This introductory short course is targeted to those with little or no knowledge of the Bayesian approach to data analysis, and focuses on applications rather than on theory. We will begin from scratch, with an overview of Bayesian inference, and will rapidly move on to discussion of single and multiparameter models including examples. Because implementation of Bayesian methods in realistically complex examples is often carried out via simulation, we will then discuss Markov chain Monte Carlo methods (particularly the Gibbs sampler and the Metropolis-Hastings algorithms) and will demonstrate the use of these methods using WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml). We will spend much of the rest of the day discussing the analysis of different types of data, including binary, multinomial, normal and count data and will therefore talk about estimation, prediction and diagnostics for logistic, multinomial, Poisson and normal regression models, perhaps formulated in a hierarchical manner. Time permitting, we might attempt to briefly discuss models for spatially distributed data and mixtures.

Emphasis will be on implementation of the methods and on interpretation of results using numerical approaches. Course notes will be based on Gelman, Carlin, Stern and Rubin, 2nd ed, 2004 and on the notes used in the course "Bayesian Methods for Data Analysis" taught at Iowa State University (http://www.stat.iastate.edu/stat544/homepage.html). The WinBUGS or R code used in all examples will be included in the notes.

Prerequisites: Course participants should have some familiarity with basic mathematical statistics concepts such as conditional expectation and likelihood, and some experience with data analysis from a classical point of view. Those with knowledge of R or SPlus will find it easy to learn WinBUGS, but utter ignorance will not be a hindrance! Little will be assumed to be known.

SC3: Bayesian Adaptive Designs in Drug Development
(Full Day: 8:30 a.m. – 5:00 p.m.)
Regency 5

Instructor: Andy Grieve (Pfizer Global R&D, Sandwich, UK)
Co-Author: Mike Smith (Pfizer Global R&D, Sandwich, UK)

Description: Nearly 40 years ago Cornfield, Halperin and Greenhouse recognized the potential benefits of response-adaptive designs: "The usual justification for not administering an agent of possible efficacy to all patients is the absence of definite information about its effectiveness. However satisfactory as this justification may be before the trial starts, it rapidly loses cogency as evidence for or against the agent accumulates during the course of the trial. But any solution ….. which permits adaptive behavior ….. at least reduces this ethical problem." (Journal of the American Statistical Association, 1969). Recently, there has been considerable interest in both adaptive trials and Bayesian methods in pharmaceutical drug development and in this course issues in the implementation of Bayesian adaptive designs are covered.

In the early part of the course, we review all aspects of adaptation in clinical trials, including, for example, adaptive interim analyses and covariate adaptive randomization. We put Bayesian response-adaptive designs into context, as well as looking at ethical arguments in favour of response-adaptation.

In the second part of the course, we review various simple response adaptive designs, including up-and-down designs, play-the-winner designs and the continuous reassessment method (CRM) and its various modifications, including designs adapting to both efficacy and safety.

In the final part of the course, we cover issues in delayed response and how this can affect adaptive designs. We conclude the course with two case studies.

Prerequisites: This will be an applied course, with emphasis placed on the practical aspects of adaptive procedures rather than theory. Some previous experience of, or exposure to, Bayesian methods would be useful, but is not absolutely essential as it will be reviewed in the course.

ENAR Short Courses

SC4: Introduction to Microarray Technology, Experimental Design, and Data Analysis

Instructor: Dan Nettleton (Iowa State University)
(Full Day: 8:30 a.m. – 5:00 p.m.)
Buccaneer A–B

Description: Microarray technology has become a central tool in functional genomics research. Microarrays allow researchers to measure the abundance of thousands of messenger RNA transcripts in multiple biological samples. By understanding how transcript abundance changes across multiple conditions, researchers gain clues about gene function and learn how genes work together to carry out biological processes. This course will provide an introduction to microarray technology, experimental design, and data analysis for statisticians who are interested in working with scientists on microarray experiments and/or developing statistical methodology for this popular application area.

The course will provide a brief introduction to the basic biology important for understanding microarray technology and an introduction to the technology that will focus on the two most popular microarray platforms: two-color glass-slide microarrays and Affymetrix GeneChips. Data analysis topics will include data normalization, methods for detecting differentially expressed genes (mixed linear model analyses, resampling-based methods, empirical Bayes approaches), controlling the false discovery rate when conducting many tests, and clustering of gene expression profiles. The basics of experimental design for microarrays will be discussed with an emphasis on the fundamental concepts of blocking, randomization, and replication. Methods for choosing between competing experimental designs will be presented.

Prerequisites: This course is targeted to statisticians and/or biological scientists with masters-level statistical training who are interested in analyzing data from microarray experiments and/or developing statistical methodology for the design and analysis of microarray experiments. An understanding of statistical theory and methods at the masters degree level is required.

SC5: Joint Modeling of Repeated Measurements and Event Time Data: Applications in Survival Analysis and Longitudinal Studies with Non-Ignorable Missing Data.
(Half Day: 8:00 a.m. - 12:00 noon)
Regency I

Instructor: Edward F. Vonesh (Baxter Healthcare Corporation)

Description: In many longitudinal trials, especially randomized clinical trials, joint information is often gathered on time to some event (e.g., survival analysis, time to first hospitalization, time to dropout) and serial outcome measures (e.g., repeated measurements, growth curves). Depending on the purpose of the study, one may wish to estimate and compare serial trends over time while accounting for possibly non-ignorable dropout (missing data), or one may wish to investigate any associations that may exist between the event time of interest and various longitudinal trends.

In this course, we examine models and methods for jointly analyzing longitudinal and event time data, primarily within the context of longitudinal studies with missing data, but also within the context of survival analysis. An overview of missing data mechanisms is presented along with an overview of pattern-mixture models and selection models—two classes of models often used in the analysis of longitudinal data with non-ignorable dropout. We present, in detail, a class of random-effects selection models known as shared parameter models that are particularly useful for jointly analyzing longitudinal and event time data. Parametric and semi-parametric survival models for continuous or discrete time survival data, together with generalized linear or nonlinear mixed-effects models for repeated measurements, are proposed for jointly modeling serial outcome measures and event times. Methods of estimation are based on a generalized nonlinear mixed-effects model that may be easily implemented using existing software (e.g., SAS, WinBUGS). This approach allows for flexible modeling of both the distribution of event times and of the relationship of the longitudinal response variable to the event time of interest. A number of examples will be used throughout the course to illustrate these various concepts and techniques with particular attention given to the implementation of these methods in SAS (e.g., NLMIXED).

Prerequisites:
Participants should have an M.S. level of familiarity with linear mixed-effects models for longitudinal data and with proportional hazards models for survival analysis. Those with a working knowledge of SAS will benefit from the examples and code used throughout the course.

ENAR Short Courses

SC6: Causal Inference in Experiments and Observational Studies
(Half Day: 1:00 p.m. – 5:00 p.m.)
Regency I

Instructor: Donald Rubin (Harvard University)

Description: This course will present the perspective for causal inference based on potential outcomes. Randomization-based approaches and Bayesian (model-based) approaches will be developed as complementary and compatible, because both build on the identical framework for the definition of causal effects, which will be presented first, before discussing methods of inference. The critical role of design of studies will be emphasized. Other topics that will receive attention include propensity score methods, model-based predictive methods, principal stratification formulations (a generalization of "Instrumental Variables" approaches), and the compatibility of these methodologies. Examples will be used throughout.

Prerequisites: A desire to understand causal inference.

ENAR 2006 Tutorials

T1: An Introduction to Statistical Modeling with PROC GLIMMIX
Date: Monday, March 27
Time: 8:30-10:15 a.m.
Regency I

Instructor: Oliver Schabenberger (SAS Institute Inc.)

Description:
Generalized linear mixed models are characterized by a conditional distribution—conditional on random effects—in the exponential family and a normality assumption for the random effects. This class of models is very rich; it encompassing linear models, linear mixed models, and generalized linear models. The GLIMMIX procedure, a recent addition to SAS/STAT for SAS 9.1 on the Windows platform, was designed to perform statistical estimation and inference for generalized linear mixed models. It is the focus of this tutorial.

This tutorial :

- discusses the classes of models to which the procedure applies
- introduces important syntax elements for various modeling tasks
- contrasts the procedure with PROC MIXED and PROC NLMIXED
- presents applications to highlight important features in estimation and inference.

The instructor is a Mixed Model Developer at SAS Institute and the author of the GLIMMIX procedure.

Prerequisites:
General familiarity with linear mixed and generalized linear models. Prior exposure to mixed model tools in SAS/STAT, in particular PROC MIXED, is a plus, but not a requirement.

T2: Statistical Modeling of Epidemics
Date: Monday, March 27
Time: 1:45 – 3:30 p.m.
Regency I

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

Description: Infectious diseases, once declared nearly obsolete, have re-emerged with a vengeance. Newspapers carry reports daily of the next new thing, yet statistical analysis of infectious disease data is only recently gaining in popularity in the U.S. 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. Thus, interventions can have many different effects. Data for estimating many of the epidemiologic measures of interest, such as the latent period (the time from infection to onset of infectiousness) or the duration of the infectious period, are difficult to obtain and pose further challenges to estimation.

This tutorial will cover many of the ideas particular to the analysis of infectious disease data and epidemic theory. Topics include the basic reproductive number, transmission probability, epidemic versus endemic infections, thresholds for transmission, latent and incubation periods, serial interval, contact network, deterministic and stochastic epidemic models, Reed-Frost model, 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, partnerships, day care centers, or classrooms for studies of infectious diseases and interventions will be discussed, as well as a few of the commonly used models and estimation procedures. Both Bayesian and likelihood methods will be presented, as well as a few example analyses. Part of the tutorial is conceptual and part is statistical.

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

T3: Intermediate R Use and Programming

Date: Tuesday, March 28
Time: 8:30-10:15 a.m.
Regency I

Instructor: Douglas Bates (University of Wisconsin – Madison and R Core Development Group)

Description:
The Open Source (and freely available) R environment is now widely used for graphics and data analysis, as well as for the computing involved in the development of statistical methodology, such as simulation studies. One of the strengths of R is its packaging system through which a collection of functions and data sets that implement and illustrate a technique can be made available to researchers worldwide. The CRAN archive, with over 500 such packages now available, and the Bioconductor archive, which provides scores of packages associated with analysis of biological experimental data, are well known and valued resources for statistical computing.

The S language, of which R is an implementation, encourages users to progress from elementary interactive usage to writing functions to designing classes and methods and eventually to producing packages for others to use. However, many researchers find it difficult to progress beyond the first step. The purpose of this tutorial is to describe some of the features of the S language and the R environment beyond elementary use. We will discuss the design of simulation studies and MCMC methods, writing functions, classes and methods and creating R packages.

Prerequisites: A working knowledge of statistical methods and some experience using R.

T4: Power and Sample Size Using SAS/STAT Software

Date: Tuesday, March 28
Time: 1:45-3:30 p.m.
Regency I

Instructor: John M. Castelloe (SAS Institute, Inc.)

Description:
This tutorial demonstrates the SAS/STAT Power and Sample Size application (a web interface) and two procedures, PROC POWER and PROC GLMPOWER. Numerous examples of sample size and power determination are illustrated with the new software, including regression, correlation, survival analyses, ANCOVA, confidence intervals, proportion tests, t-tests, and equivalence tests. After reviewing basic methodology, the workshop illustrates how to use the software to compute power and sample size, perform sensitivity analyses when varying other factors such as variability and type I error rate, and produce customized tables, graphs, and narratives. These tasks are important aspects of planning and help produce studies with useful results requiring minimum resources. More information about the software (which is production in SAS version 9.1) is available on the web at www.sas.com/statistics.

Prerequisites: It is assumed that attendees will be familiar with hypothesis testing and confidence intervals. This tutorial will be presented in mostly a learn-by-example style using SAS software for sample size analysis.

ENAR 2006 Roundtables

Roundtable Luncheons – City Center
Date – Monday, March 27, 2006
Time: 12:15 – 1:30 pm

R1: Incomplete Data in Longitudinal Studies

Discussion Leader: Geert Molenberghs, Hasselt University, Belgium

Most, if not all, clinical and non-clinical longitudinal studies are subject to incompleteness. In a biopharmaceutical context, a lot of relatively simple methods for dealing with such incomplete data have been in use. On the other hand, ever more advanced methods for handling incomplete longitudinal studies are proposed by the research community. Regulatory authority based, biopharmaceutical, and academic researchers need to work together to bridge this gap and move on. What would be viable strategies to achieve this goal?

R2: Experimental Design for Biostatisticians

Discussion Leader: Ramon Littell, University of Florida

Most graduate programs in biostatistics do not require a course in classical experimental design. The rationale is that biostatisticians deal mostly with observational studies rather than designed experiments. Some statisticians believe that many basic concepts, such as the distinction between experimental and observational units, are best understood in the framework of designed experiments. They argue that the basic problems are essentially the same for the experimenter and the investigator in an observational study. This roundtable will permit discussion and debate of both sides of the issue.

R3: How to Use Publications to Effectively Disseminate New Knowledge

Discussion Leader: Naisyin Wang, Texas A&M University

Publishing an article is perhaps the most efficient way to communicate a new finding or a new idea to a large group of people. In this roundtable, Biometrics, Naisyin Wang, the co-editor-elect of Biometrics, will lead the discussion on how we can use publication as an effective tool to communicate new findings or new ideas with others. This roundtable discussion will also provide readers/authors of Biometrics an opportunity to provide feedback on how to make Biometrics the most resourceful journal to them.

R4: Statistical Issues in Clinical Trials: Device Studies versus Drug Trials.

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

The similarities and differences between clinical studies of medical devices and clinical pharmaceutical trials are discussed. The nature of medical devices poses a number of statistical challenges, such as how to incorporate prior information, when is a randomized clinical trial absolutely essential, and what statistical methodologies can be brought to bear for the evaluation of diagnostic products, including genetic and genomic tests. The overlap between medical devices and pharmaceuticals is also discussed in terms of combination products such as drug-eluting coronary stents, as well as trials that both establish drug efficacy as well as diagnostic test capability.

R5: Some Issues in Syndromic Surveillance of Public Health Data

Discussion Leader: Andrew B. Lawson, University of South Carolina

Syndromic Surveillance has become an important focus in public health with the widespread concern over bioterrorism threats. In essence, syndromic surveillance concerns early detection of events (essentially using markers or surrogates) when the full knowledge of the outcome is limited. For example, a gastro-intestinal outbreak might take one week to be confirmed from laboratory tests, while ancillary measures (such as pharmaceutical sales and job absenteeism) could be used to make an early decision. This area has many novel statistical challenges. These challenges include multivariate time series, joint modeling of time series and spatio-temporal maps, calibration, Bayesian updating, particle filtration and other speed-ups. Some of the basic issues in this challenging area will be the focus of the discussion.

ENAR 2006 Roundtables

R6: Linking large data sets, mathematical models, and statistics for environmental and ecological applications

Discussion Leader: Lance A. Waller, Emory University

Recent years have seen an increase in data availability for environmental and ecological studies. Not only are more data sets available, but available data sets are increasing in size. Examples include remotely sensed landscape variables as well as georeferenced genomic data. Simultaneously, computational breakthroughs allow simulation and analysis of complicated mathematical models of movement, diffusion, and population interactions. In addition, statistical models have grown more complex and comprehensive. At the intersection of these developments lies a very interesting set of questions to address: How can statistical concepts incorporate uncertainty within "big" ecological models? How could one incorporate "big" data sets from multiple sources (each measured with error) into some sort of cohesive framework for inference? How can one link local sampled data to large-scale models of diffusion and competition? Applications can range from global climatology to local resource use by competing species to the spread of a new disease within a susceptible population.

R7: The Use of Bayesian Inference in the Drug Approval Process

Discussion Leader: Stacy Lindborg, Eli Lilly Corporation

In May 2004, the FDA and Johns Hopkins led an engaging discussion aimed at exploring the following question: "Can Bayesian approaches to studying new treatments improve regulatory decision-making?" The majority of the examples discussed came from experiences through the Center for Devices and Radiological Health. In this round table discussion, we will review the points raised in the May workshop and continue this discussion focusing on the use of Bayesian approaches in the development of drugs. Participants should have interest in exploring and/or experience in using Bayesian methods to aide in decisions internal to companies (e.g., early phase clinical trials) and in discussions with regulators globally.

R8: NIH Statistical Grant Applications and Reviews

Discussion Leader: Xihong Lin, Harvard University

This roundtable luncheon provides biostatisticians who are interested in applying for statistical methodological grants to NIH with information about the NIH process to apply and review grants, the NIH study section reviewing statistical methodological grant proposals (BMRD), grant review criteria, grant writing, and common problems in applications.

R9: Issues in Designing, Planning, and Implementing Adaptive Clinical Trials

Discussion Leader: Brenda L. Gaydos, Eli Lilly and Company

Experiences will be shared from practice, and the host will discuss recommendations from the PhRMA Adaptive Design working group that she co-chairs. Broadly speaking, a design is (response) adaptive when the option exists to modify the study in some way based on one or more outcome measures. Adaptive designs have the potential to increase the efficiency of drug development and to improve the treatment of patients within a trial, for example, by minimizing exposure to ineffective therapies. However, adaptive designs are not always "better" than fixed designs, and there are virtually an infinite number of adaptive design possibilities. From a theoretical perspective, knowledge of available methods is needed to select candidate designs. Then the operating characteristics of competing designs need to be assessed, and in some cases, new statistical methods need to be developed. From an implementation perspective, there are a variety of barriers that need to be addressed such as business planning issues, software gaps, and rapid access to quality data.

R10: Analysis of data with both a longitudinal and a spatial component.

Discussion Leader: Paul Rathouz, University of Chicago

Data that include both a longitudinal and spatial component present interesting modeling and analysis challenges. Some questions to start the discussion are given below.

- In what diverse areas of application can data of this type be found? Environmental epidemiology? Disease surveillance? Forestry? Ecology? What applications do you have in mind?
- What is challenging about handling both longitudinal and spatial data?
- Can the longitudinal or spatial structure of the data be used to control unmeasured confounders? (Much as longitudinal data are used in this way in classical designs.)
- In what other ways can the longitudinal and/or spatial structure be exploited to address interesting scientific questions?
- How would one begin to model such data?

ENAR 2006 Roundtables

R11: What is an "Intent to Treat" (ITT) Population? The Need for Consensus

Discussion Leader: Abie Ekangaki, Eli Lilly Corporation

The Intent-to-Treat paradigm is a core philosophy that underpins the conduct of randomized controlled experiments. The essence of an experiment is to manipulate a condition, e.g. a treatment, while controlling other conditions that are not of primary interest. But experiments may go wrong and patients may not always take their treatments as planned (lack of compliance) or investigators may include patients who don't satisfy the inclusion criteria. Decisions have to be made as to which patient population to include for analysis. Two phrases commonly used in this context are "per-protocol" and "intention-to-treat" (ITT), which provide different perspectives on the analysis population. The regulator typically requires the latter, while it is often supposed that the investigator prefers the former. A common justification for ITT is that it retains the integrity of the randomization process, as used in connection with the frequentist view on hypothesis testing. The intent of this roundtable is to bring to discussion some key issues surrounding the choice of an appropriate ITT population and to hopefully achieve some degree of consensus on this matter. Some key questions of focus will include: Is it appropriate to modify the ITT population for the primary analysis?; Is it reasonable/realistic to be less stringent with the ITT analysis depending on size and duration of trial, population characteristics, trial procedures or adverse events ?; Should trials reflect pharmacologic efficacy or mimic standard of care?; Can/should specific pre-defined subset populations be used for the primary analysis?; If ITT must be used, should complex imputation methods (instead of "last observation carried forward" (LOCF)) be used to impute missing observations?

R12: Standards of Care for Data Collection and Analysis

Discussion Leader: Cyndi Garvan, University of Florida

Ultimately the quality of research depends on the quality of data acquisition and management, yet scant attention is paid to these subjects in traditional graduate education. Too many lessons are learned the "hard way" with needless financial and human costs. Discussion in this roundtable will be focused on defining the role of a statistician in research data management, on lessons learned and pitfalls to avoid, as well as on how new regulations and new capabilities may affect database design and management in the future. We will also discuss ways to incorporate these topics into graduate education.

R13: Diagnostics for Mixed Models

Discussion Leader: Geert Verbeke, Katholieke Universiteit Leuven, Belgium

The linear mixed model has been proposed several decades ago, followed more recently by generalized linear and nonlinear extensions. Together they provide a rich framework for handling wide classes of longitudinal, multilevel, clustered, and otherwise correlated data. The development of flexible standard software tools have contributed to the popularity of mixed model analysis. The field has enjoyed lots of interest from the research community, and remains a center for investigation. A very important topic is diagnostics. What is available and in which directions would more work be welcome?

R14: Measurement Error Models

Discussion Leader: Raymond Carroll, Texas A & M University

Measurement error models in epidemiology have become increasingly sophisticated and complex, both from the frequentist and Bayesian perspectives. The idea of the roundtable is to share interesting emerging problems in the area, e.g., the combination of Berkson and classical models where the Berkson errors in particular have a shared component.

ENAR

Future Meetings of the International Biometric Society

2006 International Biometric Conference

July 16–21, 2006

Montreal, Quebec

2007 ENAR Spring Meetings

Miami, FL

2008 ENAR Spring Meetings

Crystal City, VA

Program Summary

SATURDAY, MARCH 25

9:00 a.m. – 9:00 p.m.
Buccaneer A-B
Workshop for Junior Researchers

3:00 p.m. – 5:30 p.m.
Registration Counter (2nd Floor)
Conference Registration

SUNDAY, MARCH 26

7:30 a.m. – 6:30 p.m.
Registration Counter (2nd Floor)
Conference Registration

8:30 a.m. – 12:00 p.m.
Regency 1
Short Courses
SC5: Joint Modeling of Repeated Measurements and Event Time Data: Applications in Survival Analysis and Longitudinal Studies with Non-ignorable Missing Data

8:30 a.m. – 5:00 p.m.
Short Courses
Regency 2
SC1: Continuous, Discrete, and Incomplete Longitudinal Data
Regency 6
SC2: An Introduction to Bayesian Approaches for Data Analysis
Regency 5
SC3: Bayesian Adaptive Designs in Drug Development
Buccaneer A-B
SC4: Introduction to Microarray Technology, Experimental Design, and Data Analysis

12:30 p.m. – 5:00 p.m.
Regency 3
Fostering Diversity Workshop

1:00 p.m. – 5:00 p.m.
Regency 1
Short Courses
SC6: Causal Inference in Experiments and Observational Studies

4:00 p.m. – 6:00 p.m.
Atrium
Exhibits Open

4:00 p.m. – 7:00 p.m.
Harborview (16th Floor)
ENAR Executive Committee Meeting (Closed)

4:30 p.m. – 6:30 p.m.
Ybor Room
Placement Service Opens

8:00 p.m. – 11:00 p.m.
Regency Ballroom
Social Mixer and Poster Session

MONDAY, MARCH 27

7:30 a.m. – 8:30 a.m.
City Center (Lobby Level)
Student Breakfast

7:30 a.m. – 5:00 p.m.
Registration Counter (2nd Floor)
Conference Registration

7:30 a.m. – 5:00 p.m.
Channelside 2
Speaker Ready Room

9:00 a.m. – 5:00 p.m.
Ybor Room
Placement Service

8:30 a.m. – 5:00 p.m.
Atrium
Exhibits Open

Program Summary

8:30 a.m. – 10:15 a.m.	**Tutorial**
Regency 1	**T1: An Introduction to Statistical Modeling with PROC GLIMMIX**
	Scientific Program
Regency 3	1. Advances in Spatial and Temporal Modeling
Regency 6	2. Missing Data in Longitudinal Studies: Parametric and Semiparametric Perspectives
Regency 7	3. Recent Advances in the Association Analysis for Multivariate Failure Time Data
Regency 2	4. Introductory Lecture Session: Introduction to Statistical Genetics
Regency 5	5. IMS: Non-Standard Maximum Likelihood Inference
Buccaneer A	6. Contributed Papers: Health Services Research
Buccaneer C	7. Contributed Papers: Designing Clinical Trials
Buccaneer B	8. Contributed Papers: Bayesian Methods and Applications
Buccaneer D	9. Contributed Papers: Bioassay and Biopharmaceutical Applications
Esplanade 1	10. Contributed Papers: Generalized Linear Models
Esplanade 3	11. Contributed Papers: Causal Inference
10:15 a.m. – 10:30 a.m.	**Refreshment Break**
Atrium	
10:30 a.m. – 12:15 p.m.	**Scientific Program**
Regency 3	12. New Developments in Microarrays: Identifying Differentially Expressed Genes and Methods for Building Prediction Models
Regency 6	13. Assessing Spatial Surveillance for Bioterrorism: Simulating Attacks
Regency 7	14. Statistical Leadership Under PhRMA Critical Path Initiatives
Regency 2	15. The Role of New Designs for Evaluating Vaccines and Other Prevention Programmes
Regency 5	16. IMS: Dimension Reduction
Buccaneer A	17. Contributed Papers: Environmental and Ecological Applications
Buccaneer B	18. Contributed Papers: Adaptive Clinical Trial Designs and Methods
Buccaneer C	19. Contributed Papers: Statistical and Computational Methods for Genetic Data
Buccaneer D	20. Contributed Papers: Mixed Models: Linear, Generalized, and Non-Linear
Esplanade 1	21. Contributed Papers: Semiparametric and Nonparametric Modeling
Esplanade 3	22. Contributed Papers: Spatial Modeling
12:15 p.m. – 1:30 p.m.	**Roundtable Luncheons (registration required)**
City Center (Lobby Level)	
12:30 p.m. – 4:30 p.m.	**Regional Advisory Board (RAB) Luncheon Meeting (By Invitation Only)**
Garrison Suite	
1:45 p.m. – 3:30 p.m.	**Tutorial**
Regency 1	**T2: Statistical Modeling of Epidemics**
	Scientific Program
Regency 3	23. Statistical Issues in Genetic Investigations
Regency 6	24. Statistical Models in Microparticle Remediation/Decontamination
Regency 7	25. Inference in Randomized Multi-Center Clinical Trials
Regency 2	26. Introductory Lecture Session: Introduction to Longitudinal Data
Regency 5	27. IMS: Recent Advances in Mixture Models
Buccaneer A	28. Contributed Papers: Copula and Cox Regression Models
Buccaneer C	29. Contributed Papers: Biomarkers and Surrogate Markers
Buccaneer B	30. Contributed Papers: Clustering, Classification, and Identification Methods
Buccaneer D	31. Contributed Papers: Survey Data and Sampling Methods
Esplanade 1	32. Contributed Papers: Linkage Analysis
Esplanade 3	33. Contributed Papers: Imaging Methods

Program Summary

3:30 p.m. – 3:45 p.m. *Atrium*	**Refreshment Break**
3:45 p.m. – 5:30 p.m.	**Scientific Program**
Regency 3	34. Statistical Issues in Using Exposure Estimates in Environmental Epidemiology
Regency 6	35. Solutions for Missing Data in Complex Sample Surveys Relevant in Health Policy Research
Regency 7	36. Fusing Biomedical/Environmental Data with Numerical Models
Regency 2	37. Statistical Methods for Public Health Studies in Developing Countries
Regency 5	38. IMS: Spatiotemporal Statistics
Buccaneer D	39. Contributed Session: Computational, Classification, and Model Selection Methods
Buccaneer A	40. Contributed Papers: Survival Analysis I
Esplanade 1	41. Contributed Papers: Methods in Epidemiology
Buccaneer C	42. Contributed Papers: Bayesian Methods in Genomics Data Analysis
Buccaneer B	43. Contributed Papers: Longitudinal Data Analysis
Esplanade 3	44. Contributed Papers: Measurement Error
6:30 p.m. – 7:30 p.m. *City Center (Lobby Level)*	**President's Reception (By Invitation Only)**

TUESDAY, MARCH 28

7:30 a.m. – 5:00 p.m. *Registration Counter (2nd Floor)*	**Conference Registration**
7:30 a.m. – 5:00 p.m. *Channelside 2*	**Speaker Ready Room**
9:00 a.m. – 5:00 p.m. *Ybor Room*	**Placement Service**
8:30 a.m. – 5:00 p.m. *Atrium*	**Exhibits Open**
8:30 a.m. – 10:15 a.m.	**Tutorial**
Regency 1	**T3: Intermediate R Use and Programming**
	Scientific Program
Regency 6	45. New Statistical Methods in Genetic Epidemiology
Regency 7	46. Dealing with Missing Data: Are We There Yet?
Regency 3	47. Statistical Issues in Environmental Monitoring
Regency 2	48. Introductory Lecture Session: Introduction on Data Mining and Its Recent Development
Regency 5	49. IMS: Survival Analysis and Genetic Epidemiology
Buccaneer B	50. Contributed Papers: Graphical Analysis and Reporting of Clinical Safety and Efficacy Data
Buccaneer A	51. Contributed Papers: Random Effects and Frailty Models
Buccaneer C	52. Contributed Papers: Analyzing High Dimensional Genomics Data
Buccaneer D	53. Contributed Papers: Diagnostic and Screening Tests
Esplanade 1	54. Contributed Papers: Methods for Multiple Endpoints
Esplanade 3	55. Contributed Papers: Case-Control Studies
10:15 a.m. – 10:30 a.m. *Atrium*	**Refreshment Break**

Program Summary

10:30 a.m. – 12:15 p.m. *Regency Ballroom*	**Presidential Invited Address**
12:30 p.m. – 4:30 p.m. *Garrison Suite*	**Regional Committee (RECOM) Luncheon Meeting (By Invitation Only)**
1:45 p.m. – 3:30 p.m.	**Tutorial**
Regency 1	**T4: Power and Sample Size Using SAS/STAT Software**
	Scientific Program
Regency 3	56. Censored Data in the Environmental Agricultural and Medical Sciences
Regency 6	57. Statistical Issues in Meta-Analysis of Genomic and Transcriptional Data with a Focus on Array CGH
Regency 7	58. Health Survey Data
Regency 2	59. Introductory Lecture Session: Introduction to Bayesian Analysis and Software
Regency 5	60. IMS: Recent Developments in False Discovery Rate
Buccaneer A	61. Contributed Papers: Interval-Censored Time-to-Event Data
Buccaneer C	62. Contributed Papers: Modeling Methods in Epidemiology
Buccaneer D	63. Contributed Papers: Clinical Trials
Esplanade 1	64. Contributed Papers: Missing Data Methods
Buccaneer B	65. Contributed Papers: Quantitative-Trait Linkage Analysis
Esplanade 3	66. Contributed Papers: Spatial-Temporal Modeling
3:30 p.m. – 3:45 p.m. *Atrium*	**Refreshment Break**
3:45 p.m. – 5:30 p.m.	**Scientific Program**
Regency 2	67. Recent Developments in Bayesian Bioinformatics
Regency 3	68. To Pool or Not to Pool: Systematic Reviews and Pooled Analyses
Regency 6	69. New Statistical Methods for Estimating Medical Expenditures and Cost Effectiveness from Observational Data
Regency 7	70. Statistical Challenges in Preclinical Pharmaceutical Research
Regency 1	71. IMS: Medallion Lecture
Buccaneer A	72. Contributed Papers: Semiparametric and Nonparametric Methods in Longitudinal and Survival Analysis
Buccaneer C	73. Contributed Papers: Spatial Modeling of Disease
Buccaneer B	74. Contributed Papers: Multiple Testing and False Discovery Rates
Buccaneer D	75. Contributed Papers: Competing Risks and Cure Rates
Esplanade 1	76. Contributed Papers: Gene Expression Analysis
5:30 p.m. – 6:30 p.m. *Buccaneer B*	**ENAR Business Meeting (Open to all ENAR Members)**
6:30 p.m. – 9:30 p.m.	**Tuesday Night Event – Yacht Starship Dining Cruise (registration required)**

Program Summary

WEDNESDAY, MARCH 29

7:30 a.m. – 9:00 a.m. **Spring Meeting Planning Committee Breakfast Meeting (Closed)**
Ybor Room

7:30 a.m. – 12:00 noon **Speaker Ready Room**
Channelside 2

8:00 a.m. – 12:30 p.m. **Conference Registration**
Registration Counter (2nd Floor)

8:30 a.m. – 10:15 a.m. **Scientific Program**

Regency 2 77. Recent Advances in Statistical Methods for Genetic Epidemiology
Regency 3 78. Adaptive Bayesian Modeling of Functional Biomedical Data
Regency 6 79. Current Trends in Small Area Estimation
Regency 7 80. Inference in the Presence of Non-Identifiability: Applications to the Analysis of Coarse Data
Regency 5 81. IMS: Bayesian Model Selection
Buccaneer B 82. Contributed Papers: Resampling and Robust Methods and Applications
Buccaneer A 83. Contributed Papers: Pharmacokinetics, Pharmacodynamics, and Toxicology
Buccaneer C 84. Contributed Papers: Survival Analysis II
Buccaneer D 85. Contributed Papers: Topics in Statistics: Sequential Methods, Goodness-of-Fit Tests, and Multivariate Analysis

10:15 a.m. – 10:30 a.m. **Refreshment Break**
Atrium

10:30 a.m. – 12:15 p.m. **Scientific Program**

Regency 2 86. Statistical Issues in the Design, Evaluation, and Monitoring of Clinical Trials with Longitudinal and Survival Endpoints
Regency 3 87. Latent Variables and Multivariate Analysis
Regency 6 88. ROC Analysis in Biomedical Informatics
Regency 7 89. Statistical Contributions to the Frontiers of HIV/AIDS Research
Buccaneer B 90. Contributed Papers: Missing Data in Longitudinal Data Analysis
Buccaneer A 91. Contributed Papers: Categorical Data Analysis and Experimental Design
Buccaneer C 92. Contributed Papers: Analyzing Microarray Data
Buccaneer D 93. Contributed Papers: Nonparametric and Semiparametric Methods

12:15 p.m. **Meeting Adjourns**

Scientific Program: Poster Session Summary

Sunday, March 26
8:00 p.m. – 11:00 p.m.
Regency Ballroom

1. A Bayesian Model for Non-Inferiority Study Design Involving Ethnic Comparisons
Fanni Natanegara*, Eli Lilly and Company, John W. Seaman, Baylor University

2. Modeling Diabetes Incidence in the National Health Interview Survey
Theodore J. Thompson* and James P. Boyle, Centers for Disease Control & Prevention

3. Comparison of Estimators for Average Medical Costs in a Markov Model
Lin Liu*, Zhehui Luo and Joseph C. Gardiner, Michigan State University

4. Receiver Operating Characteristic Curve Analysis Using the Jackknife: An Application to Diagnostic Accuracy for Pigmented Lesions
Paul Kolm*, Michelle L. Pennie and Suephy C. Chen, Emory University

5. A Dynamic Forecasting Model of Diagnosed Diabetes in the U.S. (2005-2050)
James P. Boyle* and Theodore J. Thompson, Centers for Disease Control and Prevention

6. A Comprehensive MALDI-TOF MS Data Preprocessing Method Using Feedback Concept
Shuo Chen*, Ming Li, Huiming Li, Don Hong, Dean Billheimer and Yu Shyr, Vanderbilt University

7. Tests for Comparison of Two Poisson Means
Kangxia Gu and Hon Keung Tony Ng*, Southern Methodist University, Man Lai Tang, Hong Kong Baptist University-Kowloon, Hong Kong

8. Stochastic Optimization for Parameter Estimation in Frailty Models
Tim C. Hesterberg*, Insightful Corporation

9. A Nonparametric Method of Background Correction for Microarray Data Analysis
Zhongxue Chen* and Monnie McGee, Southern Methodist University

10. False Discovery Rate and Multiple Testing Corrections in Disease-Marker Association Studies
Julia Kozlitina* and William R. Schucany, Southern Methodist University, Patrick S. Carmack, UT Southwestern Medical School

11. Honeycomb Designs Computing and Analysis
Andy Mauromoustakos*, University of Arkansas, Vasilia Fasoula*, University of Georgia, Kevin Thompson, University of Arkansas

12. Graph and Hypergraph-Based Models in Bioinformatics: Present and Future
Sujay Datta*, Northern Michigan University

13. Using Robust Estimators Can Increase the Power of the Shapiro-Wilk Test Against Heavy-Tailed Alternatives
Joseph L. Gastwirth*, George Washington University, Weiwen Miao, Macalester College, Yulia Gel, University of Waterloo

14. A Numerical Study of the Problem of Insufficient Overlap in Propensity Scores when Estimating a Causal Treatment Effect
Yuliya Lokhnygina*, Duke University and Duke Clinical Research Institute, Karen Chiswell, North Carolina State University and Duke Clinical Research Institute

15. Spatial Analysis of Probe Level Intensities in Affymetrix Genechip Microarrays
Kinfemichael A. Gedif*, Andrew Hardin, William R. Schucany and Monnie McGee, Southern Methodist University

Scientific Program: Oral Session Summary

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate paper presenters. ENAR Distinguished Student Award Winner presentations appear in **boldface**. The winner of the John van Ryzin prize for best student paper appears in ***boldface italics***.

Monday, March 27
8:30 a.m. – 10:15 a.m.

1. ADVANCES IN SPATIAL AND TEMPORAL MODELING

Regency 3

Sponsors: *ASA Section on Statistics and the Environment*
Organizer: *Marc G. Genton, Texas A&M University*
Chair: *Marc G. Genton, Texas A&M University*

8:30 Space-Time Modeling of Global Ozone Levels
Mikyoung Jun*, Texas A&M University, Michael L. Stein, University of Chicago

8:55 Covariance Tapering for Interpolation of Large Spatial Datasets
Reinhard Furrer*, Colorado School of Mines, Marc G. Genton, Texas A&M University, Douglas Nychka, National Center for Atmospheric Research

9:20 Sequential Estimation of Spatio-Temporal Models
Jonathan R. Stroud*, University of Pennsylvania

9:45 On Optimal Point and Block Prediction in Log-Gaussian Random Fields
Victor De Oliveira*, University of Arkansas

10:10 Floor Discussion

2. MISSING DATA IN LONGITUDINAL STUDIES: PARAMETRIC AND SEMIPARAMETRIC PERSPECTIVES

Regency 6

Sponsors: *ASA Section on Epidemiology/ASA Biopharmaceutical Section*
Organizers: *Annie Qu, Oregon State University, Peter Song, University of Waterloo*
Chairs: *Annie Qu, Oregon State University, Peter Song, University of Waterloo*

8:30 Issues in Multiple Imputation for Hierarchical Data
James R. Carpenter* and Mike G. Kenward, London School of Hygiene & Tropical Medicine

8:55 Multiple Imputation for Nonignorably Missing Data Using a Bayesian Latent-Class Selection Model
Joseph L. Schafer* and Hyekyung Jung, The Pennsylvania State University

9:20 Testing for Missing Data Mechanisms Using Quadratic Inference Function
Grace Y. Yi*, University of Waterloo, Annie Qu, Oregon State University , and Peter Song, University of Waterloo

9:45 Nonlinear Mixed-Effects Models with Dropouts and Missing Covariates
Wei Liu, and Lang Wu*, University of British Columbia

10:10 Floor Discussion

3. RECENT ADVANCES IN THE ASSOCIATION ANALYSIS FOR MULTIVARIATE FAILURE TIME DATA

Regency 7

Sponsors: *IMS/ENAR*
Organizer: *Joanna H. Shih, National Cancer Institute*
Chair: *Jimbo Chen, University of Pennsylvania School of Medicine*

8:30 The Kendall Distribution with Bivariate Censored Data
David Oakes*, University of Rochester, Antai Wang, Georgetown University

8:55 Functional Association Models for Multivariate Temporal Processes
Jason Fine*, University of Wisconsin-Madison

9:20 Analysis of Failure Time Data with Multi-Level Clustering, with Application to the Child Vitamin A Intervention Trial in Nepal
Joanna Shih*, National Cancer Institute, National Institutes of Health, Shou-En Lu, University of Medicine and Dentistry of New Jersey

9:45 Discussant: Karen Bandeen-Roche, Johns Hopkins University, Bloomberg School of Public Health

4. ILS✧: INTRODUCTION TO STATISTICAL GENETICS

Regency 2

Sponsor: *ASA Sections on Teaching Statistics in the Health Sciences/Statistical Education*
Organizers: *Todd G. Nick Cincinnati Children's Hospital Medical Center and Dahlia Nielsen, North Carolina State University*
Chair: *Todd G. Nick, Cincinnati Children's Hospital Medical Center*

8:30 Introduction to Association Studies: Basic Concepts & Methods
Rudy Guerra*, Rice University

9:00 Overview on Structural Association Testing and Regional Admixture Mapping
David B. Allison*, T.M. Beasley, Jose R. Fernandez, David T. Redden, Hemant K. Tiwari, Jasmin Divers and Robert Kimberly, The University of Alabama at Birmingham

✧ ILS stands for Introductory Lecture Session.

9:30 Introduction to Microarrays
Russell D. Wolfinger*, SAS Institute, Inc.
10:00 Floor Discussion

5. IMS: NON-STANDARD MAXIMUM LIKELIHOOD INFERENCE

Sponsor: *IMS*
Organizer: *Moulinath Banerjee, University of Michigan*
Chair: *Ji Zhu, University of Michigan* *Regency 5*

8:30 Rates of Convergence for Current Status Data with Competing Risks
Marloes Maathuis*, University of Washington
8:55 Inference Under Right Censoring for Transformation Models with a Change-Point Based on a Covariate Threshold
Michael R. Kosorok* and Rui Song, University of Wisconsin-Madison
9:20 Nonconcave Penalized Likelihood Inference for Multivariate Survival Data
Jianwen Cai, University of North Carolina at Chapel Hill, Jianqing Fan*, Princeton University, Runze Li, Pennsylvania State University, Haibo Zhou, University of North Carolina at Chapel Hill
9:45 Likelihood Inference under Monotonicity Constraints: Some Recent Developments
Moulinath Banerjee*, University of Michigan
10:10 Floor Discussion

6. CONTRIBUTED PAPERS:
HEALTH SERVICES RESEARCH *Buccaneer A*

Sponsor: *ENAR*
Chair: *Daniel F. Heitjan, University of Pennsylvania*

8:30 Bayesian Inference of the Lead Time in Periodic Cancer Screening
Dongfeng Wu*, Mississippi State University, Gary L. Rosner, and Lyle D. Broemeling, University of Texas, M.D. Anderson Cancer Center
8:45 The Multi-Phase Optimization Strategy: A New Way to Develop Multi-Component 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
9:00 Robust Estimation for the Mean Medical Expenditure
Kenny Shum* and Scott L. Zeger, Johns Hopkins University
9:15 Sample Size Requirements for Studying Small Populations in Gerontology
Robert B. Noble, A. John Bailer*, Suzanne R. Kunkel and Jane K. Straker, Miami University
9:30 Modeling Differentiated Associations Between Physiological Dysregulation and Frailty in Older Women
Hongfei Guo* and Karen Bandeen-Roche, Johns Hopkins University
9:45 A Locally Weighted Regression Approach to Evaluating the Effects of Smoking Reduction on Birth Weight
Jeff M. Szychowski*, J. Michael Hardin and Michael D. Conerly, University of Alabama, Wendy Horn, Cooper Green Hospital, Jefferson Health System (CCOE), Lesa Woodby, University of Alabama at Birmingham
10:00 Adjusting Longitudinal Confounding Variables
Haiqun Lin*, Yale University

7. CONTRIBUTED PAPERS:
DESIGNING CLINICAL TRIALS *Buccaneer C*

Sponsor: *ENAR*
Chair: *Fanni Natanegara, Eli Lilly and Company*

8:30 Clinical Trials Simulation: Overview and Demonstration of a New System
Stephan Ogenstad*, Vertex Pharmaceuticals Incorporated, Peter H. Westfall, Texas Tech University, Kuenhi Tsai, Leif Bengtsson, Scott Moseley and Min Yao, Vertex Pharmaceuticals Incorporated, Alin Tomoiaga and Lan Zhang, Texas Tech University
8:45 Predicting Event Times in Clinical Trials when Randomization is Masked and Blocked
J. Mark Donovan*, University of Pennsylvania, Michael R. Elliott, University of Michigan, Daniel F. Heitjan, University of Pennsylvania
9:00 Effect of Dropouts on Cost-Efficiency of Higher-Order Crossover Designs in Comparative Bioavailability Clinical Trials
Jihao Zhou*, Allergan, Inc., Jane Li, University of Michigan
9:15 Stochastic Curtailment in Multi-Armed Trials
Xiaomin He*, University of Rochester
9:30 Examination of the Efficiency of the Sequential Parallel Design in Psychiatric Clinical Trials
Roy N. Tamura* and Xiaohong Huang, Eli Lilly and Company
9:45 Single-Stage Simultaneous Testing of Superiority and Non-Inferiority in Active Control Clinical Trials
Yongzhao Shao, Vandana Mukhi* and Judith D. Goldberg, New York University School of Medicine
10:00 SCPRT Design for Clinical Trials with Survival Data
Xiaoping Xiong*, St. Jude Children's Research Hospital, Ming Tan, University of Maryland-Greenebaum Cancer Center, James Boyett, St. Jude Children's Research Hospital

Scientific Program: Oral Session Summary

8. CONTRIBUTED PAPERS:
BAYESIAN METHODS AND APPLICATIONS

Buccaneer B

Sponsor: *ENAR*
Chair: *Bo Cai, NIEHS*

8:30 Bayesian Image Analysis of Changes in Brain/Tumor Permeability Induced by Radiotherapy Using Reversible Jump Markov Chain Monte Carlo
Xiaoxi Zhang* and Timothy D. Johnson, University of Michigan

8:45 Reconsidering the Variance Parameterization in Multiple Precision Models
Yi He*, James S. Hodges and Bradley P. Carlin, University of Minnesota

9:00 A Bayesian Screening Procedure for Identifying 'Signals' Obtained by Data Mining from Spontaneous Report Adverse Event Databases
A. Lawrence Gould*, Merck Research Laboratories

9:15 Combining Bootstrap and Bayesian Inferences
Yan Zhou*, Jack D. Kalbfleisch and Roderick J.A. Little, University of Michigan

9:30 Empirical Bayes Estimation for Additive Hazards Regression Models
Sinha Debajyoti* and Stuart Lipsitz, Medical University of South Carolina, Brent McHenry, Bristol-Meyer & Squibb

9:45 Dose-Finding Based on the Maximum Difference in the Probability of Response and the Probability of Toxicity
Yuan Ji, Yisheng Li* and B. Nebiyou Bekele, M.D. Anderson Cancer Center

10:00 A Distance Approach to Bayesian Model Diagnostics
Guan Xing* and J. Sunil Rao, Case Western Reserve University

9. CONTRIBUTED PAPERS:
BIOASSAY AND BIOPHARMACEUTICAL APPLICATIONS

Buccaneer D

Sponsor: *ENAR*
Chair: *Sumithra J. Mandrekar, Mayo Clinic, Division of Biostatistics*

8:30 Stochastic Modeling of Human Colon Cancers: A Mixture Approach
Wai-Yuan Tan and Lijun Zhang*, University of Memphis, Chao-Wen Chen, EPA, Junmei Zhu, University of Memphis

8:45 An Analytical Tool for Assay Development on Protein Chip Platform
Steven Novick*, GlaxoSmithKline

9:00 Assessing Individual Agreement via Individual Equivalence
Huiman X. Barnhart* and Andrzej S. Kosinski, Duke University, Michael J. Haber, Emory University

9:15 Extending Tolerance Intervals for Prediction Interval Coverage
Jacqueline R. Wroughton*, and Erin E. Blankenship, University of Nebraska, James R. Schwenke, Boehringer-Ingelheim Pharmaceuticals Inc., Walter W. Stroup, University of Nebraska

9:30 On Searching for Trend in Gene Expression Using ORIOGEN
Shan Chen, Irene B. Helenowski, Raymond C. Bergan and Borko D. Jovanovic*, Northwestern University

9:45 Model Averaging in Dichotomous Dose-Response Risk Estimation
Matthew W. Wheeler*, Risk Evaluation Branch, NIOSH, A. John Bailer, Risk Evaluation Branch, NIOSH and Miami University

10:00 Assessing Drug Interaction under Different Experimental Conditions
Maiying Kong*, J. Jack Lee and Dan Ayers, M. D. Anderson Cancer Center

10. CONTRIBUTED PAPERS:
GENERALIZED LINEAR MODELS *Esplanade I*

Sponsor: *ENAR*
Chair: *Brisa N. Sanchez, Biostatistics, Harvard School of Public Health*

8:30 Combining Studies to Calculate a Bound on a Regression Coefficient
Chand K. Chauhan*, Yvonne M. Zubovic, Indiana-Purdue University-Fort Wayne

8:45 A Practical Approach to Computing Power for Generalized Linear Models with Nominal, Count, or Ordinal Responses
Robert H. Lyles*, Emory University, Hung-Mo Lin, Penn State College of Medicine, John M. Williamson, Centers for Disease Control and Prevention

9:00 **A Class of Markov Models for Longitudinal Ordinal Data**
Keunbaik Lee* and Michael Daniels, University of Florida

9:15 Testing Homogeneity in Finite Mixtures and Mixture Regression Models
Hongying Dai* and Richard Charnigo, University of Kentucky

9:30 Robust Estimation for Zero-Inflated Regression Models
Jing Shen* and Daniel B. Hall, University of Georgia

Scientific Program: Oral Session Summary

9:45 A Comparison of Models for Immunological Correlates of Protection
Fabrice Bailleux*, Sanofi Pasteur-Lyon, France, Andrew Dunning, Sanofi Pasteur-Swiftwater, USA

10:00 Model Selection for Generalized Linear Model
Bo Hu*, Jun Shao, Mari Palta, University of Wisconsin-Madison

11. CONTRIBUTED PAPERS: CAUSAL INFERENCE *Esplanade 3*

Sponsor: *ENAR*
Chair: *Jesse Berlin, Johnson and Johnson Pharmaceutical Research and Development*

8:30 Estimation and Confidence Regions for Multi-Dimensional Effective Dose
Jialiang Li*, Erik V. Nordheim, Chunming Zhang and Charles E. Lehner, University of Wisconsin

8:45 **Structural Nested Mean Models for Assessing Time-Varying Effect Moderation: An Illustration**
Daniel Almirall*, University of Michigan, Thomas R. Ten Have, University of Pennsylvania School of Medicine, Susan A. Murphy, University of Michigan

9:00 Predicting Treatment Means in a One-Way Factorial Design Based on a Potential Observable Random Variable Framework
Bo Xu* and Edward J. Stanek III, University of Massachusetts

9:15 Machine Learning Methods for Observational Studies
Debashis Ghosh*, University of Michigan

9:30 Causal Analysis of Binary Responses
Haihong Li* and P. V. Rao, University of Florida

9:45 Selection of Average Causal Effect (ACE) Measures for Binary Outcomes Using Propensity Score Subclassification
Yi Huang*, Karen Bandeen-Roche and Constantine Frangakis, Johns Hopkins University Bloomberg School of Public Health

10:00 Heterogeneous Variances in Principal Stratification Models
Robert J. Gallop*, West Chester University, Thomas R. Ten Have, University of Pennsylvania

Monday, March 27
10:15a.m. – 10:30 a.m.

Refreshment Break *Atrium*

Monday, March 27
10:30 a.m. – 12:15 p.m.

12. NEW DEVELOPMENTS IN MICROARRAYS: IDENTIFYING DIFFERENTIALLY EXPRESSED GENES AND METHODS FOR BUILDING PREDICTION MODELS *Regency 3*

Sponsor: *ASA Sections on Teaching Statistics in the Health Sciences/Statistical Education*
Organizer: *Todd G. Nick, Cincinnati Children's Hospital Medical Center*
Chair: *Todd G. Nick, Cincinnati Children's Hospital Medical Center*

10:30 Selection and Use of Pathways for Prognosis
Hans C. van Houwelingen* and Jelle J. Goeman, Leiden University Medical Center-The Netherlands

10:55 Regularized Inference for Microarrays
Hemant Ishwaran*, Cleveland Clinic Foundation

11:20 A Parametric Bootstrap Method for Model Selection in Penalized Logistic Regression for Disease Classification using Microarray Data
Jason Liao* and Yong Lin, University of Medicine and Dentistry of New Jersey

11:45 Variants of the Support Vector Machine and Their Applications to Microarray Classification
Ji Zhu*, University of Michigan

12:10 Floor Discussion

13. ASSESSING SPATIAL SURVEILLANCE FOR BIOTERRORISM: SIMULATING ATTACKS *Regency 6*

Sponsor: *ASA Section on Statistics in Defense and National Security*
Organizer: *Ken Kleinman, Harvard Medical School/Harvard Pilgrim Health Care*
Chair: *Ken Kleinman, Harvard Medical School/Harvard Pilgrim Health Care*

10:30 Detecting Simple Simulated Anthrax Attacks: Comparison of Cluster Identification Methods via New Assessment Metrics
Ken Kleinman* and Allyson Abrams, Harvard Medical School/Harvard Pilgrim Health Care

11:00 A Simulation Model for Evaluating Outbreak Detection
David L. Buckeridge*, McGill University

11:30 Protecting Public Health Through Advanced Spatio Temporal Epidemiological Modeling
James H. Kaufman* and Daniel A. Ford, IBM Research Division

12:00 Discussant: Michael Soto, Rand Corporation

Scientific Program: Oral Session Summary

14. STATISTICAL LEADERSHIP UNDER PHRMA CRITICAL PATH INITIATIVES *Regency 7*

Sponsor: *Biometrics/ASA Biopharmaceutical Section*
Organizer: *Frank Shen, Bristol-Myers Squibb Co.*
Chair: *Frank Shen, Bristol-Myers Squibb Co.*

10:30 Efficiency of Late-Stage Clinical Research (ECR)
Walter W. Offen*, Eli Lilly and Company, Joe Camardo, Wyeth Pharmaceuticals
11:00 Novel Adaptive Clinical Trial Design
Brenda L. Gaydos*, Eli Lilly and Company, Michael Krams, Pfizer, Inc.
11:30 Rolling Dose Studies
José C. Pinheiro*, Novartis Pharmaceuticals, Rick Sax, Astrazeneca
12:00 Discussant: Bob O'Neill, Food and Drug Administration

15. THE ROLE OF NEW DESIGNS FOR EVALUATING VACCINES AND OTHER PREVENTION PROGRAMMES *Regency 2*

Sponsors: *ASA Section on Risk Analysis/ASA Biopharmaceutical Section*
Organizers: *Constantine Frangakis, John Hopkins University*
Chair: *Dennis O. Dixon, NIH/NIAID*

10:30 Statistical Challenges in Combining Human with Animal Studies: The Case of Anthrax Vaccines
Donald B. Rubin*, Harvard University
11:00 Robust Analysis of Therapeutic HIV Vaccination Trials
Devan V. Mehrotra* and Robin Mogg, Merck Research Laboratories
11:30 Augmented Designs to Assess Immune Response in Vaccine Trails
Dean A. Follmann* , NIAID
12:00 Discussant: Constantine Frangakis, John Hopkins University

16. IMS: DIMENSION REDUCTION *Regency 5*

Sponsor: *IMS*
Organizer: *Dennis Cook, University of Minnesota*
Chair: *Dennis Cook, University of Minnesota*

10:30 Dimension Reduction: An Overview
Bing Li*, Pennsylvania State University
10:55 Using Intra-Slice Information for Improved Dimension Reduction
Liqiang Ni*, University of Central Florida
11:20 Sufficient Dimension Reduction for the Small-n-Large-p Problems
Lexin Li*, North Carolina State University, Dennis Cook, University of Minnesota
11:45 An Interactive Method for Sufficient Dimension Reduction
Francesca Chiaromonte*, Penn State University, Dennis Cook, University of Minnesota, Bing Li, Penn State University
12:10 Floor Discussion

17. CONTRIBUTED PAPERS: ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS *Buccaneer A*

Sponsor: *ENAR*
Chair: *Myron J. Katzoff, National Center for Health Statistics*

10:30 Identifying Effect Modifiers in Air Pollution Time-Series Studies Using a Two-Stage Analysis
Sandrah P. Eckel* and Thomas A. Louis, Johns Hopkins University
10:45 Semi-Parametric Modeling of Effects of Air Quality on Respiratory Health in Chicago Medicaid Population
Chava E. Zibman*, Vanja Dukic and Paul Rathouz, University of Chicago
11:00 Bayesian Modeling of Air Pattern for Two-Zone Fields
Yufen Zhang*, Sudipto Banerjee, Gurumurthy Ramachandran and Rui Yang, University of Minnesota
11:15 Use of GAMs to Assess Effects of Air Pollution on Human Health: Our ACAPS Experience
Vincent C. Arena*, Ya-Hsiu Chuang and Sati Mazumdar, University of Pittsburgh
11:30 Experimental Designs for Evaluating the Effectiveness of Rehabilitation Actions in Creating Fish Habitat in the Trinity River
Darcy C. Pickard*, Simon Fraser University & ESSA Technologies
11:45 Bayesian Hierarchical Models in Nest Survival Studies
Jing Cao*, Southern Methodist University, Chong He, Virginia Tech University
12:00 Bayesian Distributed Lag Models: Estimating Effects of Particulate Matter Air Pollution on Daily Mortality
Leah J. Welty*, Northwestern University, Scott L. Zeger and Francesca Dominici, Johns Hopkins Bloomberg School of Public Health

Scientific Program: Oral Session Summary

18. CONTRIBUTED PAPERS:
ADAPTIVE CLINICAL TRIAL DESIGNS AND METHODS
Buccaneer B

Sponsor: *ENAR*
Chair: *Peter B. Imrey, Cleveland Clinic Foundation*

10:30 Screening Designs for Drug Development
Peter Mueller and Gary Rosner, M.D. Anderson Cancer Center, David Rossell*, Rice University/M.D. Anderson Cancer Center

10:45 An Adaptive Phase I Design for Identifying a Dose-Outcome Region for Two Drug Combinations
Sumithra J. Mandrekar*, Daniel J. Sargent and Yue Cui, Mayo Clinic

11:00 Optimal Two-Stage Designs in Phase-II Clinical Trials
Anindita Banerjee* and Anastasios A. Tsiatis, North Carolina State University

11:15 Designing Covariate Adjusted Response Adaptive Randomized Trials in the Presence of Covariate by Treatment Interactions
Ayanbola O. Ayanlowo* and David T. Redden, University of Alabama at Birmingham

11:30 An Adaptive Design in a Dose-Finding Study for the Acute Treatment of Migraine
Vladimir Dragalin*, GlaxoSmithKline

11:45 Adaptive Treatment Allocation with Continuous Covariates: A Comparison of Methods
Nora J. Graber*, Rho, Inc.

12:00 Bayesian Dose-Finding Designs Based on a New Statistical Framework for Phase I Clinical Trials
Yuan Ji*, University of Texas M.D. Anderson Cancer Center

19. CONTRIBUTED PAPERS:
STATISTICAL AND COMPUTATIONAL METHODS FOR GENETIC DATA
Buccaneer C

Sponsor: *ENAR*
Chair: *Jinfeng Xu, Columbia University*

10:30 Relatedness Estimation for Structured Populations
Amanda B. Hepler*, North Carolina State University, Bruce S. Weir, University of Washington

10:45 Coalescent Analysis of Modeling Mutation Progression in Colorectal Cancer
Hui Zhao* and Qingyi Wei, M.D. Anderson Cancer Center, Yun-Xin Fu, University of Texas-Houston Health Science Center, School of Public Health

11:00 Context Dependent Models for Discovery of Transcription Factor Binding Sites
Chuancai Wang*, Penn State College of Medicine, Jun Xie and Bruce A. Craig, Purdue University

11:15 Incorporating Medical Interventions into Mendelian Mutation Prediction Models
Hormuzd A. Katki*, Johns Hopkins Bloomberg School of Public Health and Division of Cancer Epidemiology and Genetics, NCI, NIH, DHHS

11:30 Identifying Novel NF-kB-Regulated Immune Genes in the Human Genome using Structured Support Vector Machine with Discrete Kernel
Insuk Sohn*, Korea University, Seoul, Korea, Sujong Kim, Skin Research Institute, AmorePacific Corporation R&D Center, Kyounggi-do, Korea, Jae Won Lee, Korea University-Seoul, Korea

11:45 A Weighted Regression Model for a Global Test of Haplotype Effects in Case-Control Samples
Anbupalam Thalamuthu* and Daniel E. Weeks, University of Pittsburgh

12:00 Comparing the Joint Distribution of Multiple Categorical Variables between Two Groups: with Application to Analysis of Pre/Post HIV-1 Genotype Sequences
Greg Di Rienzo*, Harvard School of Public Health

20. CONTRIBUTED PAPERS:
MIXED MODELS: LINEAR, GENERALIZED, AND NON-LINEAR
Buccaneer D

Sponsor: *ENAR*
Chair: *Yolanda Munoz, UT-HSC School of Public Health at Houston*

10:30 Bivariate Random Effect Model using Skew Normal Distribution with Application to HIV-RNA
Pulak Ghosh*, Georgia State University, Marcia D. Branco, University of São Paulo, Hrishikesh Chakraborty, RTI International, North Carolina

10:45 Generalized Monotonic Functional Mixed Models for the Effects of Radiation Dose Histograms on Normal Tissue Complications
Matthew J. Schipper* and Jeremy M.G. Taylor, University of Michigan, Xihong Lin, Harvard University

11:00 Semiparametric Approach for the Misaligned Measurements in Colon Carcinogenesis Study
Zonghui Hu*, National Institute of Health, Naisyin Wang, Texas A&M University

11:15 Empirical Bayes Linear Mixed Model Analyses for Two-Color Microarray Experiments
Lan Xiao* and Robert J. Tempelman, Michigan State University

11:30 Statistical Analysis of Dendritic Branching in Hippocampal Neurons
Rebecka J. Jornsten*, Rutgers University

11:45 Floor Discussion

Scientific Program: Oral Session Summary

21. CONTRIBUTED PAPERS:
SEMIPARAMETRIC AND NONPARAMETRIC MODELING
Esplanade 1

Sponsor: *ENAR*
Chair: *Aiyi Liu, NICHD/NIH*

10:30 A Nonparametric Estimate of the Cumulative Incidence Function under Time-Dependent Treatment Assignments
Chung-Chou H. Chang*, University of Pittsburgh, Wei Tian, Inspire Pharmaceuticals, Inc.

10:45 **Efficient Estimation of Population-Level Summaries in General Semiparametric Regression Models with Missing Response**
Arnab Maity*, Yanyuan Ma and Raymond J. Carroll, Texas A&M University

11:00 A Novel Approach to Testing Equality of Survival Distributions when the Population Membership Information is Censored
Dipankar Bandyopadhyay*, University of Georgia, Somnath Datta, University of Louisville

11:15 Measuring Lateral Control in Driving Studies
Jeffrey D. Dawson*, Joseph E. Cavanaugh, K.D. Zamba, Matthew Rizzo, University of Iowa

11:30 A Geometric Approach to Estimation of the Number of Species
Changxuan Mao*, University of California, Riverside

11:45 Estimation of the Mean Function of Panel Count Data Using Monotone Polynomial Splines
Minggen Lu*, Ying Zhang and Jian Huang, University of Iowa

12:00 Nonparametric Ecological Inference: Incorporating Marginal Covariate Information in a Nonparametric Regression Model for Aggregate Data
Joan G. Staniswalis*, University of Texas at El Paso

22. CONTRIBUTED PAPERS:
SPATIAL MODELING *Esplanade 3*

Sponsor: *ENAR*
Chair: *Bradley P. Carlin, University of Minnesota*

10:30 A Composite Likelihood Cross-Validation Approach in Selecting Bandwidth for the Estimation of the Pair Correlation Function
Yongtao Guan*, University of Miami

10:45 Improved Detection of Differentially Expressed Genes through Incorporation of Gene Locations
Guanghua Xiao*, Cavan Reilly, Betsy M. Martinez-Vaz, Wei Pan and Arkady Khodursky, University of Minnesota

11:00 **Hierarchical and Joint Site-Edge Methods for Areal Boundary Analysis**
Haijun Ma*, Bradley P. Carlin and Sudipto Banerjee, University of Minnesota

11:15 The Effect of Aggregation on Inferences using Small Area Health Data
Sandy Burden* and David G. Steel, University of Wollongong

11:30 Statistical Comparison of Observed and Multi-Resolution CMAQ Modeled Ozone Concentrations
Li Chen* and Michael L. Stein, University of Chicago

11:45 Parameterization of Spatial Models and Stability of Estimates
Petruta C. Caragea, Mark S. Kaiser, and Kyoji Furukawa, Iowa State University

12:00 Modeling the Evolution of an Air-Borne Contaminant Release in an Urban Environment
Margaret B. Short*, Los Alamos National Laboratory

Monday, March 27
12:15 p.m. – 1:30 p.m.

ROUNDTABLE LUNCHEONS *City Center*
(registration required)

Scientific Program: Oral Session Summary

Monday, March 27
1:45 p.m. – 3:30 p.m.

23. STATISTICAL ISSUES IN GENETIC INVESTIGATIONS
Regency 3

Sponsor: *ASA Biopharmaceutical Section*
Organizers: *Boris Iglewicz, Temple University*
Chair: *Boris Iglewicz, Temple University*

1:45 Controlling False Discoveries and False Non-Discoveries in Microarray Analysis
Sanat K. Sarkar*, Temple University
2:15 Changing Expressions: The Evolution of Information Technology Applied to Gene Expression
Daniel J. Holder*, Merck Research Laboratories
2:45 Approaches to Analysis of Non-Gaussian Clustered Data from Genetic Animal Studies
Inna Chervoneva*, Thomas Jefferson University
3:15 Discussant: Sue-Jane Wang, FDA

24. STATISTICAL MODELS IN MICROPARTICLE REMEDIATION/DECONTAMINATION *Regency 6*

Sponsor: *ASA Section on Statistics in Defense and National Security*
Organizer: *Myron Katzoff, National Center for Health Statistics*
Chair: *Joe Fred Gonzalez, Jr., National Center for Health Statistics*

1:45 Experimental Validation of Contaminant Concentrations Predicted by a Deterministic Model
Myron J. Katzoff* and Abera Wouhib, National Center for Health Statistics/CDC, Stanley A. Shulman, James S. Bennett and William K. Sieber, National Institute for Occupational Safety and Health
2:15 Estimating Tracer Gas Distribution in a Ventilation Chamber
James S. Bennett, Stanley A. Shulman* and W. Karl Sieber, National Institute for Occupational Safety and Health, Myron Katzoff and Abera Wouhib, National Center for Health Statistics, Brian Adams, South Dakota School of Mines
2:45 A Comparison of Room Contamination Fields Estimated via Kriging and Deterministic Air Flow Models
James S. Bennett*, National Institute for Occupational Safety and Health, Sean A. McKenna and Patrick D. Finley, Sandia National Laboratories, Stanley A. Shulman and W. Karl Sieber, National Institute for Occupational Safety and Health, Myron Katzoff and Abera Wouhib, National Center for Health Statistics, John E. Brockman and Richard O. Griffith, Sandia National Laboratories
3:15 Floor Discussion

25. INFERENCE IN RANDOMIZED MULTI-CENTER CLINICAL TRIALS *Regency 7*

Sponsor: *ASA Biopharmaceutical Section*
Organizer: *Marvin Zelen, Harvard School of Public Health*
Chair: *L.J. Wei Harvard School of Public Health*

1:45 Conditioning on the Sample Space: A Method to Adjust for Large Numbers of Institutions without Introducing Parameters
Lu Zheng* and Marvin Zelen, Harvard School of Public Health
2:15 Design vs Model Based Analysis
John M. Lachine*, The George Washington University
2:45 Local vs Global Inference
Marvin Zelen*, Harvard University
3:15 Discussant David DeMets, University of Wisconsin

26. ILS✧: INTRODUCTION TO LONGITUDINAL DATA
Regency 2

Sponsor: *ASA Sections on Teaching Statistics in the Health Sciences/Statistical Education*
Organizer: *Paul Rathouz, University of Chicago*
Chair: *Paul Rathouz, University of Chicago*

1:45 Introduction to Longitudinal Data
Marie Davidian, North Carolina State University
3:15 Floor Discussion

27. IMS: RECENT ADVANCES IN MIXTURE MODELS
Regency 5

Sponsor: *IMS*
Organizer: *Bruce Lindsay, The Pennsylvania State University*
Chair: *Bruce Lindsay, The Pennsylvania State University*

1:45 Semiparametric Analysis in Conditional Independence Latent Class Models
Jing Qin*, National Institute of Allergy and Infectious Disease, NIH, Denis Leung, Singapore Management University
2:15 Clustering Based on a Multi-Layer Mixture Model
Jia Li*, The Pennsylvania State University
2:45 Generalizing Hodges-Lehmann: Nonparametric Inference for Location Mixtures
David R. Hunter*, The Pennsylvania State University
3:15 Floor Discussion

✧ ILS stands for Introductory Lecture Session.

Scientific Program: Oral Session Summary

28. CONTRIBUTED PAPERS:
COPULA AND COX REGRESSION MODELS

Buccaneer A

Sponsor: *ENAR*
Chair: *Jay Mandrekar, Division of Biostatistics, Mayo Clinic*

1:45 Sensitivity Analysis for Shared Parameter Models using Copulas
Dimitris Rizopoulos*, Geert Verbeke and Emmanuel Lesaffre, Catholic University of Leuven

2:00 Modeling Bivariate Survival Times by Copulas
Rui Qin* and Michael P. Jones, The University of Iowa

2:15 An Analog Parameter Estimator for Copula Models
Antai Wang*, Georgetown University, David Oakes, University of Rochester

2:30 Multi-Dimensional Copula Regression Models for Correlated Mixed/Censored Outcomes
Mingyao Li*, University of Pennsylvania, Peter X.K. Song, University of Waterloo-Canada

2:45 Estimation of Survival Functions and Covariate Effects Based on an Assumed Copula Accounting for Dependent and Independent Censoring
Xuelin Huang*, The University of Texas-MD Anderson Cancer Center, Gretchen A. Fix and Katherine B. Ensor, Rice University

3:00 New Approach to Directional Dependence using Copula Function
Yoonsung Jung*, Kansas State University, Jong-Min Kim and Engin A. Sungur, University of Minnesota-Morris

3:15 Inferences of Change Point in Piecewise Cox Model
Zhiying Xu* and Pingfu Fu, Case Western Reserve University

29. CONTRIBUTED PAPERS:
BIOMARKERS AND SURROGATE MARKERS *Buccaneer C*

Sponsor: *ENAR*
Chair: *Yan D. Zhao, Eli Lilly and Company*

1:45 Surrogate Marker Validation from an Information Theory Perspective
Ariel A. Abad* and Geert Molenberghs, Hasselt University

2:00 Combining Logistic Regression Models for Multiple Biomarkers
Zheng Yuan* and Debashis Ghosh, University of Michigan

2:15 Joint Analysis of Multiple Longitudinal Biomarkers and Tumor Count Data
Yulin Zhang* and KyungMann Kim, University of Wisconsin-Madison

2:30 On Combining Diagnostic Markers
Ruth Pfeiffer, National Cancer Institute, Efstathia Bura*, George Washington University

2:45 The Use of Surrogate Markers on Early Treatment Comparison in a Meta-Analysis Framework
Yun Li* and Jeremy M.G. Taylor, University of Michigan

3:00 Floor Discussion

30. CONTRIBUTED PAPERS:
CLUSTERING, CLASSIFICATION, AND IDENTIFICATION METHODS *Buccaneer B*

Sponsor: *ENAR*
Chair: *A. James O'Malley, Harvard Medical School*

1:45 Modification of Conventional Edge Detectors for Segmentation of Spotted Microarray Images
Jingran Sun*, Amgen Inc., Peihua Qiu, University of Minnesota

2:00 MICE: Multiple-Peak Identification, Characterization and Estimation
Nicoleta Serban*, Georgia Institute of Technology

2:15 ***Mixed Membership Stochastic Block Models for Relational Data With Application to Protein-Protein Interactions***
Edoardo M. Airoldi*, David M. Blei, Stephen E. Fienberg and Eric P. Xing, Carnegie Mellon University

2:30 **A Framework for Kernel Regularization with Application to Protein Clustering**
Fan Lu*, Sunduz Keles, Stephen J. Wright and Grace Wahba, University of Wisconsin-Madison

2:45 Comparison of Classification Methods to Predict Complications to Liver Surgery
Leah Ben-Porat*, Mithat Gonen and William Jarnigan, Memorial Sloan Kettering Cancer Center

3:00 Model-Based Projection Pursuit Clustering
Jie Ding*, GlaxoSmithKline

3:15 Multinomial Group Testing Model with Small-Sized Pools and Application to California HIV Data: Bayesian and Bootstrap Approaches
Jong-Min Kim*, University of Minnesota-Morris, Tae-Young Heo, Electronics and Telecommunications Research Institute-South Korea

Scientific Program: Oral Session Summary

31. CONTRIBUTED PAPERS: SURVEY DATA AND SAMPLING METHODS

Buccaneer D

Sponsor: *ENAR*
Chair: *Timothy R. Church, University of Minnesota*

1:45 The Job Outlook for Biostatistics Graduates
Joseph L. Hagan* and Stephen W. Looney, Biostatistics Program LSU HSC School of Public Health

2:00 Computing Inclusion Probabilities for Constructing Horvitz-Thompson Estimators of Sampling Plans Excluding Neighboring Units
Kyoungah See*, Robert Noble and A. John Bailer, Miami University

2:15 Estimating the Distribution Function Using k-Tuple Ranked Set Samples
Kaushik Ghosh*, George Washington University, Ram C. Tiwari, National Cancer Institute

2:30 Covariate Adjustment May Not Be Better: Thresholds of Relative Risk for Reductions in MSE
Wenjun Li* and Edward J. Stanek III, University of Massachusetts

2:45 Modelling Rare Events Using Generalized Inverse Sampling Scheme
Soumi Lahiri* and Sunil K. Dhar, New Jersey Institute of Technology

3:00 Weighted Proportional Hazards Models for Biased Samples with Estimated Weights
Qing Pan* and Douglas E. Schaubel, University of Michigan

3:15 Predicting Realized Cluster Means in Unequally Sized Cluster Populations
Edward J. Stanek III*, University of Massachusetts, Julio M. Singer, University of Sao Paulo-Brazil

32. CONTRIBUTED PAPERS: LINKAGE ANALYSIS

Esplanade I

Sponsor: *ENAR*
Chair: *Jialiang Li, Department of Statistics, University of Wisconsin*

1:45 Ascertainment Adjustment in Genetic Studies of Ordinal Traits
Rui Feng*, University of Alabama at Birmingham, Heping Zhang, Yale University

2:00 **Interval Mapping for Expression Quantitative Trait Loci**
Meng Chen*, Christina Kendziorski and Alan Attie, University of Wisconsin-Madison

2:15 Linkage Tests for Affected Relative-Pairs with Incomplete IBD and Known IBS
Dennis W. Buckman*, IMS Inc., Zhaohai Li, George Washington University

2:30 Evaluating Admixture Estimation and Mapping Techniques through Plasmodes
Laura K. Vaughan*, Jasmin Divers, Miguel Padilla, Hemant K. Tiwari, David T. Redden and David B. Allison, University of Alabama at Birmingham

2:45 Extension of Variance Component Linkage Analysis to Incorporate Repeated Measurements
Wei-Min Chen* and Liming Liang, University of Michigan, Pak C. Sham, University of London and University of Hong Kong, Gonçalo R. Abecasis, University of Michigan

3:00 Confidence Set Inference on Maximum Load Score Statistic in Linkage Analysis: Solving the Problem of Multiple Testing
Ritwik Sinha* and Yuqun Luo, Case Western Reserve University

3:15 Semiparametric Transformation Models for Mapping Quantitative Trait Loci with Censored Data
Guoqing Diao* and Danyu Lin, University of North Carolina at Chapel Hill

Scientific Program: Oral Session Summary

33. CONTRIBUTED PAPERS:
IMAGING METHODS *Esplanade 3*

Sponsor: *ENAR*
Chair: *Howard Bondell, North Carolina State University*

1:45 High Dimension, Low Sample Size Principal Components
Keith E. Muller, University of North Carolina-Chapel Hill, Yueh-Yun Chi*, University of Washington, Jeongyoun Ahn and Steve Marron, University of North Carolina-Chapel Hill

2:00 Analyzing Diffusion Tensor Imaging Data
Meagan E. Clement*, Rho, Inc.

2:15 Estimation Efficiency and Statistical Power in Arterial Spin Labeling FMRI
Jeanette A. Mumford*, Luis Hernandez-Garcia, Gregory R. Lee and Thomas E. Nichols, University of Michigan

2:30 **Functional Principal Component Regression and Functional Partial Least Squares**
Philip T. Reiss* and R. Todd Ogden, Columbia University

2:45 Spatiotemporal Modeling of Functional Magnetic Resonance Imaging Data
Qihua Lin*, Southern Methodist University, Patrick S. Carmack, University of Texas Southwestern Medical Center at Dallas, Richard F. Gunst and William R. Schucany, Southern Methodist University, Jeffrey S. Spence, University of Texas Southwestern Medical Center at Dallas

3:00 Prediction of Post-Treatment Brain Activity using a Bayesian Hierarchical Model
Ying Guo* and DuBois Bowman, Emory University"

3:15 Floor Discussion

Monday, March 27
3:30 p.m. – 3:45 p.m.

Refreshment Break *Atrium*

Monday, March 27
3:45 p.m. – 5:30 p.m.

34. STATISTICAL ISSUES IN USING EXPOSURE ESTIMATES IN ENVIRONMENTAL EPIDEMIOLOGY *Regency 3*

Sponsors: *ASA Sections on Epidemiology/Statistics and the Environment*
Organizer: *Chris Paciorek, Harvard School of Public Health*
Chair: *Chris Paciorek, Harvard School of Public Health*

3:45 Semiparametric Dynamic Structural Models for Multivariate Exposures
Amy H. Herring*, The University of North Carolina at Chapel Hill, David B. Dunson, National Institute of Environmental Health Sciences-National Institutes of Health

4:10 Source Apportionment: From Characterization to Imputation
Thomas Lumley*, University of Washington

4:35 Exposure Measurement Error Caused by Spatial Misalignment in Environmental Epidemiology
Alexandros Gryparis*, Christopher Paciorek and Brent A. Coull, Harvard University

5:00 Model Choice in Time Series Studies of Air Pollution and Mortality
Roger D. Peng*, Francesca Dominici and Thomas A. Louis, Johns Hopkins University

5:25 Floor Discussion

35. SOLUTIONS FOR MISSING DATA IN COMPLEX SAMPLE SURVEYS RELEVANT IN HEALTH POLICY RESEARCH *Regency 6*

Sponsor: *ASA Health Policy Statistics Section*
Organizer: *Recai Yucel, University of Massachusetts*
Chair: *James O'Malley, Harvard Medical School*

3:45 Integrated Design and Estimation Strategies to Correct for Missing Data in the Medical Expenditure Panel Survey
Steven B. Cohen*, AHRQ

4:15 Multiple Imputation for Non-Normal Continuous Missing Variables in Complex Surveys
Yulei He*, Harvard University, Trivellore E. Raghunathan, University of Michigan

4:45 Missing Data in Cluster Samples: Design-Based and Bayesian Perspectives
Recai M. Yucel*, University of Massachusetts, Joseph L. Schafer, The Pennsylvania State University

5:15 Floor Discussion

Scientific Program: Oral Session Summary

36. FUSING BIOMEDICAL / ENVIRONMENTAL DATA WITH NUMERICAL MODELS *Regency 7*

Sponsors: *Joint IMS/ENAR and ASA Statistics and the Environment*
Organizers: *Dennis D. Boos and Montserrat Fuentes, North Carolina State University*
Chair: *Dennis Boos, North Carolina State University*

3:45 Estimating Parameters for Huge Systems: Tuning the Community Atmosphere Model
Douglas W. Nychka*, National Center for Atmospheric Research
4:15 Statistical Methods for Data Assimilation Applied to Hurricane Forecasting
Montserrat Fuentes* and Kristen Foley, North Carolina State University
4:45 Spatio-Temporal Dynamics of the Spread of Raccoon Rabies. Lance Waller, Emory University
5:15 Discussant: Sudipto Banerjee, University of Minnesota

37. STATISTICAL METHODS FOR PUBLIC HEALTH STUDIES IN DEVELOPING COUNTRIES *Regency 2*

Sponsor: *ASA Section on Epidemiology*
Organizer: *Dylan Small, Wharton School/University of Pennsylvania*
Chair: *Dylan Small, Wharton School/University of Pennsylvania*

3:45 Accounting for Variability in Sample Size Estimation with Application to a Malaria Vaccine Phase 2 Trial
Michael P. Fay*, National Institute of Allergy and Infectious Diseases, M.E. Halloran, Emory University, Dean A. Follmann, National Institute of Allergy and Infectious Diseases
4:10 Design and Analysis of Cluster-Randomized Phased Implementation Studies
Lawrence H. Moulton*, Johns Hopkins Bloomberg School of Public Health
4:35 Analysis of Treatment Effects When the Treatment and Outcome Are Spatially Correlated with Application to a Government Food Relief Program in Bangladesh
Dylan S. Small*, The Wharton School, University of Pennsylvania
5:00 Measurement Issues Related to Quantifying Oligosaccharides in Human Milk to Determine their Association with Infant Diarrhea
Mekibib Altaye*, Cincinnati Children's Hospital Medical Center and University of Cincinnati
5:25 Floor Discussion

38. IMS: SPATIOTEMPORAL STATISTICS *Regency 5*

Sponsor: *IMS*
Organizers: *Richard Smith, University of North Carolina at Chapel Hill*
Chair: *Jason Fine, University of Wisconsin*

3:45 A Modification of the EM Algorithm with Application to Spatio-Temporal Models
Stanislav Kolenikov*, University of Missouri
4:15 Estimating Deformations of Isotropic Gaussian Random Fields
Ethan B. Anderes*, University of California-Berkeley
4:45 Bayesian Modeling of Extreme Precipitation Return Levels
Daniel S. Cooley*, NCAR/Colorado State University, Douglas Nychka, NCAR, Philippe Naveau, University of Colorado/LSCE-CNRS-IPSL
5:15 Discussant: Richard L. Smith, University of North Carolina at Chapel Hill

39. CONTRIBUTED PAPERS: COMPUTATIONAL, CLASSIFICATION, AND MODEL SELECTION METHODS *Buccaneer D*

Sponsor: *ENAR*
Chair: *Benjamin E. Leiby, University of Pennsylvania*

3:45 Model Complexity and the AIC Statistic for Neural Networks
Doug Landsittel* and Dustin Ferris, Duquesne University
4:00 On Creating Model Assessment Tools Independent of Sample Size
Jiawei Liu*, Georgia State University, Bruce G. Lindsay, Penn State University
4:15 Classification of Psychotropic Drugs Based on Sleep - Waking Behavior in Rats
Kristien Wouters and José Cortiñas Abrahantes*, Hasselt University-Diepenbeek, Belgium, Abdellah Ahnaou and Helena Geys, J&J PRD Janssen Pharmaceutica-Beerse, Belgium, Geert Molenberghs, Hasselt University-Diepenbeek, Belgium, Pim Drinkenburg, J&J PRD Janssen Pharmaceutica-Beerse, Belgium
4:30 Estimation of Stochastically Ordered Survival Functions by Geometric Programming
Johan Lim, Texas A&M University, Xinlei Wang*, Southern Methodist University, Seung Jean Kim, Stanford University

Scientific Program: Oral Session Summary

4:55 RANOVA: A New Method of Detecting Differentially Expressed Genes through Probe Level Data from Oligonucleotide Arrays
Jin Xu*, Timothy S. Davison and Charles D. Johnson, Ambion Services, Ambion Inc.
5:00 Floor Discussion

40. CONTRIBUTED PAPERS:
SURVIVAL ANALYSIS I *Buccaneer A*

Sponsor: *ENAR*
Chair: *Kirsten Doehler, North Carolina State University*

3:45 Estimating Time to Event from Longitudinal Categorical Data: An Analysis of Multiple Sclerosis Progression
Micha Mandel*, Harvard School of Public Health, Susan A. Gauthier, Charles R.G. Guttmann and Howard L. Weiner, Brigham and Women's Hospital, Rebecca A. Betensky, Harvard School of Public Health
4:00 Methods for the Accelerated Failure Time Model
Zhezhen Jin*, Columbia University
4:15 A Dynamic Model for Survival Data with Longitudinal Covariates
Gary L. Rosner, The University of Texas-M.D. Anderson Cancer Center, Krzysztof J. Rudnicki*, Rice University
4:30 Checking the Censored Two-Sample Accelerated Life Model using Integrated Cumulative Hazard Difference
Seung-Hwan Lee*, Illinois Wesleyan University
4:45 Semiparametric Survival Models with Censored Covariates
Gina M. D'Angelo* and Lisa Weissfeld, University of Pittsburgh
5:00 Nonparametric Regression using Kernel Estimating Equations for Correlated Failure Time Data
Zhangsheng Yu*, University of Michigan, Xihong Lin, Harvard School of Public Health
5:15 On Sample Size Selection in Clinical Trials with Both Accrual and Follow-Up Periods for Several Treatment Groups
Susan Halabi*, Duke University, Bahadur Singh, Cancer Center Biostatistics, Duke University and Linberger Cancer Center, UNC

41. CONTRIBUTED PAPERS:
METHODS IN EPIDEMIOLOGY *Esplanade I*

Sponsor: *ENAR*
Chair: *Barbra Richardson, Department of Biostatistics, University of Washington*

3:45 Constructing Better Binomial Confidence Intervals by Remembering Two Techniques for Normal Confidence Intervals
Craig B. Borkowf*, US Centers for Disease Control and Prevention
4:00 Some Thoughts on the Relative Survival Rate
Chris M. Drake*, University of California, Davis, Julie Smith-Gagen, Center for Health Data and Research
4:15 A Likelihood-Based Approach to Quantification of the Spread of an Infectious Disease
Laura F. White* and Marcello Pagano, Harvard School of Public Health
4:30 More Realistic Assumptions for Controlling Confounding in Observational Studies of Time Varying Exposures
Marshall M. Joffe*, University of Pennsylvania
4:45 Floor Discussion

42. CONTRIBUTED PAPERS:
BAYESIAN METHODS IN GENOMICS DATA ANALYSIS
Buccaneer C

Sponsor: *ENAR*
Chair: *Tapan K. Nayak, George Washington University*

3:45 Bayesian Analysis of Loss of Heterozygosity by Modeling of Frequency of Allelic Loss Data
Hanwen Huang*, Fei Zou and Fred A. Wright, University of North Carolina at Chapel Hill
4:00 Bayesian Hierarchical Model to Detect QTL
Susan J. Simmons*, Edward Boone and Ann E. Stapleton, University of North Carolina Wilmington
4:15 Semiparametric Bayesian Inference for Sage Data-Model Based Clustering for Count Data
Michele Guindani* and Peter Mueller, The University of Texas M.D. Anderson Cancer Center
4:30 A Compositional Retrospective Analysis of Microarray Data
Jingqin Luo*, Edwin S. Iversen and Merlise A. Clyde, Duke University
4:45 Using Clustering to Enhance Hypothesis Testing
David B. Dahl*, Texas A&M University, Michael A. Newton, University of Wisconsin-Madison
5:00 Floor Discussion

Scientific Program: Oral Session Summary

43. CONTRIBUTED PAPERS: LONGITUDINAL DATA ANALYSIS

Buccaneer B

Sponsor: *ENAR*
Chair: *James Carpenter, Institut für Medizinische Biometrie*

3:45 Role of Alternative Statistical Methods of CD4 Cell Count Extrapolation in Quantifying HIV Related Morbidity
Bingxia Wang*, Boston University

4:00 Signal Intensity Processing Based on Non-Linear Mixed Modeling to Study Changes in Neuronal Activity
Jan Serroyen* and Geert Molenberghs, Hasselt University-Diepenbeek, Belgium, Marleen Verhoye, Vincent Van Meir and Annemie Van der Linden, University of Antwerp-Antwerp, Belgium

4:15 Prediction of Renal Graft Failure using Multivariate Longitudinal Profiles
Steffen Fieuws* and Geert Verbeke, K.U.Leuven, Belgium

4:30 Robustness in Joint Modeling of a Primary Regression Model and a Longitudinal Process
Xianzheng Huang*, Leonard Stefanski and Marie Davidian, North Carolina State University

4:45 Multivariate Hidden Markov Processes: Application to Coronary Vascular Disease Progression
Melanie M. Wall*, University of Minnesota, Judith Rousseau, Université Paris and Chantal Guihenneuc-Jouyaux, Université Paris 5

5:00 Failed Clinical Trials and Failed Analyses: New Help from Trajectory-Based Analyses
Ralitza Gueorguieva*, Ran Wu, Brian Pittman, Stephanie O'Malley and John Krystal, Yale University

5:15 Floor Discussion

44. CONTRIBUTED PAPERS: MEASUREMENT ERROR

Esplanade 3

Sponsor: *ENAR*
Chair: *Jeffrey D. Dawson, University of Iowa*

3:45 An Estimating Equations Approach to Fit Latent Exposure Models
Brisa N. Sánchez* and Louise M. Ryan, Harvard School of Public Health

4:00 Locally Efficient Estimators for Semiparametric Models With Measurement Error
Yanyuan Ma* and Raymond J. Carroll, Texas A&M University

4:15 Statistical Methods for Measurement Comparison in Clinical Studies
Jing Han*, St. Francis Hospital

4:30 **Operating Characteristics of Group Testing Algorithms for Case Identification in the Presence of Test Error**
Hae-Young Kim*, Michael G. Hudgens, Jonathan Dreyfuss, Daniel J. Westreich and Christopher D. Pilcher, University of North Carolina at Chapel Hill

4:45 Methods for Cox Regression with Non-Classical Measurement Error in the Covariates
Pamela A. Shaw*, University of Washington, Ross L. Prentice, Fred Hutchinson Cancer Research Center

5:00 Estimation for Generalized Structural Equation Models Without Normality Assumptions on the Continuous Factors
Jia Guo* and Melanie M. Wall, University of Minnesota, Yasuo Amemiya, IBM T.J. Watson Research Center

5:15 Measurement Error in Population Dynamic Models
John P. Buonaccorsi* and John Staudenmayer, University of Massachusetts

Scientific Program: Oral Session Summary

Tuesday, March 28
8:30 a.m. – 10:15 a.m.

45. NEW STATISTICAL METHODS IN GENETIC EPIDEMIOLOGY *Regency 6*

Sponsors: *Biometrics/ASA Section on Epidemiology*
Organizer: *Paul Rathouz, University of Chicago*
Chair: *Mingyao Li, University of Pennsylvania School of Medicine*

8:30 Detecting Linkage Disequilibrium in the Presence to Locus Hetergeneity
Jian Huang*, University of Iowa, Deli Wang, University of Alabama at Birmingham
8:55 Analysis of Complex Traits with Ordinal Outcome Data
Heping Zhang*, Xueqin Wang and Yuanqing Ye, Yale University
9:20 Nonparametric Pathway Based Regression Methods for Assessing Gene-Gene and Gene-Environment Interactions
Hongzhe Li*, University of Pennsylvania
9:45 Family-Based Haplotype Studies
Glen A. Satten*, Centers for Disease Control and Prevention, Andrew S. Allen, Duke University, Anastasios A. Tsiatis, North Carolina State University
10:10 Floor Discussion

46. DEALING WITH MISSING DATA: ARE WE THERE YET? *Regency 7*

Sponsor: *Statistics in Medicine*
Organizer: *Jay N. Mandrekar, Mayo Clinic, Division of Biostatistics*
Chair: *Jay N. Mandrekar, Mayo Clinic, Division of Biostatistics*

8:30 Missing Data in Infectious Disease Research
Barbra A. Richardson*, University of Washington
9:55 Strategies for Missing Patient Reported Outcomes
Diane L. Fairclough*, University of Colorado Health Sciences Center
9:20 Sensitivity Analysis for Incomplete Clinical Trial Data
Geert Molenberghs*, Universiteit Hasselt-Diepenbeek, Belgium
9:45 The Complexities, Complications and Contributions of Dealing with Missing Data in Clinical Trials
Ralph B. D'Agostino, Sr.*, Joseph M. Massaro and Lisa Sullivan, Boston University
10:10 Floor Discussion

47. STATISTICAL ISSUES IN ENVIRONMENTAL MONITORING *Regency 3*

Sponsor: *ASA Section on Statistics and the Environment*
Organizer: *Jun Zhu, University of Wisconsin*
Chair: *Linda Young, University of Florida*

8:30 Statistics Issues in Designing an Optimal Detection System with Multiple Sensors
Carol Y. Lin*, Lance A. Waller, Robert H. Lyles and Barry P. Ryan, Emory University
8:55 Optimal Network Design for Spatial Prediction, Covariance Parameter Estimation, and Empirical Prediction
Dale L. Zimmerman*, University of Iowa
9:20 Estimation for Lonitudinal Surveys with Repeated Panels of Observations
Jason C. Legg*, Wayne A. Fuller and Sarah M. Nusser, Iowa State University
9:45 Spatial Lasso with Application to GIS Model Selection
Jay Breidt*, Colorado State University, Nan-Jung Hsu, National Tsing-Hua University, Hsin-Cheng Huang, Academica Sinica, Dave Theobald, Colorado State University
10:10 Floor Discussion

48. ILS✧: INTRODUCTION ON DATA MINING AND ITS RECENT DEVELOPMENT *Regency 2*

Sponsor: *ASA Sections on Teaching Statistics in the Health Sciences/Statistical Education*
Organizer: *Hao Helen Zhang, North Carolina State University*
Chair: *Bin Cheng, Columbia University*

8:30 Classification with Support Vector Machine and Psi-Learning
Yufeng Liu*, University of North Carolina
8:55 Feature Selection and Clustering for High Dimensional Data
Xiaodong Lin*, University of Cincinnati
9:20 High Dimensional, Low Sample Size Data Analysis: Data Piling and Geometric Representation
Jeongyoun Ahn* and Steve Marron, University of North Carolina at Chapel Hill
9:45 Visualization Challenges in Internet Traffic Research
Cheolwoo Park*, University of Georgia, Barbara Gonzalez, University of Louisiana at Lafayette, Felix Hernandez-Campos and Steve Marron, University of North Carolina at Chapel Hill
10:10 Floor Discussion

✧ ILS stands for Introductory Lecture Session.

Scientific Program: Oral Session Summary

49. IMS: SURVIVAL ANALYSIS AND GENETIC EPIDEMIOLOGY

Regency 5

Sponsor: *IMS*
Organizer: *David Glidden, University of California*
Chair: *Nilanjan Chatterjee, NIH*

8:30 Semiparametric Normal Models for Multivariate Survival Data in Family Studies
Malka Gorfine*, Technion Institute-Israel, Ross L. Prentice and Li Hsu, Fred Hutchinson Cancer Research Center

9:00 Nonparametric Association Analysis of Multivariate Competing Risks Data
Yu Cheng*, Jason P. Fine and Michael R. Kosorok, University of Wisconsin-Madison

9:30 New Developments in the Analysis of Familial Aggregation
Rebecca A. Betensky*, Harvard School of Public Health

10:00 Floor Discussion

50. CONTRIBUTED PAPERS: GRAPHICAL ANALYSIS AND REPORTING OF CLINICAL SAFETY AND EFFICACY DATA

Buccaneer B

Sponsor: *ENAR*
Chair: *Tim Hesterberg, Insightful Corporation.*

8:30 Graphical Approaches to the Analysis of Safety Data from Clinical Trials
Ohad Amit*, Lane W. Peter and Shi-tao Yeh, GlaxoSmithKline, Richard Heiberger, Temple University

8:45 Visual Representations of Data Used During the NDA Review Cycle
Mat Soukup*, Food and Drug Administration

9:00 Now Look at This: Concepts for Visualizing Clinical Data
Andreas Krause*, Pharsight Corporation

9:15 The State of Data Visualization in the Reporting of Clinical Results
Matthew D. Austin*, Amgen, Inc

9:30 Statistical Graphics in Drug Discovery and Development
Michael A. O'Connell*, Insightful Corporation

9:45 Graphical Analysis and Reporting of Clinical Safety and Efficacy Data
Tom Filloon, Proctor & Gamble

10:00 Floor Discussion

51. CONTRIBUTED PAPERS: RANDOM EFFECTS AND FRAILTY MODELS

Buccaneer A

Sponsor: *ENAR*
Chair: *Christiana Drake, University of California-Davis*

8:30 Assessing the Effectiveness of Potential Longitudinal Biomarkers in Multivariate Survival Analysis
Feng-shou Ko* and Stewart J. Anderson, University of Pittsburgh Graduate School of Public Health

8:45 Semiparametric Analysis of Correlated Recurrent and Terminal Events
Yining Ye*, John D. Kalbfleisch and Douglas E. Schaubel, University of Michigan

9:00 Application of Recurrent Event Data Analysis Methodologies in Clinical Trial Studies
Xiaohong Zhang*, Iowa State University, Matt Austin and Li Chen, Amgen, Inc.

9:15 Shared Frailty Models for Grouped Multivariate Survival Data
Denise A. Esserman*, Columbia University, Andrea B. Troxel, University of Pennsylvania

9:30 A Model-Based Measure of Inter-Rater Agreement
Kerrie P. Nelson*, Max Planck Institute for Demographic Research, Don Edwards, University of South Carolina

9:45 Longitudinal, Multivariate Trajectory Models for Estimating American Disability
Jason T. Connor*, Stephen A. Fienberg and Daniel S. Nagin, Carnegie Mellon University

10:00 Multiple Comparisons with Several Methods
Jixiang Wu*, Mississippi State University, Johnie N. Jenkins and Jack C. McCarty, USDA-ARS-Mississippi State

Scientific Program: Oral Session Summary

52. CONTRIBUTED PAPERS:
ANALYZING HIGH DIMENSIONAL GENOMICS DATA
Buccaneer C

Sponsor: *ENAR*
Chair: *Steven Novick, GlaxoSmithKline*

8:30 Characterizing the Genetic Structure from Single-Nucleotide Polymorphism Data
Xi Chen*, North Carolina State University, Bruce S. Weir, University of Washington

8:45 Analysis Methods for Illumina DASL Data
Karla V. Ballman*, Mayo Clinic College of Medicine

9:00 **A Pseudolikelihood Approach for Simultaneous Analysis of Array Comparative Genomic Hybridizations (aCGH)**
David A. Engler*, Harvard University, Gayatry Mohapatra and David N. Louis, Massachusetts General Hospital, Rebecca A. Betensky, Harvard University

9:15 High Dimensional Phenotype Scoring of Hand Osteoarthritis Data using Exploratory Multivariate Analysis
Sergio Eslava*, Kwan R. Lee, Keith Crowland and Uzma Atif, GlaxoSmithKline

9:30 A Wavelet-Based Approach to Clustering Time-Dependent Gene Expression Profiles
Bong-Rae Kim*, Ramon C. Littell and Rongling Wu, University of Florida

9:45 Statistical Performance of Cladistic Strategies for Haplotype Grouping in Pharmacogenetics
Jared K. Lunceford* and Nancy Liu, Merck Research Laboratories

10:00 Floor Discussion

53. CONTRIBUTED PAPERS:
DIAGNOSTIC AND SCREENING TESTS *Buccaneer D*

Sponsor: *ENAR*
Chair: *A. Lawrence Gould, Merck Research Laboratories*

8:30 Accuracy of Biometric Authentication/Identification Systems
Peter B. Imrey*, The Cleveland Clinic Foundation

8:45 Bayesian Adaptation of the Summary ROC Curve Model for Meta-Analysis of Diagnostic Test Performance
Scott W. Miller*, Debajyoti Sinha, Elizabeth Slate, Don Garrow and Joseph Romagnuolo, Medical University of South Carolina

9:00 A Unified Family of Nonparametric ROC Area Estimators in Group Sequential Designs
Liansheng Tang*, Xiao-Hua Zhou and Scott S. Emerson, University of Washington

9:15 Sequential Evaluation of a Medical Diagnostic Test with Binary Outcomes
Yu Shu*, The George Washington University, Aiyi Liu, National Institute of Child Health and Human Development-Department of Health and Human Services, Zhaohai Li, The George Washington University, National Cancer Institute, Department of Health and Human Services

9:30 ROC Analysis with Non-Binary Reference Standard
Shang-Ying Shiu* and Constantine Gatsonis, Brown University

9:45 Recent Developments in the Dorfman-Berbaum-Metz (DBM) Procedure for Multireader ROC Study Analysis
Stephen L. Hillis*, Iowa City V.A. Medical Center, Kevin S. Berbaum, University of Iowa

10:00 Floor Discussion

54. CONTRIBUTED PAPERS:
METHODS FOR MULTIPLE ENDPOINTS *Esplanade I*

Sponsor: *ENAR*
Chair: *Ronald Gangnon, University of Wisconsin-Madison*

8:30 New Confidence Bounds for QT Studies
Dennis D. Boos*, North Carolina State University, David Hoffman, Robert Kringle and Ji Zhang, Sanofi-Aventis, New Jersey

8:45 Joint Models for a Primary Endpoint and Multivariate Longitudinal Data
Erning Li*, Texas A&M University, Nae-Yuh Wang, The Johns Hopkins University School of Medicine, Naisyin Wang, Texas A&M University

9:00 **Identification of Responders in an Interstitial Cystitis Clinical Trial**
Benjamin E. Leiby*, Mary D. Sammel, Thomas R. Ten Have and Kevin G. Lynch, University of Pennsylvania School of Medicine

9:15 Deletion Diagnostics for Alternating Logistic Regressions
John S. Preisser*, Jamie Perin and Bahjat F. Qaqish, University of North Carolina

9:30 Multivariate Gaussian Power Confidence Intervals Due to Estimating Covariance in One or Two Groups
Sola Park* and Keith E. Muller, University of North Carolina at Chapel Hill

9:45 A Global Test to Detect Regulated Genes in Oligonucleotide Gene Chip
Dung-Tsa Chen*, University of Alabama at Birmingham, James Chen, NCTR, FDA, ChenAn Tsai, Academia Sinica, Seng-jaw Soong, University of Alabama at Birmingham

10:00 Regression Analysis for Modeling 16O/18O Stable-Isotope Distributions for Mass-Spectrometry Analysis
Jeanette E. Eckel-Passow*, Ann L. Oberg, Christopher J. Mason, Douglas W. Mahoney, Robert H. Bergen, Janet E. Olson and Terry M. Therneau, Mayo Clinic

55. CONTRIBUTED PAPERS:
CASE-CONTROL STUDIES *Esplanade3*

Sponsor: *ENAR*
Chair: *David L. Buckeridge, McGill University*

8:30 MLE Method for Case-Control Studies with Longitudinal Covariates
Honghong Zhou*, University of Michigan, Xihong Lin, Harvard University, Bin Nan, University of Michigan

8:45 Comparisons of Sequential Testing Approaches for Detection of Association Between Disease and Candidate Genes: A Simulation Study
Andres Azuero*, University of Alabama at Birmingham

9:00 Weighted Estimating Equations for Case-Control Study within Cohort with Correlated Failure Times
Sangwook Kang* and Jianwen Cai, University of North Carolina at Chapel Hill

9:15 A Minimum Distance Approach to Logistic Regression via the Case-Control Formulation
Howard D. Bondell*, North Carolina State University

9:30 Saddlepoint Approximations in Matched Case-Control Study
Malay Ghosh, Bhramar Mukherjee and Upasana Santra*, University of Florida

9:45 Case-Control Follow-Up Studies: A New Approach to Sampling from a Cohort
Wenguang Sun* and Marshall M. Joffe, University of Pennsylvania

10:00 Floor Discussion

Tuesday, March 28
10:15 a.m. – 10:30 a.m.

Refreshment Break *Atrium*

Tuesday, March 28
10:30 a.m. – 12:15 p.m.

PRESIDENTIAL INVITED ADDRESS *Regency Ballroom*

Sponsor: ENAR
Organizer/Chair: Jane Pendergast, University of Iowa

10:30 Introduction: Jane Pendergast, University of Iowa

10:35 Distinguished Student Paper Awards: Marie Davidian, North Carolina State University

10:55 Statistical Science – Knowledge from Information
Scott L. Zeger, Frank Hurley and Catharine Dorrier Professor in Biostatistics and Chair of the Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health

Tuesday, March 28
1:45 p.m. – 3:30 p.m.

56. CENSORED DATA IN THE ENVIRONMENTAL, AGRICULTURAL AND MEDICAL SCIENCES *Regency 3*

Sponsor: *ASA Section on Statistics and the Environment*
Organizer: *Mary C. Christman, University of Florida*
Chair: *Mary C. Christman, University of Florida*

1:45 Statistical Methods for Censored (Nondetect) Environmental Data
Dennis R. Helsel*, US Geological Survey

2:15 Analysis of Designed Experiments in the Presence of Censored Data
Linda J. Young*, Mary C. Christman and Ramon C. Littell, University of Florida

2:45 Survival Analysis in Two-Stage Randomization Designs
Abdus S. Wahed*, University of Pittsburgh

3:15 Floor Discussion

Scientific Program: Oral Session Summary

57. STATISTICAL ISSUES IN META-ANALYSIS OF GENOMIC AND TRANSCRIPTIONAL DATA WITH A FOCUS ON ARRAY CGH

Regency 6

Sponsor: *Biometrics*
Organizers: *Jane Fridlyand, Department of Epidemiology and Biostatistics, Comprehensive Cancer Center, University of California-San Francisco and Adam Olshen, Memorial Sloan-Kettering Cancer Center*
Chair: *E.S. Venkatraman, Memorial Sloan-Kettering Cancer Center*

1:45 Methods for the Joint Analysis of Array CGH and Gene Expression Data
Adam B. Olshen*, Memorial Sloan-Kettering Cancer Center, E. S. Venkatraman, Memorial Sloan-Kettering Cancer Center"

2:10 Combining Copy Number and Gene Expression Data for the Analysis of Cancer Data
Jane Fridlyand*, Department of Epidemiology and Biostatistics, Comprehensive Cancer Center, University of California-San Francisco, Ritu Roydasgupta, Sandy DeVries, Koei Chin and Fred Waldman University of California-San Francisco, Joe Gray, LBNL, Donna Albertson, University of California-San Francisco

2:35 Detection of the DNA Copy Number Changes Using High Density Oligonucleotide Arrays
Jing Huang*, Affymetrix Inc., Wen Wei, Roche Molecular Systems, Inc., Joyce Chen, Jane Zhang, Guoying Liu, XiaoJun Di and Rui Mei, Affymetrix Inc., Shumpei Ishikawa, University of Tokyo, Keith W. Jones and Michael H. Shapero, Affymetrix Inc.

3:00 Visualizing and Analyzing High Density SNP Data with SNPscan
Ingo Ruczinski* and Rob Scharpf, Johns Hopkins University, Jason Ting and Jonathan Pevsner, Kennedy Krieger Institute

3:25 Floor Discussion

58. HEALTH SURVEY DATA

Regency 7

Sponsor: *ASA Section on Survey Research Methods*
Organizer: *Jai Won Choi, Centers for Disease Control and Prevention*
Chair: *Partha Lahiri, University of Maryland*

1:45 Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health
Nathaniel Schenker*, National Center for Health Statistics

2:15 Using National Surveys to Compute the Number of Deaths Attributable to a Risk Factor
Barry I. Graubard*, National Cancer Institute, Katherine M. Flegal, National Center for Health Statistics/Centers for Disease Control and Prevention, David F. Williamson, Centers for Disease Control and Prevention, Mitchell H. Gail, National Cancer Institute

2:45 Correlation in Multistage Health Survey Designs
Jai W. Choi*, National Center for Health Statistics, Balgobin Nandram, Worcester Polytechnic Institute

3:15 Floor Discussion

59. ILS✧: INTRODUCTION TO BAYESIAN ANALYSIS AND SOFTWARE

Regency 2

Sponsor: *ASA Sections on Teaching Statistics in the Health Sciences/Statistical Education*
Organizer: *Montserrat Fuentes, North Carolina State University*
Chair: *Montserrat Fuentes, North Carolina State University*

1:45 Introduction to Bayesian Analysis and Software
Bradley P. Carlin*, University of Minnesota

3:15 Floor Discussion

60. IMS: RECENT DEVELOPMENTS IN FALSE DISCOVERY RATE

Regency 5

Sponsor: *IMS*
Organizer: *Jonathan Taylor, Stanford University*
Chair: *Jason Fine, University of Wisconsin.*

1:45 Hierarchical FDR Controlling Procedures
Daniel Yekutieli*, Tel Aviv University

2:15 Tail Strength of a Dataset
Jonathan Taylor* and Robert Tibshirani, Stanford University

2:45 Sensitivity and Specificity of FDR Methods in Neuroimaging
Thomas E. Nichols* and Wei Xie, University of Michigan

3:15 Floor Discussion

✧ ILS stands for Introductory Lecture Session.

Scientific Program: Oral Session Summary

61. CONTRIBUTED PAPERS: INTERVAL-CENSORED TIME-TO-EVENT DATA

Buccaneer A

Sponsor: *ENAR*
Chair: *Mireya Diaz, Case Western Reserve University*

1:45 Maximum Likelihood Analysis of Repeated Left- and Interval-Censored Bioassay Data
Jonathan S. Hartzel*, Merck & Co., Inc.

2:00 A Conditional Approach for Regression Analysis of Case 2 Interval-Censored Failure Time Data
Lianming Wang*, Jianguo Sun and Xingwei Tong, University of Missouri-Columbia

2:15 Survival Curve Estimation for Informatively Coarsened Discrete Event-Time Data
Michelle D. Shardell*, University of Maryland School of Medicine, Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health, Sam A. Bozzette, San Diego Veterans Affairs Medical Center

2:30 A Nonparametric Test for Interval-Censored Failure Time Data with Unequal Censoring
Chao Zhu* and Jianguo Sun, University of Missouri-Columbia

2:45 **Sensitivity of Kaplan-Meier Estimate to Nonignorable Censoring**
Tao Liu* and Daniel F. Heitjan, University of Pennsylvania

3:00 'Smooth' Inference for Survival Functions with Arbitrarily Censored Data
Kirsten Doehler* and Marie Davidian, North Carolina State University

3:15 The Impact of Censoring Patterns on the Analysis of Interval-Censored Data
Guozhi Gao*, Xiang Zhang, Steven Snapinn and Qi Jiang, Amgen Inc.

62. CONTRIBUTED PAPERS: MODELING METHODS IN EPIDEMIOLOGY

Buccaneer C

Sponsor: *ENAR*
Chair: *Michael P. Fay, National Institute of Allergy and Infectious Diseases*

1:45 Statistical Modeling and Its Evaluation of Reference Values for Pulmonary Function Test: A Multivariate Approach
JungBok Lee*, Chol Shin and Jae Won Lee, Korea University

2:00 Cancer Risk Assessment of Environmental Agents by Stochastic Models of Carcinogenesis
Wai-Yuan Tan* and Wenyen Zhao, University of Memphis, Chao W. Chen, US EPA, Li-jun Zhang, University of Memphis

2:15 Flexible Bayesian Multistate Models for Multivariate Longitudinal Data
Bo Cai* and David B. Dunson, NIEHS, Joseph B. Stanford, University of Utah

2:30 Comparing Smoothing Techniques for Modeling Exposure-Response Curves in Cox Models
Usha S. Govindarajulu* and Donna Spiegelman, Harvard School of Public Health, Sally W. Thurston, University of Rochester Medical Center, Ellen A. Eisen, Harvard School of Public Health

2:45 Unifying Regression Approaches for Estimating Chronic Effects of Air Pollution on Human Health
Sorina E. Eftim* and Francesca Dominici, Johns Hopkins Bloomberg School of Public Health

3:00 Floor Discussion

63. CONTRIBUTED PAPERS: CLINICAL TRIALS

Buccaneer D

Sponsor: *ENAR*
Chair: *Susan Halabi, Duke University*

1:45 Power Approximation for the van Elteren Test Based on Location-Scale Family of Distributions
Yan Zhao* and Yongming Qu, Eli Lilly and Company, Dewi Rahardja, University of Indianapolis

2:00 Working with the Data Safety Monitoring Board for a Clinical Trial: A Question of Power
Felicity B. Enders*, Jeffrey A. Schmoll and Tanya L. Hoskin, Mayo Clinic

2:15 Bayesian Design for a Second Clinical Trial
Elizabeth A. Johnson*, Scott L. Zeger and Jay Herson, Bloomberg School of Public Health-Johns Hopkins University

2:30 Longitudinal Nested Compliance Class Model in the Presence of Time-Varying Noncompliance
Julia Y. Lin* and Thomas R. Ten Have, University of Pennsylvania School of Medicine, Michael R. Elliott, University of Michigan School of Public Health

2:45 Composite Design for Dose-Finding under Bivariate Probit Model
Yuehui Wu*, Vladimir Dragalin and Valerii V. Fedorov, GlaxoSmithKline

3:00 Weighted Log Rank Subtraction
John L. Bryant*, University of Pittsburgh

3:15 Floor Discussion

64. CONTRIBUTED PAPERS:
MISSING DATA METHODS *Esplanade 1*

Sponsor: *ENAR*
Chair: *Recai M. Yucel, University of Massachusetts-Amherst*

1:45 Multiple Imputation in Survival Analysis with Incomplete Covariates
Jia Li* and Stewart Anderson, University of Pittsburgh

2:00 Shared Parameter Models with a Flexible Random Effects Distribution
Roula Tsonaka*, Geert Verbeke and Emmanuel Lesaffre, Catholic University of Leuven

2:15 An Imputation Strategy for Binary Data
Hakan Demirtas* and Don Hedeker, University of Illinois at Chicago

2:30 Sieve Maximum Likelihood Estimation for Missing Covariates in Regression Models
Qingxia Chen*, Vanderbilt University, Donglin Zeng and Joseph G. Ibrahim, University of North Carolina at Chapel Hill

2:45 A CAR-BART Model to Merge Two Datasets
Song Zhang*, Peter Muller and Tina Shih, University of Texas M.D. Anderson Cancer Center

3:00 A Pseudolikelihood Method for Semiparametric Regression Data with Nonignorable Non-Response
Gong Tang*, University of Pittsburgh

3:15 Floor Discussion

65. CONTRIBUTED PAPERS:
QUANTITATIVE-TRAIT LINKAGE ANALYSIS
Buccaneer B

Sponsor: *ENAR*
Chair: *Mariza de Andrade, Mayo Clinic, Division of Biostatistics*

1:45 Position-Dependent Correlations in eQTL Mapping
Kwang-Youn A. Kim*, Todd E. Scheetz, Ruth Swiderski, Alisdair R. Philp, Thomas L. Casavant, Edwin M. Stone, Val C. Sheffield and Jian Huang, University of Iowa

2:00 Efficient Markov Chain Monte Carlo Algorithms for Mapping Genome-Wide Interacting QTL
Nengjun Yi*, University of Alabama at Birmingham

2:15 Variable Selection for Large p Small n Regression Models with Incomplete Data: Mapping QTL with Epistasis
Min Zhang* and Dabao Zhang, Purdue University, Martin T. Wells, Cornell University

2:30 Poor Performance of Bootstrap Confidence Intervals for the Location of Quantitative Trait Loci
Ani W. Manichaikul*, Karl W. Broman, Johns Hopkins University

2:45 Nonparametric Functional Interval Mapping of Quantitative Trait Loci
Jie Yang* and George Casella, University of Florida

3:00 Genomewide Functional Mapping for Genetic Control of Programmed Cell Death: A Semiparametric Model
Yuehua Cui*, Michigan State University, Rongling Wu, University of Florida

3:15 Nonlinear Mixed-Effect Mixture Models for Functional Mapping of Longitudinal Traits
Wei Hou* and Rongling Wu, University of Florida

66. CONTRIBUTED PAPERS:
SPATIAL-TEMPORAL MODELING *Esplanade 3*

Sponsor: *ENAR*
Chair: *Mikyoung Jun, Texas A&M University*

1:45 Multivariate Spatiotemporal Models for Environmental Epidemiological Data
Brian Reich* and Montserrat Fuentes, North Carolina State University, David Holland, U.S. Environmental Protection Agency

2:00 Spatial and Temporal Analysis on Missouri Bladderpod (Lesquerella filiformis)
William B. Leeds*, Elizabeth R. Bobzien, Hyun-Joo Kim and Michael I. Kelrick, Truman State University

2:15 Nonparametric Estimation of Correlation Functions in Longitudinal and Spatial Data, with Application to Colon Carcinogenesis Experiments
Yehua Li*, Naisyin Wang and Raymond J. Carroll, Texas A&M University
2:30 Partitioning Statistical Evidence of Causal Association in Observational Studies: An Illustration Using Spatiotemporal Data
Holly Janes*, Scott L. Zeger and Francesca Dominici, Johns Hopkins Bloomberg School of Public Health
2:45 Floor Discussion

Tuesday, March 28
3:30 p.m. – 3:45 p.m.

Refreshment Break *Atrium*

Tuesday, March 28
3:45 p.m. – 5:30 p.m.

67. RECENT DEVELOPMENTS IN BAYESIAN BIOINFORMATICS *Regency 2*

Sponsor: *ASA Biopharmaceutical Section*
Organizer: *Cavan Reilly, University of Minnesota*
Chair: *Cavan Reilly, University of Minnesota*

3:45 Bayesian Calibration of Mass Spectra
Cavan Reilly*, University of Minnesota
4:15 Regulatory Binding Site Detection from High-Density Sequence and ChIP-chip Array Data using Hidden Markov Models
Mayetri Gupta*, Jonathan Gelfond and Joseph Ibrahim, University of North Carolina at Chapel Hill
4:45 Bayesian Robust Inference for Differential Gene Expression
Raphael Gottardo*, University of British Columbia, Adrian Raftery, Ka Yee Yeung and Roger Bumgarner, University of Washington
5:15 Floor Discussion

68. TO POOL OR NOT TO POOL: SYSTEMATIC REVIEWS AND POOLED ANALYSES *Regency 3*

Sponsor: *ENAR*
Organizer: *Sumithra J. Mandrekar, Mayo Clinic*
Chair: *Sumithra J. Mandrekar, Mayo Clinic*

3:45 Case Studies of the Use of Meta-Analysis in Pharmacoepidemiology
Jesse A. Berlin*, Johnson and Johnson Pharmaceutical Research and Development
4:15 Recent Advances in Surrogate Endpoint Evaluation
Tomasz Burzykowski*, Hasselt University-Belgium
4:45 A Simple Meta-Analytic Approach for Binary Surrogate Endpoints
Stuart G. Baker*, National Cancer Institute
5:15 Discussant: Daniel J. Sargent, Mayo Clinic

69. NEW STATISTICAL METHODS FOR ESTIMATING MEDICAL EXPENDITURES AND COST EFFECTIVENESS FROM OBSERVATIONAL DATA *Regency 6*

Sponsor: *ASA Health Policy Statistics Section*
Organizers: *Paul Rathouz, University of Chicago, Daniel Heitjan, University of Pennsylvania-School of Medicine*
Chair: *Paul Rathouz, University of Chicago*

3:45 Bayesian Cost Effectiveness Analysis
Daniel F. Heitjan*, University of Pennsylvania
4:10 Estimating Medical Expenditures for Smoking-Related Diseases via Smooth Quantile Ratio Estimation
Francesca Dominici*, Johns Hopkins Bloomberg School of Public Health
4:35 Estimating the Cost-Effectiveness of Medical Therapies from Observational Data via Propensity Scores
Nandita Mitra, University of Pennsylvania, Alka Indurkhya*, University of Massachusetts
5:00 Covariate Adjustment in Censored Cost Data
Andrew R, Willan*, SickKids Research Institute and University of Toronto, Danyu Lin, University of North Carolina, Andrea Manca, University of York
5:25 Floor Discussion

Scientific Program: Oral Session Summary

70. STATISTICAL CHALLENGES IN PRE-CLINICAL PHARMACEUTICAL RESEARCH *Regency 7*

Sponsor: *ASA Biopharmaceutical Section*
Organizer: *David Potter, Pfizer Global Research & Development*
Chair: *David Potter, Pfizer Global Research & Development*

3:45 Analysis in Preclinical Pharmaceutical Research: Challenges and Opportunities
J. Alan Menius*, GlaxoSmithKline

4:10 How to Find Drugs with Trees: Applications of Ensemble Methods in QSAR Modeling
Andy Liaw*, Christopher Tong, Ting-chuan Wang and Vladimir Svetnik, Merck Research Laboratories

4:35 Searching for Optimums in High Dimensional Space: An Application of the SELC Algorithm in Drug Discovery
Kjell Johnson*, Pfizer, Inc., Abhyuday Mandal, University of Georgia, C.F. Jeff Wu, Georgia Institute of Technology

5:00 Open Issues in the Classification of Oligonucleotide Microarray Data
Max Kuhn*, Pfizer Global Research and Development

5:25 Floor Discussion

71. IMS: MEDALLION LECTURE *Regency 1*

Sponsor: *IMS*
Organizer: *Michael Kosorok, University of Wisconsin-Madison*
Chair: *Michael Kosorok, University of Wisconsin-Madison*

3:45 Shrinkage Estimation: An Expanding Statistical Theme
Lawrence D. Brown*, University of Pennsylvania

5:00 Floor Discussion

72. CONTRIBUTED PAPERS: SEMIPARAMETRIC AND NONPARAMETRIC METHODS IN LONGITUDINAL AND SURVIVAL ANALYSIS *Buccaneer A*

Sponsor: *ENAR*
Chair: *Lang Wu, University of British Columbia*

3:45 Identifying Latent Clusters of Variability in Longitudinal Data
Michael R. Elliott*, University of Michigan School of Public Health

4:00 **Bayesian Semiparametric Regression for Longitudinal Binary Process Data**
Li Su* and Joseph W. Hogan, Brown University

4:15 Semiparametric Analysis of Longitudinal Data with Informative Observation Times
Jianguo Sun, University of Missouri-Columbia, Do-Hwan Park*, University of Nevada-Reno, Liuquan Sun, Chinese Academy of Sciences, Xingqiu Zhao, McMaster University

4:30 On Semiparametric Regression Models for Nonhomogeneous Birth-Birth Process
Hao Liu*, University of California-Davis

4:45 A Bayesian Model for Sparse Functional Data
Wesley K. Thompson*, University of Pittsburgh, Ori Rosen, University of Texas-El Paso

5:00 Floor Discussion

73. CONTRIBUTED PAPERS: SPATIAL MODELING OF DISEASE *Buccaneer C*

Sponsor: *ENAR*
Chair: *Veera Baladandayuthapani, M.D. Anderson Cancer Center*

3:45 A Spatial Scan Statistic for Ordinal Data
Inkyung Jung* and Martin Kulldorff, Harvard Medical School-Harvard Pilgrim Health Care, Ann C. Klassen, Johns Hopkins Bloomberg School of Public Health

4:00 Cancer Cluster Detection in Semi-Parametric Models with Random Effects: A Score-Based Testing Approach
Matteo Bottai and Marco Geraci*, Arnold School of Public Health-University of Southern California

4:15 Spatial Association on Cancer Incidence in the Community Surrounding the Rocketdyne Facility in Southern California
Sunkyung Yu*, Jennifer B. Dimmer and Hal Morgenstern, University of Michigan

4:30 Geostatistical Hierarchical Model for Temporally Integrated Data Measured with Error
Brian J. Smith* and Jacob J. Oleson, The University of Iowa

4:45 Signal Quality Measurements for cDNA Microarray Data
Tracy L. Bergemann*, University of Minnesota, Lue Ping Zhao, Fred Hutchinson Cancer Research Center

5:00 Approximate Methods in Bayesian Point Process Spatial Models
M. M. Hossain* and Andrew B. Lawson, University of South Carolina

5:15 Effects of Air Pollution, Social Economical Status, and Spatial Clustering Effects on Lung Cancer Mortality Rates of North Carolina Counties of 2000
Kuo-Ping Li* and Chirayath Suchindran, University of North Carolina at Chapel Hill

Scientific Program: Oral Session Summary

74. CONTRIBUTED PAPERS: MULTIPLE TESTING AND FALSE DISCOVERY RATES

Buccaneer B

Sponsor: *ENAR*
Chair: *Yuehua Cui, Michigan State University*

3:45 A Modified Spatial Scan Statistic for Cluster Detection
Ronald E. Gangnon*, University of Wisconsin-Madison

4:00 Screening and Replication Using the Same Data Set: Testing
Amy J. Murphy* and Matthew B. McQueen, Harvard School of Public Health, Benjamin A. Raby, Brigham and Women's Hospital-Harvard Medical School, Kady Schneider, Jessica Su and Juan Celedon, Harvard School of Public Health, Edwin K. Silverman, Brigham and Women's Hospital-Harvard Medical School, Nan M. Laird, Harvard School of Public Health, Scott T. Weiss, Brigham and Women's Hospital, Harvard Medical School, Christoph Lange, Harvard School of Public Health

4:15 An Adaptive Alpha-Spending Algorithm Improves the Power of Statistical Inference in Microarray Data Analysis
Jacob P. Brand*, Pennington Biomedical Research Center, Lan Chen, Xiangqin Cui, Alfred A. Bartolucci, Grier P. Page, Kyoungmi Kim, Stephen Barnes, Vinodh Srinivasasainagendra, Mark T. Beasley and David B. Allison, University of Birmingham at Alabama

4:30 The Effect of Correlated Gene Expression on Tests of Functional Category Enrichment
William T. Barry*, Andrew B. Nobel and Fred A. Wright, University of North Carolina

4:45 A Note on Using Permutation Based False Discovery Rate Estimate to Compare Different Analysis Methods for Microarray Data
Yang Xie*, Wei Pan and Arkady B. Khodursky, University of Minnesota

5:00 False Discovery Rate Adjustment for Tree Models
Carol J. Etzel* and Sumesh Kachroo, University of Texas-M.D. Anderson Cancer Center

5:15 Finite Sample Properties of Estimators of the False Discovery Rate
Naim U. Rashid*, Duke University and Anindya Roy, University of Maryland

75. CONTRIBUTED PAPERS: COMPETING RISKS AND CURE RATES

Buccaneer D

Sponsor: *ENAR*
Chair: *Abdus Wahed, University of Pittsburgh*

3:45 Impact of Change in Level of Risk Factor(s) and Proportion of Cured/Immune Individuals on the Population Attributable Risk: A Simulation Based Study
Jayawant N. Mandrekar*, Mayo Clinic, Melvin L. Moeschberger, The Ohio State University

4:00 Bayesian Additive-Multiplicative Cure Rate Model
Guosheng Yin*, M. D. Anderson Cancer Center–The University of Texas, Luis E Nieto-Barajas, Departmento de Estadistica ITAM

4:15 Effects of Competing Causes of Death in MA.17, a Placebo-Controlled Trial of Letrozole as Extended Adjuvant Therapy for Breast Cancer Patients
Daniel Q. Meng*, Judith-Anne W. Chapman, Lois E. Shepherd and Wendy Parulekar, NCIC-CTG, Queen's University-Canada, James N. Ingle, Mayo Clinic, Paul E. Goss, Massachusetts General Hospital Cancer Center

4:30 Flexible Cure Rate Modeling Under Latent Activation Schemes
Freda W. Cooner*, Sudipto Banerjee and Bradley P. Carlin, University of Minnesota, Debajyoti Sinha, Medical University of South Carolina

4:45 Parametric Regression on Cumulative Incidence Function
Jong-Hyeon Jeong*, University of Pittsburgh, Jason Fine, University of Wisconsin-Madison

5:00 Modeling Bivariate Competing Risk Events via Markov Chains
Mireya Diaz*, Case Western Reserve University

5:15 Floor Discussion

Scientific Program: Oral Session Summary

76. CONTRIBUTED PAPERS:
GENE EXPRESSION ANALYSIS *Esplanade I*

Sponsor: *ENAR*
Chair: *Russ Wolfinger, SAS Institute Inc.*

3:45 Flexible Temporal Expression Profile Modelling Using the Gaussian Process
Ming Yuan*, Georgia Institute of Technology

4:00 **Normalization of Microarrays in Transcription Inhibition**
Yan Zheng*, Cavan Reilly, University of Minnesota

4:15 Comparison of Various Statistical Methods for Identifying Differential Gene Expression in Replicated Microarray Data
Seo Young Kim, Chonnam National University, Jae Won Lee* and In Suk Sohn, Korea University

4:30 Applications of Reliability Coefficients in cDNA Microarray Data Analysis
Wenqing He*, University of Western Ontario, Shelley B. Bull, Samuel Lunenfeld Research Institute and University of Toronto

4:45 Improved Parameter Estimation for RMA in Background Correction of Gene Expression Microarrays
Monnie McGee* and Zhongxue Chen, Southern Methodist University

5:00 Identifying Functional Gene Categories in Microarray Experiments with Nonparametric Methods
Hua Liu*, Christopher P. Saunders, Constance L. Wood and Arnold J. Stromberg, University of Kentucky

5:15 Floor Discussion

Tuesday, March 28
5:30 p.m. – 6:30 p.m.

ENAR Business Meeting *Buccaneer B*
(Open to all ENAR Members)

Wednesday, March 29
8:30 a.m. – 10:15 a.m.

77. RECENT ADVANCES IN STATISTICAL METHODS FOR GENETIC EPIDEMIOLOGY *Regency 2*

Sponsors: *ASA Section on Risk Analysis/ASA Section on Epidemiology*
Organizer: *Bhramar Mukherjee, University of Florida*
Chair: *Bhramar Mukherjee, University of Florida*

8:30 Robust Estimation of Haplotype/Environment Interactions
Andrew S. Allen*, Duke University, Glen A. Satten, Centers for Disease Control and Prevention

9:00 Association Testing with Related Individuals
Mary Sara McPeek*, University of Chicago

9:30 Two-Phase Designs in Studies of Gene-Environment Interaction
Nilanjan Chatterjee*, National Cancer Institute

10:00 Floor Discussion

78. ADAPTIVE BAYESIAN MODELING OF FUNCTIONAL BIOMEDICAL DATA *Regency 3*

Sponsor: *Biometrics*
Organizer: *Jeffrey S. Morris, The University of Texas-M.D. Anderson Cancer Center*
Chair: *Nebiyou Bekele, The University of Texas-M.D. Anderson Cancer Center*

8:30 Adaptive Bayesian Smoothing of Functional Predictors in Linear Mixed Models Using Wavelet Shrinkage
Elizabeth J. Malloy* and Brent A. Coull, Harvard School of Public Health, Jeffrey S. Morris, M.D. Anderson Cancer Center, Sara D. Dubowsky and Helen H. Suh, Harvard School of Public Health

8:55 Bayesian Adaptive Regression Splines for Hierarchical Data
Jamie L. Bigelow* , University of North Carolina at Chapel Hill and National Institute of Environmental Health Sciences, David B. Dunson, National Institute of Environmental Health Sciences

9:20 Bayesian Hierarchical Spatially Correlated Functional Data Analysis with Application to Colon Carcinogenesis
Veera Baladandayuthapani*, The University of Texas M.D. Anderson Cancer Center, Bani K. Mallick, Mee Young Hong and Raymond J. Carroll, Texas A&M University

9:45 Wavelet-Based Functional Mixed Models
Jeffrey S. Morris*, University of Texas M.D. Anderson Cancer Center, Raymond J. Carroll, Texas A&M University

10:10 Floor Discussion

Scientific Program: Oral Session Summary

79. CURRENT TRENDS IN SMALL AREA ESTIMATION *Regency 6*

Sponsor: *ASA Section on Survey Research Methods*
Organizer: *Partha Lahiri, University of Maryland*
Chair: *Paul D. Williams, National Center for Health Statistics*

8:30 Bayesian Methodology which Accounts for Uncertainty About the Commonality of a Set of Small Area Estimates
Guofen Yan* , University of Virginia, J. Sedransk, Case Western Reserve University
9:00 Using Small Area Estimation Method to Combine Two Health Surveys
Shijie Chen*, RTI International
9:30 A Small Area Estimation Approach in Reducing Survey Costs
Partha Lahiri*, University of Maryland, Paul D. Williams, National Center for Health Statistics
10:00 Floor Discussion

80. INFERENCE IN THE PRESENCE OF NON-IDENTIFIABILITY: APPLICATIONS TO THE ANALYSIS OF COARSE DATA *Regency 7*

Sponsors: *IMS/ENAR*
Organizer: *Daniel O. Scharfstein, Johns Hopkins University*
Chair: *Daniel Scharfstein, Johns Hopkins University*

8:30 Bayesian Inference in the Presence of Non-Identifiability
Paul Gustafson*, University of British Columbia
9:00 Drawing Inference from Regions of Estimates: Ignoring Bounds or Bounding Ignorance?
Stijn Vansteelandt* and Els Goetghebeur, Ghent University-Belgium, Mike Kenward, London School of Hygiene and Tropical Medicine-U.K., Geert Molenberghs, Hasselt University-Belgium
9:30 A Distributional Approach for Sensitivity Analysis in Observational Studies
Zhiqiang Tan*, Johns Hopkins University
10:00 Floor Discussion

81. IMS: BAYESIAN MODEL SELECTION *Regency 5*

Sponsor: *IMS*
Organizer: *Adrian Raftery, University of Washington*
Chair: *Jason Fine, University of Wisconsin*

8:30 MCMC with Mixtures of Singular Distributions: Application to Bayesian Model Selection
Adrian E. Raftery*, University of Washington, Raphael Gottardo, University of British Columbia
9:00 Using Bayesian Model Averaging to Assess the Effect of Social Interactions on Recidivism
Sibel Sirakaya*, Departments of Economics and Statistics, and Center for Statistics and the Social Sciences, University of Washington
9:30 Bayesian Model Selection for Discriminant Analysis
Russell J. Steele* and Michelle E. Ross, McGill University
10:10 Floor Discussion

82. CONTRIBUTED PAPERS: RE-SAMPLING AND ROBUST METHODS AND APPLICATIONS *Buccaneer B*

Sponsor: *ENAR*
Chair: *Philip Reiss, Columbia University*

8:30 General Outlier Detection for a Homogeneous Poisson Process with Sum-Quota Accrual Scheme
Jonathan T. Quiton*, Edsel A. Peña and James D. Lynch, University of South Carolina
8:45 Determining Optimal Experimental Designs For Nonlinear Models Using Likelihood Ratio Based Inference
Sharon D. Yeatts* and Chris Gennings, Virginia Commonwealth University
9:00 Components of the Bootstrap-Variance of the Area under the ROC Curve
Andriy I. Bandos*, Howard E. Rockette and David Gur, University of Pittsburgh
9:15 **Marginal Analysis of Correlated Failure Time Data with Informative Cluster Sizes**
Xiuyu Cong*, Rice University, Guosheng Yin and Yu Shen, University of Texas-M. D. Anderson Cancer Center
9:30 **Permutation-Based Test for Identifying Longitudinal Gene Expressions Associated with the Time to an Event**
Natasa Rajicic*, Harvard School of Public Health, Dianne M. Finkelstein and David A. Schoenfeld, MGH Biostatistics and Harvard School of Public Health
9:45 Estimating the Location from a Skewed Sample: To Transform or Not to Transform
Abutaher M. Minhajuddin* and Xian-Jin Xie, University of Texas-Southwestern Medical Center
10:00 A Novel Statistical Approach Identifying and Limiting the Effect of Influential Observations
Tamekia L. Jones* and David T. Redden, University of Alabama at Birmingham

Scientific Program: Oral Session Summary

83. CONTRIBUTED PAPERS:
PHARMACOKINETICS, PHARMACODYNAMICS, AND TOXICOLOGY *Buccaneer A*

Sponsor: *ENAR*
Chair: *Andreas Krause, Pharsight Corp.*

8:30 Threshold Dose-Response Model with Random Effects for Teratological Data
Daniel L. Hunt* and Shesh N. Rai, St. Jude Children's Research Hospital

8:45 Interval Estimation of Effective Doses in Tobit Regression Model
Nan Lin*, Washington University in St. Louis, Douglas Simpson, University if Illinois at Urbana-Champaign, Ringo M. Ho, Nanyang Technology University

9:00 **Stochastic Models for Compliance Analysis Using Inter-Dosing Times**
Junfeng Sun*, University of Nebraska Medical Center, Haikady N. Nagaraja, The Ohio State University

9:15 Nonparametric Bayes Testing of Changes in a Response Distribution with an Ordinal Predictor
Michael L. Pennell*, University of North Carolina at Chapel Hill, NIEHS, David B. Dunson, NIEHS

9:30 A Physiologically Based Pharmacokinetic Model for Gavage and IV Administration of Methyleugenol in F344/N Rats and B6C3F1 Mice
Petra K. LeBeau*, Rho, Inc., Shree Y. Whitaker, Christopher J. Portier, NIEHS

9:45 A Regression Based Approach for Developing a Limited Sample Model for Pharmacokinetic Data
Alfred F. Furth*, Sumithra J. Mandrekar, Andrea Rau, Joel M. Reid, Angelina Tan, Sara J. Felten, Charles Erlichman, Matthew M. Ames and Alex A. Adjei, Mayo Clinic

10:00 Relationship Assessment between PK and PD
Tao Liu, University of Pennsylvania, Longlong Gao*, GlaxoSmithKline Company

84. CONTRIBUTED PAPERS:
SURVIVAL ANALYSIS II *Buccaneer C*

Sponsor: *ENAR*
Chair: *Hans C. Van Houwelingen, LUMC*

8:30 Nonparametric Estimation of the Bivariate Survivor Function Under Right Truncation with Application to Panic Disorder
Xiaodong Luo* and Wei-Yann Tsai, Columbia University

8:45 A Penalized Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data
Wen Ye*, University of Michigan, Xihong Lin, Harvard University, Jeremy M.G. Taylor, University of Michigan

9:00 New Methodology: Challenging the Logrank and Wilcoxon Tests for Nonparametric Survival Comparison
Gabriel P. Suciu*, Nova Southeastern University

9:15 **A Sample Size Formula for Recurrent Events Data Using Robust Log-Rank Statistics**
Rui Song*, University of Wisconsin-Madison, Michael R. Kosorok, University of Wisconsin-Madison

9:30 Regression Models for the Mean of the Quality-of-Life-Adjusted Restricted Survival Time Using Pseudo-Observations
Adin-Cristian Andrei* and Susan Murray, University of Michigan

9:45 An Evaluation of Pseudo Observations in Multi State Models
Pinaki Biswas* and Jack D. Kalbfleisch, University of Michigan, Ann Arbor

85. CONTRIBUTED PAPERS:
TOPICS IN STATISTICS: SEQUENTIAL METHODS, GOODNESS-OF-FIT TESTS, AND MULTIVARIATE ANALYSIS *Buccaneer D*

Sponsor: *ENAR*
Chair: *Guosheng Yin, Department of Biostatistics, M. D. Anderson Cancer*

8:30 Comparison of Sequential Experiments for Estimating the Number of Classes in a Population
Tapan K. Nayak* and Subrata Kundu, George Washington University

8:45 An Effective Maximum Likelihood Estimation Method for a Finite Mixture Model in High Dimensional Biology
Qinfang Xiang* and Gary L. Gadbury, University of Missouri-Rolla

9:00 New Tests of Uniformity and Normality
David B. Kim*, Manhattan College

9:15 The Asymptotic Distribution of Modified Shapiro-Wilk Statistics for Testing Multivariate Normality
Christopher P. Saunders* and Constance L. Wood, University of Kentucky

9:30 Estimating Equations for Canonical Correlations
Hye-Seung Lee*, Myunghee Cho Paik and Joseph H. Lee, Columbia University

9:45 ArrayBlast: Data-Mining Tool for Gene Expression Signatures
Yajun Yi, Chun Li* and Alfred L. George, Vanderbilt University

10:00 Randomized Discontinuation Trials: Design and Efficiency
Tao Liu, University of Pennsylvania, Valeri V. Fedorov*, GlaxoSmithKline Co.

Wednesday, March 29
10:15 a.m. – 10:30 a.m.

Refreshment Break *Atrium*

Wednesday, March 29
10:30 a.m. – 12:15 p.m.

86. STATISTICAL ISSUES IN THE DESIGN, EVALUATION, AND MONITORING OF CLINICAL TRIALS WITH LONGITUDINAL AND SURVIVAL ENDPOINTS

Regency 2

Sponsors: *Biometrics and ASA Sections on Teaching Statistics in the Health Sciences, Statistical Education and the Biopharmaceutical Section*
Organizer: *Dan Gillen, University of California*
Chair: *Dan Gillen, University of California*

10:30 Bayesian Evaluation of Longitudinal/Survival Trials
Donald A. Berry*, University of Texas M. D. Anderson Cancer Center

10:55 Sample Size Re-Estimation in Survival Studies
Thomas D. Cook*, University of Wisconsin-Madison

11:20 Stochastic Curtailment Estimation in Survival Studies
Dan L. Gillen*, University of California-Irvine

11:45 Evaluation of Stopping Rules and Secondary Endpoints for Longitudinal Studies
John M. Kittelson*, University of Colorado Health Sciences Center

12:10 Floor Discussion

87. LATENT VARIABLES AND MULTIVARIATE ANALYSIS

Regency 3

Sponsor: *ASA Health Policy Statistics Section*
Organizer: *A. James O'Malley, Harvard Medical School*
Chair: *Recai M. Yucel, Ph.D., University of Massachusetts*

10:30 Variable Selection in Nonparametric Random Effects Models
David B. Dunson* and Bo Cai, NIEHS

10:55 Longitudinal Profiling of Health Care Units Based on Continuous and Discrete Patient Outcomes
Michael J. Daniels*, University of Florida, Sharon-Lise Normand, Harvard Medical School

11:20 Latent Variable Models for Multiple Non-Commensurate Outcomes
Armando Teixeira-Pinto*, Harvard Graduate School of Arts and Science and Faculty of Medicine, University of Porto, Sharon-Lise T. Normand, Harvard Medical School and Harvard School of Public Health

11:45 Bayesian Approaches to Hierarchical Factor Analysis
A. James O'Malley* and Alan M. Zaslavsky, Harvard Medical School

12:10 Floor Discussion

88. ROC ANALYSIS IN BIOMEDICAL INFORMATICS

Regency 6

Sponsor: *ASA Health Policy Statistics Section*
Organizer: *Kelly H. Zou, Harvard Medical School*
Chair: *Kelly H. Zou, Harvard Medical School*

10:30 Optimal Estimation of ROC Curves of Continuous-Scale Tests
Xiao-Hua A. Zhou*, University of Washington and VA Puget Sound Health Care System, Huazhen Lin, University of Washington and Sichuan University

11:00 Non-Parametric Sequential Testing of the Area under the ROC Curves
Aiyi Liu*, Chengqing Wu and Enrique F. Schisterman, National Institute of Child Health and Human Development

11:30 Assessing Rater Performance in Image Segmentation
Simon K. Warfield*, Kelly H. Zou and William M. Wells, Harvard Medical School

12:00 Discussant: Kelly H. Zou, Harvard Medical School

Scientific Program: Oral Session Summary

89. STATISTICAL CONTRIBUTIONS TO THE FRONTIERS OF HIV/AIDS RESEARCH *Regency 7*

Sponsor: *ASA Biopharmaceutical Section*
Organizer: *Craig B. Borkowf, Centers for Disease Control and Prevention*
Chair: *Dr. Lillian S. Lin, Centers for Disease Control and Prevention*

10:30 A General Gamma-Based History of Survival after AIDS: 1984-2004
Alvaro Muñoz, Christopher Cox*, Haitao Chu and Michael Schneider, Johns Hopkins Bloomberg School of Public Health

10:55 Challenges in Designing a Study to Evaluate Whether Use of Antriretroviral Therapy to Prevent Mother to Child HIV Transmission Impacts Future Maternal Treatment Options
Michael D. Hughes*, Harvard School of Public Health

11:20 Modeling the Population Level Effects of an HIV-1 Vaccine in an Era of HAART
Wasima Rida*, Statistics Collaborative, Inc., Sonja Sandberg, Framingham State College

11:45 Investigating Associations between Functional Patterns of Immune Response to HIV and Disease Progression
Martha Nason*, Biostatistics Research Branch, NIAID, NIH

12:10 Floor Discussion

90. CONTRIBUTED PAPERS: MISSING DATA IN LONGITUDINAL DATA ANALYSIS *Buccaneer B*

Sponsor: *ENAR*
Chair: *Yulei He, Harvard University*

10:30 A Latent-Class Mixture Model for Incomplete Longitudinal Data
Caroline Beunckens* and Geert Molenberghs, Hasselt University, Geert Verbeke, Biostatistical Center-K. U. Leuven

10:45 A Censored Multinomial Model for Binary, Longitudinal Survey Data with Missing Values
Steven J. Mongin and Timothy R. Church*, University of Minnesota School of Public Health

11:00 Missing Phenotype Data Imputation for Longitudinal Pedigree Data Analysis
Mariza de Andrade and Brooke Fridley*, Mayo Clinic

11:15 **Sensitivity Analysis and Informative Priors for Longitudinal Binary Data with Outcome-Related Dropout**
Joo Yeon Lee* and Joseph W. Hogan, Brown University

11:30 An Extension of Latent Variable Model for Informative Intermittent Missing Data
Li Qin*, Lisa A. Weissfeld, Melissa Kalarchian and Marsha Marcus, University of Pittsburgh

11:45 Imputation Approach for Responders Analysis in Longitudinal Studies with Random Missing Data
Liqiu Jiang*, North Carolina State University, Kaifeng Lu, Merck & Co., Inc., Anastasios A. Tsiatis North Carolina State University

12:00 A Selection Model for Functional Mapping of Longitudinal Traits with Non-Ignorable Missing Data
Hongying Li* and Rongling Wu, University of Florida

91. CONTRIBUTED PAPERS: CATEGORICAL DATA ANALYSIS AND EXPERIMENTAL DESIGN *Buccaneer A*

Sponsor: *ENAR*
Chair: *Jong-Min Kim, University of Minnesota-Morris*

10:30 Sampling Weighted Relative Risk Regression
Rickey E. Carter* and Stuart R. Lipsitz, Medical University of South Carolina

10:45 Misspecification Tests for Discrete Data Models
Marinela Capanu*, Memorial Sloan-Kettering Cancer Center, Brett Presnell, University of Florida

11:00 **A Test for Non-Inferiority with a Mixed Multiplicative/Additive Null Hypothesis**
Xiaodan Wei* and Richard J. Chappell, University of Wisconsin-Madison

11:15 A Penalized Latent Class Model for Ordinal Responses
Stacia M. DeSantis*, E. Andres Houseman, Brent A. Coull and Rebecca A. Betensky, Harvard School of Public Health

11:30 Asymptotically Optimal Inference with Partially Observable Binary Data: Applications to Plant Disease Assessment
Joshua M. Tebbs*, University of South Carolina, Melinda H. McCann, Oklahoma State University

11:45 Multivariate Logistic Models
Bahjat F. Qaqish*, Anastasia Ivanova and Eugenio Andraca, University of North Carolina

12:00 Bayesian Estimate of Odds Ratios for Small Sample Size 2 x 2 Tables with Incompletely Classified Data
Yan Lin*, Stuart R. Lipsitz, Debajyoti Sinha, Barbara C. Tilley and Rickey Carter, Medical University of South Carolina

Scientific Program: Oral Session Summary

92. CONTRIBUTED PAPERS:
ANALYZING MICROARRAY DATA — *Buccaneer C*

Sponsor: *ENAR*
Chair: *Edoardo Airoldi, School of Computer Science, Carnegie Mellon University*

10:30 Predictive Model Building for Microarray Data Using Generalized Partial Least Squares Model
Baolin Wu*, University of Minnesota

10:45 Overview on Structural Association Testing and Regional Admixture Mapping
David B. Allison*, T.M. Beasley, Jose R. Fernandez, David T. Redden, Hemant K. Tiwari, Jasmin Divers and Robert Kimberly, University of Alabama at Birmingham

11:00 Incorporating Gene Functions into Regression Analysis of DNA-Protein Binding Data and Gene Expression Data to Construct Transcriptional Networks
Peng Wei* and Wei Pan, University of Minnesota

11:15 Haplotype Frequency Estimation from Pooled Genotypes: A Contingency Table Perspective
Yaning Yang, University of Science and Technology of China, Jinfeng Xu* and Zhiliang Ying, Columbia University, Jurg Ott, Laboratory of Statistical Genetics, Rockefeller University

11:30 Eigengene Based Linear Discriminant Model for Gene Expression Data Analysis
Ronglai Shen*, University of Michigan, Zhaoling Meng, Sanofi-Aventis, Debashis Ghosh and Arul M. Chinnaiyan, Comprehensive Cancer Center-University of Michigan

11:45 Incorporating Biological Knowledge into Tumor Classifications with Microarray Data
Feng Tai* and Wei Pan, University of Minnesota

12:00 Floor Discussion

93. CONTRIBUTED PAPERS:
NONPARAMETRIC AND SEMIPARAMETRIC METHODS
Buccaneer D

Sponsor: *ENAR*
Chair: *Chung-Chou H. Chang, University of Pittsburgh*

10:30 Mixed-Effects, Posterior Means and Penalized Least Squares
Yolanda Munoz Maldonado*, UT-HSC School of Public Health at Houston

10:45 **Penalized Functional Principal Components Analysis Using a Kullback-Leibler Criterion**
Robert T. Krafty* and Wensheng Guo, University of Pennsylvania School of Medicine

11:00 Estimating Linear Functionals of Indirectly Observed Input Functions
Eun-Joo Lee*, Illinois College

11:15 A Nonparametric Likelihood Ratio Test to Identify Differentially Expressed Genes from Microarray Data
Sunil Mathur* and Sankar Bokka, University of Mississippi

11:30 Modeling of Hormone Secretion-Generating Mechanisms: A Spline and Pseudo-Likelihood Approach
Anna Liu*, University of Massachusetts, Yuedong Wang, University of California-Santa Barbara

11:45 Rank Estimation of Accelerated Lifetime Models with Dependent Censoring
Limin Peng*, Emory University, Jason P. Fine, University of Wisconsin-Madison"

12:00 Floor Discussion

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

POSTER SESSION

A BAYESIAN MODEL FOR NON-INFERIORITY STUDY DESIGN INVOLVING ETHNIC COMPARISONS

Fanni Natanegara*, Eli Lilly and Company
John W. Seaman, Baylor University

A problem of considerable interest in clinical trials is to compare the efficacy of a treatment between different ethnic groups. Traditional non-inferiority designs compare an active treatment to an established one. We consider a unique non-inferiority design in which the same treatment is compared in different ethnic groups. For a reference group we use a specific ethnic group, C, comprised of a significant portion of patients who have demonstrated efficacy in past placebo-controlled trials. In this novel non-inferiority design, a previously underrepresented ethnic group, U, will be compared to C. The goals of such a study design are two fold: (1) to provide direct efficacy comparison of ethnic groups U vs C and (2) to establish treatment efficacy in U compared to a putative placebo P. We use a Bayesian approach to account for uncertainty in historical data based on ethnic groups that provides flexibility in data interpretation based on the posterior distribution. We present the posterior probability that U is non-inferior to C and U is superior to P with application to clinical trial data evaluating treatment efficacy for erectile dysfunction.

email: natanegara_fanni@lilly.com

MODELING DIABETES INCIDENCE IN THE NATIONAL HEALTH INTERVIEW SURVEY

Theodore J. Thompson*, Centers for Disease Control & Prevention
James P. Boyle, Centers for Disease Control & Prevention

National Health Interview Survey (NHIS) data was used to estimate diabetes incidence in the United States by age, race/ethnicity, sex, and body mass index (BMI). Bayesian multilevel logistic regression models including random effects for year were fit using Markov chain Monte Carlo methods. Cubic splines were used to obtain smooth estimates by age. Model fit was assessed via graphical methods and posterior predictive p-values. NHIS data from 1997 to 2004 including 1514 incident cases among 228,628 observations were analyzed. Diabetes incidence is strongly associated with BMI. At 60 years of age the relative risk for diabetes among obese vs. normal BMI individuals is 4.9 (95% posterior interval 3.6-6.7) for White females, 4.6 (3.3-6.2) for White males, 4.8 (3.5-6.5) for Black females, 4.5 (3.3-6.1) for Black males, 4.9 (3.6-6.6) for Hispanic females, and 4.5 (3.3-6.1) for Hispanic males. Annual incidence rates at 60 years of age range from 0.4% to 3.7% in these groups. Bayesian multilevel analysis accounts for sources of variation when combining multiple years of NHIS survey data and allows estimation of the posterior distribution of general functions of model parameters.

email: tat5@cdc.gov

COMPARISON OF ESTIMATORS FOR AVERAGE MEDICAL COSTS IN A MARKOV MODEL

Lin Liu*, Michigan State University
Zhehui Luo, Michigan State University
Joseph C. Gardiner, Michigan State University

We combine a multi-state, non-homogeneous Markov model and a mixed-effects model to estimate net present value for costs accumulated over a finite time period. Costs are incurred through medical care use while patients transit between, and sojourn in health states. The Markov model governs transitions between health states and a multiplicative intensity model is used to estimate transition probabilities. A mixed-effects model is used to estimate the average transition and sojourn costs. Due to the incompleteness of medical cost data, we propose a new estimator for average transition or sojourn cost under the informative censoring assumption that the number of transitions depends on individual subject covariates. Through simulation, we assess the properties of this estimator compared to an unweighted generalized least squares estimator regardless of censoring, and a weighted ordinary least squares estimator based on uninformative censoring.

email: liulin@msu.edu

RECEIVER OPERATING CHARACTERISTIC CURVE ANALYSIS USING THE JACKKNIFE: AN APPLICATION TO DIAGNOSTIC ACCURACY FOR PIGMENTED LESIONS

Paul Kolm*, Emory University
Michelle L. Pennie, Emory University
Suephy C. Chen, Emory University

The area under a receiver operating characteristic (ROC) curve is a standard measure of accuracy of a diagnostic system. When multiple observers or readers rate the same set of cases, comparison of the mean AUC for two or more groups of readers ignores random error variance associated with an AUC. The purpose of this study was to apply a jackknife method to compare dermatologists and primary care physicians (PCP) with respect to accuracy in the diagnosis and management of melanoma. Dermatologists and PCPs were rated 30 pictures of randomly selected pigmented lesions in response to A) What is the likelihood you would biopsy (dermatologists) / refer (PCPs) this patient?, and B) What is your suspicion that this is a melanoma? The jackknife method was used to compute AUC pseudovalues, and these data were then analyzed to compare mean AUCs of dermatologists and PCPs using linear mixed model analysis. For Question A, the AUC was 0.833 for dermatologists and 0.755 for PCPs ($p < 0.001$). For Question B, the AUC was 0.894 for dermatologists and 0.797 for PCPs ($p < 0.001$). The jackknife method allows an analysis that properly accounts for random patient variance in the assessment of diagnostic accuracy.

email: paul.kolm@emory.edu

A DYNAMIC FORECASTING MODEL OF DIAGNOSED DIABETES IN THE U.S. (2005-2050)

James P. Boyle*, Centers for Disease Control and Prevention
Theodore J. Thompson, Centers for Disease Control and Prevention

A Markov model was used to generate forecasts by age, race/ethnicity and sex of diagnosed diabetes prevalence in the United States through 2050. The model forecasts the number of individuals in each of three states (diagnosed with diabetes, not diagnosed with diabetes, and death) in each year using the following inputs: estimated diagnosed diabetes prevalence and incidence for 2004; the relative risk of mortality from diabetes compared with no diabetes; and U.S. Census Bureau current population (2004) and projections of live births, net migration, and total mortality rates. Logistic regression models with noninformative priors yielded 5000 posterior draws for prevalences, incidences, and relative risks through MCMC simulation. The Markov projection model is then run and results saved using each of the 5000 samples yielding a sample from the posterior distributions of all quantities of interest including forecasts (posterior means) and Bayesian confidence intervals. For example, the projected number of people with diagnosed diabetes increases from 16.2 million in 2005 (95% confidence interval of 15.6-16.8 million) to 48.3 million in 2050 (45.1-51.5 million), implying an increase in total U.S. diagnosed diabetes prevalence from 5.6% in 2005 (5.4-5.8%) to 12.0% (11.2-12.8%) in 2050.

email: jboyle@cdc.gov

A COMPREHENSIVE MALDI-TOF MS DATA PREPROCESSING METHOD USING FEEDBACK CONCEPT

Shuo Chen*, Vanderbilt University
Ming Li, Vanderbilt University
Huiming Li, Vanderbilt University
Don Hong, Vanderbilt University
Dean Billheimer, Vanderbilt University
Yu Shyr, Vanderbilt University

Mass Spectrometry(MS) can generate high throughput protein profiles for biomedical applications. Most existing MS data analysis tools contain two parts: the preprocessing part which extracts and quantifies the features from raw spectra, and a statistical analysis part. It is crucial to perform the preprocessing part well, because subsequent analyses are determined by its output. For MALDI-TOF MS Data, a consistent, sensitive and robust preprocessing method is needed. In this paper, we propose a new comprehensive MALDI-TOF MS data preprocessing method using feedback concepts and new algorithms. All preprocessing steps and reference information work together dynamically. This new "Wave-Spec" package successfully resolves many conventional difficulties such as systematic/random spectrum shifts caused by different experiment time and location, objectively setting denosing parameters, and peak alignment across spectra. Also, we compare this new method with existing approaches through simulation and real data.

email: shuo.chen@vanderbilt.edu

TESTS FOR COMPARISON OF TWO POISSON MEANS

Kangxia Gu, Southern Methodist University
Hon Keung Tony Ng*, Southern Methodist University
Man Lai Tang, Hong Kong Baptist University-Kowloon, Hong Kong

In this paper, we investigate different test procedures for comparing two Poisson means. Asymptotic tests, tests based on approximate p-value method and the likelihood ratio test are considered. Size and power performances of these tests are studied by means of Monte Carlo simulation under different settings. Some recommendations are made based on these simulation results. We illustrate these testing procedures with a breast cancer example.

email: kangxiag@mail.smu.edu

STOCHASTIC OPTIMIZATION FOR PARAMETER ESTIMATION IN FRAILTY MODELS

Tim C. Hesterberg*, Insightful Corporation

We begin with an introduction to stochastic optimization (SO), including abuse of SO by applying it to a deterministic problem -- one-pass estimation of logistic regression parameters. We continue with the use of SO for parameter estimation in frailty models, where the likelihood is estimated using Monte Carlo integration. This provides substantial computational saving, roughly by a factor of 50 when 50 iterations are used to maximize the likelihood, and can be used for very large data sets.

email: timh@insightful.com

A NONPARAMETRIC METHOD OF BACKGROUND CORRECTION FOR MICROARRAY DATA ANALYSIS

Zhongxue Chen*, Southern Methodist University
Monnie McGee, Southern Methodist University

Probe level data preprocessing is very important for microarray data analysis. This will affect the following high-level analysis, such as gene selection, classification and clustering. Background correction is one of the three steps of preprocessing and it has a great influence on the next steps. The three most commonly used background correction methods are MAS5.0, RMA and dchip. Mas5.0 uses the information of perfect match and mismatch, while the other two methods are usually only based on perfect match. It has been shown that MAS5.0 has a poorer performance compared with the other methods based on spikein dataset. However, RMA method has its own problems, such as parameter estimation and model fitting. We propose a new background correction method, which will use information from both of PM and MM. We use the lowest q2 percentile of MM that associated with the lowest q1 percentile of PM to estimate the background noise. Based on the estimated background noise, we also propose a normalization method. This new method is compared with other methods by using the spikein dataset. The results show that our method has a very good performance.

email: zhongxue@mail.smu.edu

FALSE DISCOVERY RATE AND MULTIPLE TESTING CORRECTIONS IN DISEASE-MARKER ASSOCIATION STUDIES

Julia Kozlitina*, Southern Methodist University
William R. Schucany, Southern Methodist University
Patrick S. Carmack, UT Southwestern Medical School

Current advances in genome-sequencing technology have allowed researchers to generate large numbers of single nucleotide polymorphisms (SNPs) mapped to the entire genome or to specific genes. Disease-marker association studies based on dense SNP maps test thousands and thousands of features in order to identify possible disease-causing variations. Consequently, appropriate multiple testing adjustments must be made to avoid an abundance of false positive conclusions. When the number of expected significant SNPs is relatively large, false discovery rate (FDR) has been used successfully as an alternative to FWER control providing a sensible balance between the number of false positive and false negative results. When the number of truly significant markers is extremely small, however, even the FDR-based significance cutoffs may turn out to be too restrictive, leading to virtually no rejections. This becomes critical in the presence of positive dependency among test statistics causing an overabundance of large p-values. Using the data from 11, 600 SNPs typed on a sample of over 2500 individuals we demonstrate the use of FDR in conjunction with other methods, such as cross-validation, in order to increase power of discoveries and avoid spurious effects.

email: jkozliti@mail.smu.edu

HONEYCOMB DESIGNS COMPUTING AND ANALYSIS

Andy Mauromoustakos*, University of Arkansas
Vasilia Fasoula*, University of Georgia
Kevin Thompson, University of Arkansas

The primary objective of plant breeders is to identify and select superior genotypes from among and within a broad array of genetic entries. Honeycomb designs (HD) are a set of systematic designs capable of handling a large number of genetic entries and a large number of replications. The designs sample effectively for environmental diversity by means of large number of moving replicates. These designs were developed to carry out efficient selection among genetic entries through the partition of crop yield into three genetic components and efficient selection of the best plants within the selected entries by means of the moving-ring single-plant selection. Single-plant selection in honeycomb trials starts from the very early generations of the plant breeding program, which speeds up the process of cultivar improvement and development. This paper will focus on the development of a JSL script (JSL is the scripting language of JMP) for creating the design, field layout and subsequently performing the required calculations that would lead to selecting best entries and best plants within entry. The HD selection and analysis are implemented by the breeder with a powerful visual interface and easy to use interface to carrying out the needed calculations. Examples of previously published data on honeycomb trials will be provided.

email: amauro@uark.edu

GRAPH AND HYPERGRAPH-BASED MODELS IN BIOINFORMATICS: PRESENT AND FUTURE

Sujay Datta*, Northern Michigan University

Models based on graphs, directed graphs and multigraphs have been around for quite some time with interesting applications in, for example, computer science, statistics, sociology and market research. Nowadays a generalization of graphs, called hypergraphs (directed and undirected), are being increasingly used in modeling complex biological networks and interactions. Examples of biological datasets that lend themselves to hypergraphical models are metabolic pathways, signaling pathways, gene regulatory networks and protein interaction networks. Various graph-based models differ in terms of their chosen view of reality, their coverage and their granularity (or resolution). Some available graphical models are compound graph models, reaction graph models, hypergraph models and object oriented models. Here, after an overview of the basics of graphical modeling, we focus on hypergraph models (related to bipartite graph models), examine their pros and cons and discuss several general applications such as path finding, pathway synthesis, comparison of metabolic pathways, signal transduction pathway modeling, etc. We end with a special example of the yeast protein complex network.

email: sdatta@euclid.nmu.edu

USING ROBUST ESTIMATORS CAN INCREASE THE POWER OF THE SHAPIRO-WILK TEST AGAINST HEAVY-TAILED ALTERNATIVES

Joseph L. Gastwirth*, George Washington University
Weiwen Miao, Macalester College
Yulia Gel, University of Waterloo

The Shapiro-Wilk (SW) test is often used as a preliminary check that the data come from a normal distribution before the t-test and similar procedures are applied. Since p-values and confidence coefficients calculated from the standard tables are more severely affected by heavy tailed distributions, we propose that either one of two robust estimators of the standard deviation be used in place of s, in the SW test. It is shown that these robust SW (RSW) tests have higher power than the usual method and are related to a graphical method of checking normality of the authors.

email: jlgast@gwu.edu

A NUMERICAL STUDY OF THE PROBLEM OF INSUFFICIENT OVERLAP IN PROPENSITY SCORES WHEN ESTIMATING A CAUSAL TREATMENT EFFECT

Yuliya Lokhnygina*, Duke University and Duke Clinical Research Institute
Karen Chiswell, North Carolina State University and Duke Clinical Research Institute

The seminal paper by Rosenbaum and Rubin (1983) outlined the role of the propensity score in estimating causal effects based on observational data. Their key assumption of strongly ignorable treatment assignment has two components. The first is that the potential outcomes are conditionally independent of treatment assignment, given any value of the propensity score. The second is that the conditional probability of receiving treatment is strictly between 0 and 1, given any value of the propensity score. This second component has been called by various names, for example, the sufficient overlap or common support requirement. In practice, lack of sufficient overlap may be detected by comparing the distributions of the estimated propensity scores in the two treatment groups. However there appears to be no clear rule about what constitutes sufficient overlap, or whether sufficient overlap depends in any way on the method used to estimate the causal treatment effect. In this poster we numerically investigate potential effects of a lack of sufficient overlap on the estimation of causal effects. We consider various approaches to estimating causal effects including stratification and two examples of inverse-probability-of-treatment-weighted estimation: the simple weighted estimator of Rosenbaum (1998) and the "doubly-robust" weighted estimator of Robins et al. (1999).

email: yuliya.lokhnygina@duke.edu

SPATIAL ANALYSIS OF PROBE LEVEL INTENSITIES IN AFFYMETRIX GENECHIP MICROARRAYS

Kinfemichael A. Gedif*, Southern Methodist University
Andrew Hardin, Southern Methodist University
William R. Schucany, Southern Methodist University
Monnie McGee, Southern Methodist University

It is assumed that spatial effects are minimal or nonexistent in Afymetrix chips since probes for a given gene are scattered around the array. Hence, little investigation of the spatial effects in Affymetrix genechips has been done. However, spatial effects can occur both at the image level data (.DAT files) and the probe level summaries (.CEL files) due to various reasons, including the fluorescence technology used, the method used to get probe level intensities, or scanner effects. We investigated spatial dependencies between intensities of neighboring probes within the micro array at the probe level. Significant spatial autocorrelations were observed for some arrays studied. We can use the spatial correlation structure to estimate non-biological effects and account for these during background correction.

email: kgedif@smu.edu

email: zelen@hsph.harvard.edu

ENAR

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

ABSTRACTS

I. ADVANCES IN SPATIAL AND TEMPORAL MODELING

SPACE-TIME MODELING OF GLOBAL OZONE LEVELS

Mikyoung Jun*, Texas A&M University
Michael L. Stein, University of Chicago

We present how we develop a rich and flexible class of parametric space-time covariance function for processes on spheres. We focus on modeling space-time covariance structure, especially the features such as space-time asymmetry and different covariance structure at different latitude levels, which appear to be common among air pollution processes. We apply our new covariance functions to global total column ozone levels and compare the fitted results from our models to other existing covariance functions. Computational issues such as how to deal with large covariance matrices and how to explore the special structures of the covariance matrices will also be discussed.

email: mjun@stat.tamu.edu

COVARIANCE TAPERING FOR INTERPOLATION OF LARGE SPATIAL DATASETS

Reinhard Furrer*, Colorado School of Mines
Marc G. Genton, Texas A&M University
Douglas Nychka, National Center for Atmospheric Research

Interpolation of a spatially correlated random process is used in many areas. The best unbiased linear predictor, often called kriging in geostatistical science, requires the solution of a large linear system based on the covariance matrix of the observations. In this talk I show that tapering the correct covariance matrix with an appropriate compactly supported covariance function reduces the computational burden significantly and still has an asymptotic optimal mean squared error. The effect of tapering is to create a sparse approximate linear system that can then be solved using sparse matrix algorithms. Further, the manageable size of the observed and predicted fields can be far bigger than with classical approaches. The net result is the ability to analyze spatial data sets that are several orders of magnitude larger than past work in a high level interactive environment such as R. I will briefly discuss related approaches and extensions.

email: rfurrer@mines.edu

SEQUENTIAL ESTIMATION OF SPATIO-TEMPORAL MODELS

Jonathan R. Stroud*, University of Pennsylvania

We consider the problem of real-time forecasting of space-time processes. We first describe the computational challenges inherent in sequential state and parameter estimation for high-dimensional systems. We also propose some new performance metrics for space-time forecast validation. A case study of ozone monitoring in Mexico City is used to illustrate these issues.

email: stroud@wharton.upenn.edu

ON OPTIMAL POINT AND BLOCK PREDICTION IN LOG-GAUSSIAN RANDOM FIELDS

Victor De Oliveira*, University of Arkansas

This work discusses the problems of point and block prediction in log-Gaussian random fields with unknown mean. New point and block predictors are derived that are optimal in mean squared error sense within certain families of predictors that contain the corresponding lognormal kriging point and block predictors, as well as a block predictor originally motivated under the assumption of "preservation of lognormality," and hence improve upon them. A comparison between the optimal, lognormal kriging and best linear unbiased predictors is provided, as well as between the two new block predictors. Somewhat surprisingly, it is shown that the corresponding optimal and lognormal kriging predictors are almost identical under most scenarios. It is also shown that one of the new block predictors is uniformly better than the other.

email: vdo@uark.edu

2. MISSING DATA IN LONGITUDINAL STUDIES: PARAMETRIC AND SEMIPARAMETRIC PERSPECTIVES

ISSUES IN MULTIPLE IMPUTATION FOR HIERARCHICAL DATA

James R. Carpenter*, London School of Hygiene & Tropical Medicine
Mike G. Kenward, London School of Hygiene & Tropical Medicine

Multiple imputation provides a practical approach for applied statisticians faced with the analysis of partially observed hierarchical and longitudinal data. However, some key issues remain to be resolved, such as whether it is appropriate to use a multivariate normal model for imputing discrete data, how convenient are more principled alternatives, and the extent to which it is possible to provide an accessible, general approach to assess the sensitivity of inferences to the MAR assumption <p>. We illustrate our ideas with data from hierarchical social science and public health studies.

email: jrc@imbi.uni-freiburg.de

MULTIPLE IMPUTATION FOR NONIGNORABLY MISSING DATA USING A BAYESIAN LATENT-CLASS SELECTION MODEL

Joseph L. Schafer*, The Pennsylvania State University
Hyekyung Jung, The Pennsylvania State University

Data are said to be nonignorably missing if the probabilities of missingness depend on unobserved quantities. Traditional selection models for nonignorable nonresponse are outcome-based, tying these probabilities to the partially observed values directly (e.g., by a logistic regression) and are inherently unstable. With multivariate or longitudinal data, the number of distinct missingness patterns also becomes large, making outcome-based selection modeling unattractive. Information in the binary missing-data indicators is sometimes be well summarized by a simple latent-class structure, however, suggesting that the partially observed variables could be tied to probabilities of class membership. In this talk, we examine the properties of multiple imputations created under a Bayesian latent-class selection model.

email: jls@stat.psu.edu

TESTING FOR MISSING DATA MECHANISMS USING QUADRATIC INFERENCE FUNCTION

Grace Y. Yi*, University of Waterloo
Annie Qu, Oregon State University
Peter Song, University of Waterloo

Incomplete longitudinal data analysis has been receiving increasing attention in the literature. A variety of inferential methods have been proposed to address the complexity caused by missingness. Typically marginal methods such as the inverse probability weighted generalized estimating equations (IPWGEE) approaches have been extensively employed to conduct valid inference. However, these approaches are subject to the correct specification of the weights that are determined by missing data processes. In this talk we will discuss a test procedure for assessing the validity of the IPWGEE approaches. The performance of the proposed procedure will be evaluated through numerical studies.

email: yyi@likelihood.math.uwaterloo.ca

NONLINEAR MIXED-EFFECTS MODELS WITH DROPOUTS AND MISSING COVARIATES

Wei Liu, University of British Columbia
Lang Wu*, University of British Columbia

Semiparametric nonlinear mixed-effects (NLME) models provide flexible tools for analyzing longitudinal data. Covariates are often introduced in the models to partially explain inter-individual variation. Some covariates, however, may be measured with substantial errors and may have missing values. We propose two approximate likelihood methods for semiparametric NLME models with measurement errors and missing data in covariates. The first method may be more accurate but is often computationally intensive, while the second method is computationally much more efficient but may be less accurate. The methods are illustrated by a real data example. Simulations results show that both methods perform well and are better than two commonly used methods.

email: lang@stat.ubc.ca

3. RECENT ADVANCES IN THE ASSOCIATION ANALYSIS FOR MULTIVARIATE FAILURE TIME DATA

THE KENDALL DISTRIBUTION WITH BIVARIATE CENSORED DATA

David Oakes*, University of Rochester
Antai Wang,Georgetown University

For a bivariate survival model governed by an archimedean copula (for example, a bivariate frailty model) we derive a simple formula for the conditional distribution of the bivariate survivor function at the (partly) unobserved full (uncensored) observation corresponding to an observed singly or doubly censored observation. We show how this result can be used in problems of estimation and goodness-of-fit and compare our approach with techniques based on the full data.

email: oakes@bst.rochester.edu

FUNCTIONAL ASSOCIATION MODELS FOR MULTIVARIATE TEMPORAL PROCESSES

Jason Fine*, University of Wisconsin-Madison

We consider multivariate temporal processes that are continuously observed within overlapping time windows. The intended application is censored multistate and multivariate survival settings, where point processes are continuouly observed. This data differs from other discretely observed functional data, like longitudinal data. Functional mean and association regression models are studied for point processes, with unspecified time-varying coefficients. The observation scheme is exploited: the coefficients are estimated nonparametrically by extending GEE to continuously observed data. The estimators automatically converge at parametric rate, without smoothing, unlike discretely observed data. Uniform convergence properties are derived with empirical process techniques. Existing functional approaches to survival association analyses employ intensity models, which require smoothing and depend critically on smoothing parameter selection, similarly to discretely observed data. Our procedure yields new tests for associations, parametric sub-modeling of these parameters, and goodness-of-fit testing. An analysis of familial aggregation of alcoholism illustrates the methodology's practical utility in assessing dynamic associations.

email: fine@biostat.wisc.edu

ANALYSIS OF FAILURE TIME DATA WITH MULTI-LEVEL CLUSTERING, WITH APPLICATION TO THE CHILD VITAMIN A INTERVENTION TRIAL IN NEPAL

Joanna Shih*, National Cancer Institute, National Institutes of Health
Shou-En Lu, University of Medicine and Dentistry of New Jersey

We consider the problem of estimating covariate effects in the marginal Cox proportional hazard model and multi-level associations for child mortality data collected from a vitamin A supplementation trial in Nepal (Nepal Nutrition Intervention Project-Sarlahi, or NNIPS), where the data are clustered within households and villages. For this purpose, a class of multivariate survival models that can be represented by a functional of marginal survival functions and accounts for hierarchical structure of clustering is exploited. Based on this class of models, an estimation strategy involving a within-cluster resampling procedure is proposed. The asymptotic theory for the proposed estimators is established, and the simulation study shows that the estimates are consistent. The analysis of the NNIPS study data shows that the association of mortality is much greater within households than within villages.

email: jshih@mail.nih.gov

4. ILS: INTRODUCTION TO STATISTICAL GENETICS

INTRODUCTION TO ASSOCIATION STUDIES: BASIC CONCEPTS & METHODS

Rudy Guerra*, Rice University

Phenotype-genotype correlations have traditionally been investigated through linkage analysis based on pedigree data. More recently, association studies based on population data have been promoted largely on the argument of increased statistical power relative to linkage studies. This talk will serve as an introduction to association studies. The emphasis will be on basic genetic concepts, study design, data, and some of the common statistical methods used in this area. Specific topics include linkage disequilibrium, case-control studies, SNP data, haplotypes and haplotype blocks, limitations, and a summary of the success of association studies for complex traits.

email: rguerra@rice.edu

OVERVIEW ON STRUCTURAL ASSOCIATION TESTING AND REGIONAL ADMIXTURE MAPPING

David B. Allison*, The University of Alabama at Birmingham
T.M. Beasley, The University of Alabama at Birmingham
Jose R. Fernandez, The University of Alabama at Birmingham
David T. Redden, The University of Alabama at Birmingham
Hemant K. Tiwari, The University of Alabama at Birmingham
Jasmin Divers, The University of Alabama at Birmingham
Robert Kimberly, The University of Alabama at Birmingham

Individual genetic admixture estimates, determined both across the genome and at specific genomic regions, have been proposed for use in identifying specific genomic regions harboring loci influencing dichotomous phenotypes in regional admixture mapping (RAM). Estimates of individual ancestry can be used in structured association tests (SAT) to reduce confounding induced by various forms of population substructure. Although presented as two distinct approaches, we provide a conceptual framework in which both RAM and SAT are special cases of a more general linear model which allows for greater modeling flexibility, adaptation to multiple designs, inclusion of covariates, interaction terms, and multi-locus models. We clarify which variables it is sufficient to control for in analyses and also provide a simple closed-form 'semi-parametric' method of estimating the reliability of individual admixture estimates used as individual ancestry estimates that makes an inherent errors-in-variables problem tractable. This approach offers SAT and RAM methods enormous flexibility, enabling application to a richer set of phenotypes, populations, covariates, and situations.

email: dallison@uab.edu

INTRODUCTION TO MICROARRAYS

Russell D. Wolfinger*, SAS Institute, Inc.

The rapid increase in volumes of molecular data continues to drive a strong need for biostatisticians who are capable of analyzing and interpreting them, not to mention the continued flurry of statistical research in this area. This introductory lecture is for biostatisticians who are familiar with the buzz surrounding gene expression and microarray data and would like to learn more about them and become more involved. We will begin with an overview of some of the fundamental underlying principles of molecular biology and attempt to set gene expression in context with the previous introductory lectures on statistical genetics. Next, we will discuss the most common kinds of modern high-throughput instrumentation generating gene expression and related data. We will then survey the wide range of interesting statistical problems that have arisen and continue to arise from these data, including such areas as experimental design, normalization and quality control, pattern discovery, statistical modeling and inference (including multiple testing), data mining and prediction, and bioinformatics.

email: russ.wolfinger@sas.com

5. IMS: NON-STANDARD MAXIMUM LIKELIHOOD INFERENCE

RATES OF CONVERGENCE FOR CURRENT STATUS DATA WITH COMPETING RISKS

Marloes Maathuis*, University of Washington

We consider nonparametric estimation of the sub-distribution functions for current status data with competing risks. In particular, we study the asymptotic behavior of the nonparametric maximum likelihood estimator (MLE) and the 'naive estimator' proposed by Jewell, van der Laan and Henneman (2003). We present results on global and local consistency, and on the global and local rates of convergence. We show that both the MLE and the naive estimator achieve the optimal $n ^ {-1/3}$ local rate of convergence. Finally, we present a simulation study indicating that the MLE is superior to the naive estimator, both for finite sample sizes and asymptotically.

email: marloes@stat.washington.edu

INFERENCE UNDER RIGHT CENSORING FOR TRANSFORMATION MODELS WITH A CHANGE-POINT BASED ON A COVARIATE THRESHOLD

Michael R. Kosorok*, University of Wisconsin-Madison
Rui Song, University of Wisconsin-Madison

We consider linear transformation models applied to right censored survival data with a change-point in the regression coefficient based on a covariate threshold. We establish consistency and weak convergence of the nonparametric maximum likelihood estimators. The change-point parameter is shown to be n-consistent, while the remaining parameters are shown to have the expected root-n consistency. We show that the procedure is adaptive in the sense that the non-threshold parameters are estimable with the same precision as if the true threshold value were known. We also develop Monte Carlo methods of inference for model parameters and score tests for the existence of a change-point. A key difficulty here is that some of the model parameters are not identifiable under the null hypothesis of no change-point. Simulation studies establish the validity of the proposed score tests for finite sample sizes.

email: kosorok@biostat.wisc.edu

NONCONCAVE PENALIZED LIKELIHOOD INFERENCE FOR MULTIVARIATE SURVIVAL DATA

Jianwen Cai, University of North Carolina at Chapel Hill
Jianqing Fan*, Princeton University
Runze Li, Pennsylvania State University
Haibo Zhou, University of North Carolina at Chapel Hill

Nonconcave penalized likelihood has been demonstrated in Fan and Li (2001) as a viable class of variable selection procedures. It has been shown there that the resulting procedures perform as well as if the subset of significant variables were known in advance. Such a property is called an oracle property. In this talk, a penalised pseudo-partial likelihood method is proposed for variable selection with multivariate failure time data. Even with a growing number of dimensionality, under certain regularity conditions, we show the consistency and asymptotic normality of the penalised likelihood estimators. We further demonstrate that, for certain penalty functions with proper choices of regularisation parameters, the resulting estimator can correctly identify the true model, as if it were known in advance. Based on a simple approximation of the penalty function, the proposed method can easily be carried out with the Newton-Raphson algorithm. We conduct extensive Monte Carlo simulation studies to assess the finite sample performance of the proposed procedures. We illustrate the proposed method by analysing a dataset from the Framingham Heart Study.

email: jqfan@princeton.edu

LIKELIHOOD INFERENCE UNDER MONOTONICITY CONSTRAINTS: SOME RECENT DEVELOPMENTS

Moulinath Banerjee*, University of Michigan

In this talk, I will discuss some recent developments in likelihood inference under monotonicity constraints. While it has been known for a while that maximum likelihood estimators in monotone function models exhibit cube root convergence rates with a non-Gaussian limit (Chernoff's distribution), the study of likelihood ratios in such problems is only a recent affair. It turns out that the study of likelihood ratios is much more fruitful from an estimation perspective, since the limit distribution of the likelihood ratio statistic for testing the value of a monotone function at a point is 'universal' -- it is generally free of the underlying model parameters (and the model) and is characterized in terms of slopes of convex minorants of Brownian motion plus a quadratic drift. Inversion of the likelihood ratio statistic yields confidence sets in a wide variety of problems, without the need to estimate nuisance parameters from the data. I will discuss some of my recent work in this area, highlighting the statistical implications and propose extensions and remaining challenges.

email: moulib@umich.edu

BAYESIAN INFERENCE OF THE LEAD TIME IN PERIODIC 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 develops a probability distribution for the lead time in periodic cancer screening exams. The general aim is to provide statistical inference for the lead time, the length of time the diagnosis is advanced by screening. The lead time is distributed as a mixture of a point mass and a piecewise continuous density. Simulation studies are carried out, using the HIP data, to estimate under different screening time intervals, the proportion of breast cancer patients who truly benefit from the periodic screening exams and the proportion that do not. The mean, the mode, the variance and the density curve of the lead time are also provided. This provides important information to policy makers regarding the screening period and the long-term benefit for women who take part in the periodic screening exams. Though the study is focused on breast cancer screening, it is also applicable to other kinds of chronic disease.

email: dwu@math.msstate.edu

THE MULTI-PHASE OPTIMIZATION STRATEGY: A NEW WAY TO DEVELOP MULTI-COMPONENT 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

Developing optimal multi-component treatments is a challenging issue in clinical and behavioral medicine. But design and analysis strategies to address this issue are not well-developed. In this paper we discuss a new methodology, called multi-phase optimization strategy (MOST), to proactively develop, optimize, and evaluate multi-component behavioral or medical interventions. This new method consists of three ordered phases of experimentation. Using a simulation study, we illustrate the superiority of the new procedure over the traditional approach of formulating a multi-component treatment and immediately proceeding to a confirmatory two-arm randomized trial.

email: bibhas@umich.edu

ROBUST ESTIMATION FOR THE MEAN MEDICAL EXPENDITURE

Kenny Shum*, Johns Hopkins University
Scott L. Zeger, Johns Hopkins University

Research on medical expenditure usually focuses on the total cost or difference in spending. The mean is the sole parameter of interest because other location parameters lack the relation to the total. Using the sample mean is optimal when data are normally distributed; however, the distribution of medical cost is heavily skewed. A slight change in the tail-behavior of the distribution will have large influence on the mean. An alternate approach is to fit a parametric model such as the lognormal distribution to the cost data, but the resulting maximum likelihood estimator is not robust to the distributional assumption. In this talk, we study the robustness of a class of mean estimators. We propose a novel estimator based on smoothing the quantile spacings. We also define the log-Gamma distribution, extending the generalized gamma distribution, to represent the range of shapes encountered with medical cost and other right skewed variables. Through a simulation study and asymptotic arguments, we compare the estimators' performances. Finally, we construct confidence intervals for the mean of a skewed population using these robust estimators and compare with other bootstrap intervals.

email: kshum@jhsph.edu

SAMPLE SIZE REQUIREMENTS FOR STUDYING SMALL POPULATIONS IN GERONTOLOGY

Robert B. Noble, Miami University
A. John Bailer*, Miami University
Suzanne R. Kunkel, Miami University
Jane K. Straker, Miami University

Calculating sample sizes required to achieve a specified level of precision when estimating population parameters is a common statistical task. As consumer surveys become increasingly common for nursing homes, home care agencies, other service providers, and state and local administrative agencies, standard methods to calculate sample size may not be adequate. Standard methods typically assume a normal approximation and require the specification of a plausible value of the unknown population trait. We present a strategy to estimate sample sizes for small finite populations and when a range of possible population values is specified. This sampling strategy is hierarchical, employing first a hypergeometric sampling model which directly addresses the finite population concern. This level is then coupled with a beta-binomial distribution for the number of population elements possessing the characteristic of interest. This second level addresses the concern that the opulation trait may range over an interval of values. The utility of this strategy is illustrated using a study of resident satisfaction in nursing homes.

email: baileraj@muohio.edu

MODELING DIFFERENTIATED ASSOCIATIONS BETWEEN PHYSIOLOGICAL DYSREGULATION AND FRAILTY IN OLDER WOMEN

Hongfei Guo*, Johns Hopkins University
Karen Bandeen-Roche, Johns Hopkins University

Frailty is a geriatric syndrome of physical vulnerability characterized by high susceptibility and low physiological reserve. Often frailty affects multiple physiological systems that are quantified by multiple biomediators. Researchers tend to investigate the association between frailty and individual biomediators, or a summary of a few biomediators. In contrast, this study aims to explore the association between multiple biomediators and frailty. We propose a new methodology to identify possibly differentiated associations of frailty with the multiple biomediators, i.e., to identify a few groups of biomediators, each of whose means differ similarly across frail and non-frail populations, and estimate the magnitude of mean difference per biomediator group. We consider the identification of the groups and mean difference estimation by maximum likelihood, using the MCMCEM algorithm. A simulation study revealed excellent performance of our methodology provided correctness of assumed model. Using data from the Women's Health and Aging Study (WHAS), we illustrate our methodology together with two alternative methods: summarizing the biomediators into a few scores and then applying multivariate regression, and analyzing the biomediators using multivariate regression and then summarizing the coefficients. Our methodology aims to extend the tools available for researchers to investigate the complex, multiply-measured health outcomes.

email: hfguo@jhsph.edu

A LOCALLY WEIGHTED REGRESSION APPROACH TO EVALUATING THE EFFECTS OF SMOKING REDUCTION ON BIRTH WEIGHT

Jeff M. Szychowski*, University of Alabama
J. Michael Hardin, University of Alabama
Michael D. Conerly, University of Alabama
Wendy Horn, Cooper Green Hospital, Jefferson Health System (CCOE)
Lesa Woodby, University of Alabama at Birmingham

While smoking cessation is known to improve neonatal outcomes in pregnant smokers, reducing tobacco exposure has been more controversial. Winsor et al showed that reducing tobacco exposure by 50% or more during pregnancy lead to beneficial effects on birth weight on neonatal outcomes. However, other researchers have questioned the benefit of reduction in the absence of cessation. Some researchers have argued that these conflicting reports are due to the nonlinear relationship between tobacco exposure and birth weight, e.g. England, et al, 2001. In this paper, we employ a locally weighted regression approach to evaluate the relationship between tobacco exposure and birth weight using the data from the Smoking Cessation or Reduction in Pregnancy Trial (SCRIPT).

email: jeffszychowski@yahoo.com

ADJUSTING LONGITUDINAL CONFOUNDING VARIABLES

Haiqun Lin*, Yale University

This research intends to evaluate the effect of repeatedly measured predictor of interest on longitudinal outcome in the presence of repeatedly measured confounding variable, all could be continuous variables irregularly measured over time. Our work extends marginal structural framework proposed by Robins and colleagues by using weights that are estimated from modeling the relationship between the longitudinal predictor of interest and the confounding variable. Our approach is illustrated with simulation and a real data set from health service study.

email: haiqun.lin@yale.edu

7. DESIGNING CLINICAL TRIALS

CLINICAL TRIALS SIMULATION: OVERVIEW AND DEMONSTRATION OF A NEW SYSTEM

Stephan Ogenstad*, Vertex Pharmaceuticals Incorporated
Peter H. Westfall, Texas Tech University
Kuenhi Tsai, Vertex Pharmaceuticals Incorporated
Leif Bengtsson, Vertex Pharmaceuticals Incorporated
Scott Moseley, Vertex Pharmaceuticals Incorporated
Min Yao, Vertex Pharmaceuticals Incorporated
AlinTomoiaga, Texas Tech University
Lan Zhang, Texas Tech University

While simulation analysis is well known to statisticians, interest in clinical trials simulation has recently exploded in popularity among clinicians and phamacokineticists. Applications of clinical trials simulation include design and protocol optimization, estimation of operating characteristics of nonstandard and computationally intensive procedures, and development of “mock up” trials for training review committees. Extant commercial software often require pharmacokinetic/pharmacodynamic inputs that can be difficult to specify. Instead we develop a rich probabilistic model to account for typical clinical trial scenarios, using historical data were possible to validate the outputs. The output data sets are massive and the analyses can be computationally challenging; these problems are solved through the application of an interactive, menu-driven grid computing system developed using SAS software.

email: stephan_ogenstad@vrtx.com

PREDICTING EVENT TIMES IN CLINICAL TRIALS WHEN RANDOMIZATION IS MASKED AND BLOCKED

J. Mark Donovan*, University of Pennsylvania
Michael R. Elliott, University of Michigan
Daniel F. Heitjan, University of Pennsylvania

Timing for interim or final analysis of data in an event-based trial is often determined based on the accrual of events during the course of the study. Existing Bayesian methods may be used to predict the data of the landmark event, based on current enrollment, event, and loss to follow-up. This work extends methods where the treatment arms are masked to incorporate information about a block randomization. A latent variable model with blocking information (LV with blocking) is compared with methods where blocking information is ignored (constrained LV) and methods assuming a single population (SP). Comparison of the LV model with blocking with the unconstrained LV model in our application shows that the median estimate of the landmark event dates are similar, but the LV model with blocking has narrower prediction limits. Simulation studies show that the LV model with blocking, with diffuse priors, can have better coverage probabilities for the prediction inteveral than SP models if a treatment effect is present, and prediction limits from the LV model with blocking are narrower than those for the unconstrained LV model.

email: jdonovan@cceb.upenn.edu

EFFECT OF DROPOUTS ON COST-EFFICIENCY OF HIGHER-ORDER CROSSOVER DESIGNS IN COMPARATIVE BIOAVAILABILITY CLINICAL TRIALS

Jihao Zhou*, Allergan, Inc.
Jane Li, University of Michigan

Cost-efficient trial designs have become crucial due to unprecedented high cost of drug development [FDA (2005), Drug Development Science]. Cost-efficient crossover design has drawn recent attention [Zhou, et al (2005), Clin Pharmacol Ther]. However, dropouts, often seen in higher-order crossover designs, were not considered in their investigation. This study is to explore the effect of dropouts on the cost efficiency of five commonly used, statistically optimal or nearly-optimal higher-order crossover designs. Three dropout patterns (decreasing, constant, and increasing) were simulated after generating multivariate normal data under scenarios of a wide range of variability and correlations (CV = 10% to 40%; ñ = 0.2 to 0.8). Monte Carlo simulations and mixed-effects models were carried out to obtain empirical sample sizes for each design using Schuirmann's TOST Procedure, under an 80% power and a 5% significance level, based on the FDA bioequivalence criteria (80%-125%). The five designs and the cost function were described by Zhou, et al (2005). Results show that D3x2 is generally the best with dropouts. D4x4, often the best design without dropouts, becomes the second worst with dropouts. D4x2 is the best when the screening cost is high and period cost does not vary. D2x4 is still the worst.

email: Zhou_jihao@Allergan.com

STOCHASTIC CURTAILMENT IN MULTI-ARMED TRIALS

Xiaomin He*, University of Rochester

Stochastically curtailed procedures in multi-armed trials are complicated due to repeated significance testing and multiple comparisons. From either frequentist or Bayesian viewpoints, there exists some dependence among pairwise test statistics. Investigators must consider such dependence when testing homogeneity of treatments. This paper studies the property of canonical multivariate joint distribution of test statistics in multi-armed trials. Pairwise and global monitoring are suggested based on this property. In pairwise monitoring, the Hochberg step-up procedure is recommended to strongly control the overall significance level. In global monitoring, the conditional and predictive power are calculated based on current multivariate test statistics, which reflect the dependence among pairwise test statistics. Futility monitoring in multi-armed trials is also considered. Simulation results in multi-armed trials show that, compared with the traditional group sequential and non-sequential procedures, stochastic curtailment has advantages in sample size, time and cost. An example concerning a proposed study of Coenzyme Q$_{10}$ in early Parkinson Disease is given.

email: xiaominhe@bst.rochester.edu

EXAMINATION OF THE EFFICIENCY OF THE SEQUENTIAL PARALLEL DESIGN IN PSYCHIATRIC CLINICAL TRIALS

Roy N. Tamura*, Eli Lilly and Company
Xiaohong Huang, Eli Lilly and Company

The sequential parallel design consists of two phases of a clinical trial, an initial phase in which patients are randomized to placebo and drug and a second phase in which placebo non-responders are randomized to placebo or drug. The efficiency of this design compared to the conventional two arm trial is examined for both binary and for continuous efficacy data. In the continuous data case, we propose using seemingly unrelated regression to combine the inference over the two phases of the trial. The results of our examinations suggest that for both binary and efficacy data, a reduction of the sample size around 20% is possible over the conventional design. The trade off between smaller sample size versus longer duration of trial is also discussed within the context of an antidepressant clinical trial.

email: tamura_roy_n@lilly.com

SINGLE-STAGE SIMULTANEOUS TESTING OF SUPERIORITY AND NON-INFERIORITY IN ACTIVE CONTROL CLINICAL TRIALS

Yongzhao Shao, New York University School of Medicine
Vandana Mukhi*, New York University School of Medicine
Judith D. Goldberg, New York University School of Medicine

We discuss some issues that arise in the design of clinical trials to allow simultaneous testing for both superiority and non-inferiority using a single-stage procedure. In the design of such a single-stage procedure, the sample size of the trial is generally based on the primary objective of non-inferiority or superiority; therefore, the type II errors that correspond to the secondary objective are not always controlled. When these type II errors are large, a single-stage design may not be appropriate. We propose the evaluation of power as well as the conditional level of the additional test for the secondary hypothesis, conditioning on the decision of the first test, to assess whether appropriate levels of reproducibility are achievable in a single-stage design. For various typical phase III clinical trial designs, the results of our simulation studies indicate that the proposed approach is useful to assess whether it is appropriate to design a single-stage procedure to test both superiority and non-inferiority simultaneously. When the additional test appears to have inadequate power, sample size adjustment and other strategies are discussed.

email: vandana.mukhi@med.nyu.edu

SCPRT DESIGN FOR CLINICAL TRIALS WITH SURVIVAL DATA

Xiaoping Xiong*, St. Jude Children's Research Hospital
Ming Tan, University of Maryland-Greenebaum Cancer Center
James Boyett, St. Jude Children's Research Hospital

The sequential conditional probability ratio test (SCPRT) offers several properties useful for design of clinical trials. The conclusion of the sequential test would not be reversed if the sequential test were not stopped as it should but be continued to the planned end of trial. The trial by this design is flexible such that interim looks can be ignored or added during the process of trial without affecting the characteristics of design. The sample size of the sequential test will not go beyond the sample size of the fixed sample test design with a same significance level and power. The design is efficient by having expected sample sizes close to that of Wald's SPRT. We propose SCPRT designs for clinical trials with survival data in which we considered situations of comparing two prospective arms and comparing a prospective treatment arm with an historical control arm.

email: xiaoping.xiong@stjude.org

8. BAYESIAN METHODS AND APPLICATIONS

BAYESIAN IMAGE ANALYSIS OF CHANGES IN BRAIN/TUMOR PERMEABILITY INDUCED BY RADIOTHERAPY USING REVERSIBLE JUMP MARKOV CHAIN MONTE CARLO

Xiaoxi Zhang*, University of Michigan
Timothy D. Johnson, 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. In this application, we use Hidden Markov Random Fields to model the change in vascular permeability. We assume an unobservable label underlying each voxel (the basic volume element in Magnetic Resonance Imaging) that characterizes the change, and a first order Markovian dependence structure of the labels, which is also known as the Potts model. We model the observed change in permeability as independent Gaussian random variable given the component label. The Reversible Jump Markov Chain Monte Carlo algorithm enables the Markov chain to jump between sub-models with parameters of different dimensions. Furthermore, we estimate the temperature parameter in the Potts model using Path sampling instead of fixing it beforehand, and hence incorporate extra flexibility into the model. We present simulation results and a preliminary analysis of the real data. With the success on one patient, future work will focus on longitudinal analysis across multiple patients.

email: xiaoxi@umich.edu

RECONSIDERING THE VARIANCE PARAMETERIZATION IN MULTIPLE PRECISION MODELS

Yi He*, University of Minnesota
James S. Hodges, University of Minnesota
Bradley P. Carlin, University of Minnesota

Recent developments in Bayesian computing allow accurate estimation of integrals, making advanced Bayesian analysis feasible. However, some problems remain difficult, such as estimation of posterior distributions for variance parameters. For models with three or more variances, this paper proposes a new simplex parameterization that has appealing geometric properties and also eases the related burden of specifying a prior. The simplex parameterization has at least two attractive features. First, it typically leads to MCMC algorithms that are simple and have good mixing properties, regardless of the parameterization used to specify the model's reference prior. Second, the simplex parameterization suggests a natural reference prior for the simplex parameters that is proper, scale-invariant, and appears to reduce the maximum posterior correlation with the error precision. We use simulations to compare the simplex parameterization with its reference prior to other parameterizations with their reference priors on the bases of bias, mean-squared error, and posterior 95\% credible interval coverage. The results suggest significant advantages for the simplex approach, particularly when the error precision is small. We also offer results in the context of two real data sets from the fields of periodontics and dentistry, again showing the benefit of our new approach.

email: hydinghua@gmail.com

A BAYESIAN SCREENING PROCEDURE FOR IDENTIFYING 'SIGNALS' OBTAINED BY DATA MINING FROM SPONTANEOUS REPORT ADVERSE EVENT DATABASES

A. Lawrence Gould*, Merck Research Laboratories

Surveillance of drug products in the marketplace continues after regulatory approval, to identify rare potential toxicities that are unlikely to have been observed in the clinical trials carried out before approval. This surveillance accumulates large numbers of spontaneous reports of adverse events in spontaneous report databases. Recently developed empirical Bayes and Bayes methods provide a way to summarize the data in these databases, including a quantitative measure of the strength of the reporting association between the drugs and the events. Determining which of the particular drug-event associations, of which there may be many tens of thousands, are real reporting associations and which random noise presents a substantial problem of multiplicity because the resources available for medical and epidemiologic followup are limited. This article describes a Bayesian screening method for identifying potential 'signals' from spontaneous reporting databases that appears to have attractive diagnostic properties in addition to being easy to interpret and implement computationally.

email: goulda@merck.com

COMBINING BOOTSTRAP AND BAYESIAN INFERENCES

Yan Zhou*, University of Michigan
Jack D. Kalbfleisch, University of Michigan
Roderick J.A. Little, University of Michigan

In the case of independent identically distributedsamples, the naive bootstrap yields confidence limits that are asymptotically correct to the first order, but have less certain confidence coverage in small samples. Bayesian credibility intervals based on the posterior distribution of the model parameters tend to perform better for small samples, but are more dependent on modeling assumptions than the bootstrap. A discrepancy statistic based on the difference of model and bootstrap estimates of standard error is used as a basis for combining bootstrap and Bayesian inferences. The goal is to achieve a compromise that combines the advantages of those two methods, yielding intervals that combine robustness with good small-sample confidence coverage. We assess properties of our method by some simple simulation experiments.

email: yzhouz@umich.edu

EMPIRICAL BAYES ESTIMATION FOR ADDITIVE HAZARDS REGRESSION MODELS

Sinha Debajyoti*, Medical University of South Carolina
Stuart Lipsitz, Medical University of South Carolina
Brent McHenry, Bristol-Meyer & Squibb

We develop an empirical Bayesian framework for a semiparametric additive hazards regression model of Aalen (1980) by using a gamma-process prior on the unknown baseline cumulative hazard function. The marginal likelihood obtained via integrating the prior process can be maximized using standard statistical softwares and the empirical Bayes estimates of regression parameters, survival curves and their standard errors have easy to compute closed form expressions. This marginal likelihood based methodologies have superior properties compared to currently available methods of estimation based on ordinary least squares and method of moments. We illustrate our semiparametric empirical Bayes methodology via a reanalysis of a survival dataset using existing statistical softwares such as SAS.

email: sinhad@musc.edu

DOSE-FINDING BASED ON THE MAXIMUM DIFFERENCE IN THE PROBABILITY OF RESPONSE AND THE PROBABILITY OF TOXICITY

Yuan Ji, M.D. Anderson Cancer Center
Yisheng Li*, M.D. Anderson Cancer Center
B. Nebiyou Bekele, M.D. Anderson Cancer Center

In this article, we propose a new dose-finding algorithm for cancer phase I/II clinical trials with the aim to find the maximum difference dose (MDD), i.e., the dose that maximizes the difference between the probability of response and the probability of toxicity. Based on a single flexible probability model, the proposed dose-finding algorithm is able to accommodate two types of trials in which the probability of toxicity increases or does not increase with the dose level. In addition, a new approach of penalizing assignment of too many patients at one dose and too few at another, especially in the early stage of a trial, is proposed to improve the chances of finding the MDD. Extensive simulations are carried out to compare the proposed algorithm with two methods in the recent literature.

email: ysli@mdanderson.org

A DISTANCE APPROACH TO BAYESIAN MODEL DIAGNOSTICS

Guan Xing*, Case Western Reserve University
J. Sunil Rao, Case Western Reserve University

In Bayesian data analysis, posterior predictive checking has been generally adopted for model diagnostics, including graphical checks and some statistical tests (e.g., Chi-sq test, F test). We propose a new test statistic for model consistency check using the distance between observations and the posterior sampled data. Several groups of data are sampled from the posterior predictive distribution. Then, the distance between observations and the sampled data is compared with those between-group distances and the empirical P value is calculated. Very large or small P value will indicate the inconsistency between the statistical model and observations. Several distance measures are applied to different hierarchical models of simulated and real data sets. The performance and the practical usage are discussed. Comparison with other tests is also addressed.

email: gxx4@case.edu

9. BIOASSAY AND BIOPHARMACEUTICAL APPLICATIONS

STOCHASTIC MODELING OF HUMAN COLON CANCERS: A MIXTURE APPROACH

Wai-Yuan Tan, University of Memphis
Lijun Zhang*, University of Memphis
Chao-Wen Chen, EPA
Junmei Zhu, University of Memphis

In this paper we have developed a stochastic mixture model of 5 different pathways for human colon cancers with each pathway being a stochastic multi-stage model of carcinogenesis. There pathways are: the sporadic LOH pathway (70%), the familial LOH pathway (14%), the FAP pathway (1%), the sporadic MSI pathway (10%) and the HNPCC pathway (5%). For this model, we derive an EM algorithm to estimate the proportion of different pathways and the parameters for each pathway. We derive the MLE for each pathway through the genetic algorithm. We have applied this model to fit and analyze the SEER data of human colon cancers from NCI/NIH. Our results indicate that for the time to tumor, it takes 55-60 years by the sporadic LOH pathway, about 40 years by the FAP pathway, 50-55 years for the familial LOH pathway, 65-70 years for the sporadic MSI pathway and 45-50 years for the HNPCC pathway.

email: lzhang1@memphis.edu

AN ANALYTICAL TOOL FOR ASSAY DEVELOPMENT ON PROTEIN CHIP PLATFORM

Steven Novick*, GlaxoSmithKline

The protein chip platform is a multi-plex assay system that enables simultaneous measurements of biological activities of many proteins from one sample. Our primary goal for the protein chip project is to develop high throughput, multiplexed expression arrays to analyze up to 50 proteins in a single experiment. My primary interest lies with a protein-chip platform for which the protein expression level for an unknown sample is measured as the calibrated (back-calculated) value from a standard curve. In this talk, I will be discussing a tool to assist in the development of such a platform. Two important features of interest in such a system are (1) What is the smallest quantifiable protein level (i.e., the limit of quantification, 'LOQ')? and (2) What is the accuracy of a calibrated concentration? I propose a commonly-used metric system to respond to feature (2) called an operating characteristic (OC) curve. Given specification limits on the accuracy of a calibrated concentration, one can use the OC curve to examine the potential usefulness for each protein assay. For assay development, the OC curve can be a critical tool for comparing parameter changes and optimization of experimental protocols.

email: steven.j.novick@gsk.com

ASSESSING INDIVIDUAL AGREEMENT VIA INDIVIDUAL EQUIVALENCE

Huiman X. Barnhart*, Duke University
Andrzej S. Kosinski, Duke University
Michael J. Haber, Emory University

Evaluating agreement between methods or observers is important in method comparison study and reliability study. Often we are interested in whether a new method can replace an existing invasive or expensive method, or in whether the multiple methods/observers can be used interchangeably. Ideally, interchangeability is established only if individual measurements from these methods are like replicated measurements within a method. Interchangeability between methods is similar to bioequivalence between drugs in bioequivalence studies. Following the FDA guidelines on individual bioequivalence, we propose to assess individual agreement among multiple methods via individual equivalence using the moment criteria. In the case where there is a reference method, we extend the individual bioequivalence criteria to individual equivalence criteria (IEC) to compare multiple methods to one or multiple references. In the case where no reference method is available, we propose a new IEC to assess individual agreement between multiple methods. Furthermore, a coefficient of individual agreement (CIA) is proposed to link the IEC with two recent agreement indices. A method of moments is used for estimation and the bootstrap approach is used to construct one-sided 95% confidence bound. Five examples are used for illustration.

email: huiman.barnhart@duke.edu

EXTENDING TOLERANCE INTERVALS FOR PREDICTION INTERVAL COVERAGE

Jacqueline R. Wroughton*, University of Nebraska
Erin E. Blankenship, University of Nebraska
James R. Schwenke, Boehringer-Ingelheim Pharmaceuticals Inc.
Walter W. Stroup, University of Nebraska

The traditional definition of a tolerance interval describes it as an interval estimate of a given percentile of a distribution. Alternatively, a tolerance interval can be described as a confidence interval about the bounds of a confidence interval. It can be shown that when the coverage probability of a tolerance interval is set to 50%, the noncentrality parameter controlling the width of the tolerance interval is zero, giving a tolerance interval equivalent to a confidence interval. The statistical theory which defines a tolerance interval is not directly extended to prediction interval coverage. Often prediction interval estimates are required, such as in stability studies. For stability studies, specification limits are set as a measure of product quality for current and future batches. The primary purpose of this research is to find statistical support for setting specification limits, specifically when shelf life estimation is to be based on prediction intervals. A bootstrap solution for a tolerance interval for prediction interval coverage will be presented.

email: flagqueen1@yahoo.com

ON SEARCHING FOR TREND IN GENE EXPRESSION USING ORIOGEN

Shan Chen, Northwestern University
Irene B. Helenowski, Northwestern University
Raymond C. Bergan, Northwestern University
Borko D. Jovanovic*, Northwestern University

We use ORIOGEN freeware to study trend in response to doses of genistein in PC3 and PC3M cell lines. First we briefly discuss the underlying statistical theory, and follow this by discussion of cell lines and doses involved. Since the 6 doses cross two orders of magnitude, and time of exposure to treatment is 24-72 hours, it is important to properly define parameters of search for possible trends prior to using ORIOGEN. We provide a list of genes which seem to show 'some trend' in their expression.

email: borko@northwestern.edu

MODEL AVERAGING IN DICHOTOMOUS DOSE-RESPONSE RISK ESTIMATION

Matthew W. Wheeler*, Risk Evaluation Branch, NIOSH
A. John Bailer, Risk Evaluation Branch, NIOSH and Miami University

Model averaging (MA) has been proposed as a method of accommodating model uncertainty when estimating risk. For example, MA strategies have been used to synthesize the risk estimates (benchmark doses) derived from a family of dichotomous dose-response risk models. We conducted a Monte Carlo simulation study to examine the characteristics of MA-based risk estimates where model-specific risk estimates were averaged. This MA approach was very sensitive to the underlying set of models chosen to perform the averaging calculation. We present a different approach, in which risk is estimated from the averaged model (i.e., the weighted average of the parametric dose-response models) and its corresponding lower bound is computed by bootstrap. We further study this technique through a simulation. Preliminary results suggest coverage closer to or at nominal coverage probabilities with similar bias properties. Further, these results continue to highlight the importance of choosing a proper set of models when using MA in risk estimation.

email: aez0@cdc.gov

ASSESSING DRUG INTERACTION UNDER DIFFERENT EXPERIMENTAL CONDITIONS

Maiying Kong*, M. D. Anderson Cancer Center
J. Jack Lee, M. D. Anderson Cancer Center
Dan Ayers, M. D. Anderson Cancer Center

Assessing drug interactions under experimental conditions beyond dose changes is relatively new. For example, consider a k1xk2xk3 factorial dosing structure for a three-drug synergy experiment, say, pre-treated AZA (a DNA demethylation agent), post-treated AZA, and SAHA (a histone deacetylase inhibitor). The response for each combination dose is a machine concentration of dye, which is proportional to the concentration of alive cells. Imposed on this structure, multiple cell lines are grown with or without bovine serum. Both cell line and serum effects are of interest. We present procedures to examine this data, assess drug interactions among pre-AZA, post-AZA, and SAHA, and analyze the possible interactions among serum, cell lines, and different agents. Graphical methods including trellis plots will be shown. Drug interaction will be analyzed under the framework of general linear model. In addition, both parametric and semi-parametric methods will be applied to determine whether the three drugs work synergistically, additively, or antagonistically under various experimental conditions.

email: mykong@mdanderson.org

10. GENERALIZED LINEAR MODELS

COMBINING STUDIES TO CALCULATE A BOUND ON A REGRESSION COEFFICIENT

Chand K. Chauhan*, Indiana-Purdue University-Fort Wayne
Yvonne M. Zubovic, Indiana-Purdue University-Fort Wayne

Suppose two independent studies are conducted, one providing an estimate of the regression coefficient B(XY) between the variables X and Y, and the other providing an estimate of the regression coefficient B(YZ) between Y and another variable Z. How can the results of the two studies be combined to provide information concerning the regression coefficient B(XZ) that relates X and Z? It is a well known fact that if random variables X and Y are positively (or negatively) correlated with correlation R(XY) and if Y and Z are positively (or negatively) correlated with correlation R(YZ), then X and Z are not necessarily positively (or negatively) correlated. Some authors provide a region identifying the values of R(XY) and R(YZ) which guarantee that R(XZ) is greater than or equal to a specified value. In the present paper we extend these results to identify a region in which B(XZ) assumes a specified lower limit. This region depends on the values of B(XY), B(YZ), and the standard deviations ó(X), ó(Y), and ó(Z). We characterize this region and examine different ratios of the standard deviations and their effects on the region of interest.

email: chauhan@ipfw.edu

A PRACTICAL APPROACH TO COMPUTING POWER FOR GENERALIZED LINEAR MODELS WITH NOMINAL, COUNT, OR ORDINAL RESPONSES

Robert H. Lyles*, Emory University
Hung-Mo Lin, Penn State College of Medicine
John M. Williamson, Centers for Disease Control and Prevention

Many current proposals for approximating power under generalized linear models with binary, ordinal, or count outcomes are computationally demanding, limited in terms of accommodating covariates, or have not been assessed for accuracy assuming moderate sample sizes. We present a simple method for estimating conditional power that requires only standard software for fitting the desired generalized linear model for a non-continuous outcome. The model is fit to an appropriate expanded dataset using easily calculated weights that represent response probabilities given the assumed values of the parameters. The variance-covariance matrix from this fit is used in conjunction with an established non-central chi square approximation to the distribution of the Wald statistic. Alternatively, the model can be re-fit under the null hypothesis to approximate power based on the likelihood ratio statistic. We provide guidelines for constructing a representative expanded dataset to allow close approximation of unconditional power based on the assumed joint distribution of the covariates. Relative to prior proposals, the approach proves particularly flexible for handling one or more continuous covariates without any need for discretizing.

email: rlyles@sph.emory.edu

A CLASS OF MARKOV MODELS FOR LONGITUDINAL ORDINAL DATA

Keunbaik Lee*, University of Florida
Michael Daniels, University of Florida

Generalized linear models with serial dependence are often used for short longitudinal series. Heagerty(2002) has proposed marginalized transition models for the analysis of longitudinal binary data. In this paper, we extend this work to accommodate longitudinal ordinal data. Fisher-scoring algorithms are developed for estimation. Methods are illustrated on quality of life data from a recent colorectal cancer clinical trial.

email: lee@stat.ufl.edu

TESTING HOMOGENEITY IN FINITE MIXTURES AND MIXTURE REGRESSION MODELS

Hongying Dai*, University of Kentucky
Richard Charnigo, University of Kentucky

Testing homogeneity in Finite mixtures is a challenge in statistical inference due to nonidentifiability of parameters under the null hypothesis. Chen, Chen,and Kalbfleisch (2001) proposed the modified likelihood ratio test (MLRT) and Charnigo and Sun (2004) proposed the L ^ 2-based D- test for homogeneity in two-component mixtures. In this work, we will extend the D-test and the MLRT to mixture models with arbitrary but fixed components. The limiting distributions of two test statistics along with the analytic forms of the maximum modified likelihood estimators are provided. Simulation experiments will compare the performance of two competing tests. These two tests can be applied to a wide class of parametric families because of the realistic assumptions. By postulating the link function, latent variables, we can test the homogeneity of mixture regression models, e.g. mixture Poisson regression models, mixture logistic regression models.

email: hdai2@uky.edu

ROBUST ESTIMATION FOR ZERO-INFLATED REGRESSION MODELS

Jing Shen*, University of Georgia
Daniel B. Hall, University of Georgia

Zero inflated (ZI) regression models comprise an important subclass of finite mixture models that is useful for data that contain many zeros. The maximum likelihood (ML) estimation approach is used widely for such models. It is well known that the ML estimator can become very unstable when the data are subject to contamination, such as the presence of outliers or violations of the assumed underlying data generating mechanism. We consider two alternative robust estimation approaches, minimum Hellinger distance (MHD) estimation and robust expectation-solution (RES) estimation in this paper. While simulation results indicate that both methods improve substantially on ML estimation when outliers are present and/or when the mixture components are poorly separated, the MHD method is more narrowly applicable because of identifiability problems that can arise when the mixing probability is modelled as a function of covariates. In addition, the asymptotic properties of MHD in the regression context are difficult to establish. In contrast, the RES method can easily be shown to yield consistent and asymptotically normal estimators and can be applied quite generally. An example involving data related to aggressive behavior among sixth grade school-children is presented to illustrate the methods.

email: jingshen@stat.uga.edu

A COMPARISON OF MODELS FOR IMMUNOLOGICAL CORRELATES OF PROTECTION

Fabrice Bailleux*, Sanofi Pasteur-Lyon, France
Andrew Dunning, Sanofi Pasteur-Swiftwater, USA

Vaccines prevent disease by inducing the immune system to create antibodies to disease-causing pathogens. In the process, immunological memory is created, so that if the disease-causing pathogen invades the body, the immune system is primed to attack and destroy the pathogen. The level of antibody induced by vaccination can be measured by immunological assays. A question of interest is the quantitative relationship between antibody concentrations and subsequent protection from disease. Various models have been proposed to model the relationship. Chang and Kohberger used a threshold level of antibody corresponding to the efficacy of pneumococcal vaccine. Chan et al fitted Weibull, log-normal, log-logistic and piecewise exponential models to cases of varicella among vaccinated children. Dunning has proposed a scaled logit model. Logistic regression has been frequently used. The features and results from different models will be compared and presented using historical and simulated datasets.

email: andrew.dunning@sanofipasteur.com

MODEL SELECTION FOR GENERALIZED LINEAR MODEL

Bo Hu*, University of Wisconsin-Madison
Jun Shao, University of Wisconsin-Madison
Mari Palta, University of Wisconsin-Madison

The problem of model selection in generalized linear regression is investigated. A novel selection procedure under the Bregman Divergence is proposed. The new method is obtained by peneralizing the loss with a convariance penalty measuring complexity of the candidate model. Under rather weak conditions, the new procedure is shown to be consistent in selecting the optimal model by maximizing the peneralized loss among a large class of candidate models. The method is applicable for both univarate GLM and marginal longitudinal model. Numerical results are presented to demonstrate the effectiveness of new procedure in finite sample applications.

email: boymsn@yahoo.com

11. CAUSAL INFERENCE

ESTIMATION AND CONFIDENCE REGIONS FOR MULTI-DIMENSIONAL EFFECTIVE DOSE

Jialiang Li*, University of Wisconsin
Erik V. Nordheim, University of Wisconsin
Chunming Zhang, University of Wisconsin
Charles E. Lehner, University of Wisconsin

The problem of finding confidence regions for multiple predictor variables corresponding to given expected values of a response variable has not been adequately resolved. Motivated by an example from a study on hyperbaric exposure using a logistic regression model, we develop a conceptual framework for the estimation of the multi-dimensional effective dose for binary outcomes. The k-dimensional effective dose can be determined by conditioning on k-1 components and solving for the last component as a conditional univariate effective dose. We consider various approaches for calculating confidence regions for the multi-dimensional effective dose and compare them via a simulation study for a range of possible designs. We analyze some data related to decompression sickness to illustrate our procedure. Our results provide a practical approach to finding confidence regions for predictor variables for a given response value.

email: jialiang@stat.wisc.edu

STRUCTURAL NESTED MEAN MODELS FOR ASSESSING TIME-VARYING EFFECT MODERATION: AN ILLUSTRATION

Daniel Almirall*, University of Michigan
Thomas R. Ten Have, University of Pennsylvania School of Medicine
Susan A. Murphy, University of Michigan

This article considers the problem of assessing causal effect moderation in longitudinal settings in which treatment (or exposure) is time-varying and so are the covariates said to moderate its effect. Intermediate Causal Effects that describe time-varying causal effects of treatment conditional on past covariate history are introduced and considered as part of Robins' Structural Nested Mean Model. Two estimators of the intermediate causal effects, and their standard errors, are presented and discussed: The first is a proposed 2-Stage Regression Estimator, which can be used using standard regression software. The second is Robins' G-Estimator. The methodology is illustrated using longitudinal data from the randomized controlled trial PROSPECT. Our purpose is to estimate the time-varying causal effects of adherence to the intervention on depression outcomes at the end of study, conditional on time-varying covariates that may modify these effects.

email: dalmiral@umich.edu

PREDICTING TREATMENT MEANS IN A ONE-WAY FACTORIAL DESIGN BASED ON A POTENTIAL OBSERVABLE RANDOM VARIABLE FRAMEWORK

Bo Xu*, University of Massachusetts
Edward J. Stanek III, University of Massachusetts

The potentially observable random variable framework of Little and Rubin (2002) provides a context for defining causal effects. Although the role of the treatment allocation model in developing inference is extensively discussed in the literature, the joint roles of sampling and treatment allocation are seldom considered. We introduce sampling and treatment level allocation random variables to represent the potentially observable population by a joint set of random variables for a 1 factor model. This provides a design based framework for inference about treatment level means. Treatment level means are represented as the sum of the sample and remainder random variables, with Royall's (1976) prediction theory used to develop predictors of the remainder. The approach is non-parametric, and based solely on the population sampling and treatment allocation random variables. We refer to the predictors of the realized treatment level mean as the best linear unbiased predictors. The predictors have the property of being 'shrunk' towards the overall mean, similarly to the predictors of realized random effects in mixed models. Comparison of the expected MSE of these predictors with the simple sample treatment group mean illustrates the advantage of this approach.

email: stanek@schoolph.umass.edu

MACHINE LEARNING METHODS FOR OBSERVATIONAL STUDIES

Debashis Ghosh*, University of Michigan

In many scientific studies, interest focuses on the analysis and interpretation of data from nonrandomized studies. A major issue in the consideration of such data is the fact that assignment of exposure of treatment has been done in a nonrandomized manner. This has spurred many methods for the analysis of observational data. Many have focused on the use of the propensity score, which is the probability of treatment assignment given covariates. Modelling the propensity score requires considerations of main effects and interactions. In this work, we consider flexible nonparametric and semiparametric models for the propensity score using support vector regression methods from the machine learning literature. The ideas are illustrated with data from a liver cancer study.

email: ghoshd@umich.edu

CAUSAL ANALYSIS OF BINARY RESPONSES

Haihong Li*, University of Florida
P. V. Rao, University of Florida

Most literature on recursive causal modeling concerns methods suitable for continuous response data. Under this setting, the total effects can be decomposed into the sum of direct effects and indirect effects to help understand the causal relationship. Assuming a hierarchical structural model, these methods utilize the Ordinary Least Squares (OLS) techniques for estimating and testing the direct and indirect effects of the response (dependent, endogenous) variables in the model. When dealing with binary response variables, the OLS methods have to be replaced with generalized linear models approach, such as Poisson or logistic regression methods. Quantifying the direct and indirect effects in the generalized linear models is not a trivial extension from the continuous case. Rigorous definitions are often too complicated to hold any practical value. In this work we propose an approximation to the causal effects which is intuitive and easy to evaluate. The maternal child health and education data in the state of Florida is used to illustrate the causal inference for binary responses.

email: hl_98@yahoo.com

SELECTION OF AVERAGE CAUSAL EFFECT (ACE) MEASURES FOR BINARY OUTCOMES USING PROPENSITY SCORE SUBCLASSIFICATION

Yi Huang*, Johns Hopkins University Bloomberg School of Public Health
Karen Bandeen-Roche, Johns Hopkins University Bloomberg School of Public Health
Constantine Frangakis, Johns Hopkins University Bloomberg School of Public Health

Propensity scoring (Rosenbaum, Rubin, 1983,1984) has become increasingly popular for estimating average causal effects (ACE). However, little attention has been devoted to the choice of different ACE measures in analysis of binary outcomes. Such attention is needed because different choices of ACE measures are differentially subject to the failure of collapsibility, defined in our paper as equality of the marginal ACE to a weighted average of bin-specific causal effects. Such attention is also needed because subject-specific knowledge and scientific aims may suggest reference of certain ACE measures over others. So, our first aim is to clarify the collapsibility properties of different ACE measures, and associated consequences for estimation by propensity score subclassification. These properties suggest an inherent advantage for the interpretability of the average risk difference and marginal relative risk over the marginal odds ratio. Our second aim is to reveal a connection between the choices of ACE measures and subject-specific knowledge of comparative potential risks under exposure/treatment and no exposure/no treatment. We suggest a graphical way to visualize this connection. Collectively, this work suggest how the effective choices of ACE measures can be based not only on statistical convenience, but also on their interpretability and subject specific knowledge.

email: yhuang@jhsph.edu

HETEROGENEOUS VARIANCES IN PRINCIPAL STRATIFICATION MODELS

Robert J. Gallop*, West Chester University
Thomas R. Ten Have, University of Pennsylvania

In many areas of research, the main goal is assessing the direct effect of treatment. While this is such a simple goal, arriving at an answer is challenged by the presence of mediating and confounding factors which may cause the improvement in outcome more so than the treatment effect. Imbens and Rubin (1997) considered a similar setting where the issue was focusing on the presence of counterfactual outcomes. Their setting considered patient who were randomized to treatment versus the action performed by the patient under this treatment assignment. A compliance status is determined by the patient's randomization and the patient's action under this randomization. The compliance status consists of four states: compliers, always-takers, never-takers, and defiers. One method to analyze such data is the Principal stratification models (Frangakis and Rubin, 2002). Model assumptions for the Principal Stratification are normality and constant variance across the observed and unobserved groups. This current research described here relaxes the homogeneity of variance assumption, by allowing the model to fit separate variance between subgroups.

email: rgallop@wcupa.edu

12. NEW DEVELOPMENTS IN MICROARRAYS: IDENTIFYING DIFFERENTIALLY EXPRESSED GENES AND METHODS FOR BUILDING PREDICTION MODELS

SELECTION AND USE OF PATHWAYS FOR PROGNOSIS

Hans C. van Houwelingen*, University Medical Center-The Netherlands
Jelle J. Goeman, Leiden, University Medical Center-The Netherlands

Since more and more information becomes available about pathways that might be helpful in understanding gene-expression data, it is tempting to investigate the role of pathways in existing data sets. As an example we will consider the relevance of different pathways for the breast cancer survival data set of Van de Vijver et al. (2002).First of all, we will show how the recent global test for survival data of Goeman et al. (2005) can be used as a screening test for pathways. Next we relate the outcome of the global test to the predictive value of the (genes in the) pathway when applying the cross-validated penalized Cox regression of Van Houwelingen et al. (2005).Finally, we discuss ways of combining predictive information selected pathways into a single prognostic model. References: van de Vijver, MJ; et al.,. A gene-expression signature as a predictor of survival in breast cancer.. NEW ENGLAND JOURNAL OF MEDICINE 347 (25): 1999-2009, 2002.Goeman JJ, et al., Testing association of a pathway with survival using gene expression data, Bioinformatics, 21, 1950-1957, 2005.van Houwelingen HC, et al., Cross-validated Cox-regression on microarray data, to appear in Statistics in Medicine (Preview on Wiley Interscience)

email: jcvanhouwelingen@lumc.nl

REGULARIZED INFERENCE FOR MICROARRAYS

Hemant Ishwaran*, Cleveland Clinic Foundation

DNA microarrays yield high-throughput information about a cell's proteomic composition (using nuclear RNA) and thus biologic insight into molecular differences between cells. However, analysis of microarray data is challenging because large amounts of information are collected using relatively small sample sizes. Regularization, the technique of using information across all genes, is critical and fundamental to addressing this issue. In this talk I present various methods for regularization, including new theory for rescaled spike and slab models, and discuss applications to multigroup microarray data. The methodology will be illustrated using platform independent Java software available at www.bamarray.com.

email: ishwaran@bio.ri.ccf.org

A PARAMETRIC BOOTSTRAP METHOD FOR MODEL SELECTION IN PENALIZED LOGISTIC REGRESSION FOR DISEASE CLASSIFICATION USING MICROARRAY DATA

Jason Liao*, University of Medicine and Dentistry of New Jersey
Yong Lin, University of Medicine and Dentistry of New Jersey

Building a prediction model for disease classification using microarray gene expression data presents a unique challenge in having far more predictors than the subjects. The model selection becomes the paramount issue. Recently there have been considerable progresses in this area. We shall propose a parametric bootstrap method for model selection that, for the first time, provides a unified framework for answering two key questions: how many genes to include in the logistic model and how to select the penalty parameter.

email: jg_liao@yahoo.com

VARIANTS OF THE SUPPORT VECTOR MACHINE AND THEIR APPLICATIONS TO MICROARRAY CLASSIFICATION

Ji Zhu*, University of Michigan

The support vector machine is a widely used tool for classification. In this talk, we start with a brief introduction to the standard 2-norm support vector machine, and write it as a regularized optimization problem. Based on that, we consider several variants of the support vector machine, specifically, penalized logistic regression, the 1-norm support vector machine, and the doubly regularized support vector machine. We argue that these variants may have some advantage over the standard 2-norm support vector machine under certain situations, for example, when there are redundant noise variables. We also propose efficient algorithms to solve the optimization problems posed by these variants. In the end, we compare these models on a microarray cancer dataset. This talk consists of a collection of joint work with Saharon Rosset (IBM), Hui Zou (Stanford), Trevor Hastie (Stanford) and Rob Tibshirani (Stanford).

email: jizhu@umich.edu

13. ASSESSING SPATIAL SURVEILLANCE FOR BIOTERRORISM: SIMULATING ATTACKS

DETECTING SIMPLE SIMULATED ANTHRAX ATTACKS: COMPARISON OF CLUSTER IDENTIFICATION METHODS VIA NEW ASSESSMENT METRICS

Ken Kleinman*, Harvard Medical School/Harvard Pilgrim Health Care
Allyson Abrams, Harvard Medical School/Harvard Pilgrim Health Care

There are many proposed methods of identifying clusters in surveillance data. However, there are few ways to choose amongst them. The area under the ROC curve (AUROC) is one option. Unfortunately, defining the ROC in this context is not straightforward. It also ignores the timeliness, but a method with AUROC=1 is useless if it detects outbreaks too late. We define the AUROC for simulated disease clusters added to real data and develop ROC-like tools that include timeliness. Analogous to the AUROC, each tool has a maximum of 1, suggesting perfect sensitivity and specificity and that all the signals generated immediately. Each method can be weighted to ascribe more importance to early detection. We also assess the performance of the tools in comparing seven cluster detection methods. Across a range of simulations, the tools which included timeliness were notably similar. In contrast, the area under the ROC curve resulted in much bigger values. The proposed methods provide in one number a sense of the key surveillance characteristics: sensitivity, specificity, and timeliness. This is a vast improvement over methods that output two values or ignore one key characteristic.

email: ken.kleinman@gmail.com

A SIMULATION MODEL FOR EVALUATING OUTBREAK DETECTION

David L. Buckeridge*, McGill University

Surveillance directed towards the timely detection of outbreaks relies increasingly on data collected for other purposes. Addresses, when available in these data, are usually home addresses, and work addresses are not routinely available. Many surveillance systems analyze healthcare utilization data with spatial statistics, which use the geographic location of cases to search for disease clusters. This reliance on home address may limit the power of spatial analysis for outbreak detection. If home addresses are used for analysis of cases when exposure occurred away from the home then the cases may appear falsely geographically dispersed, and this could reduce the power of detection.
In this talk I will consider modeling population spatial mobility for the purpose of evaluating its impact on outbreak detection. I will identify sources of data that are useful for modeling population mobility and describe briefly different approaches to developing models from available data. In addition, I will present results from a simulation study that estimates the impact of population mobility on outbreak detection using temporal and spatial statistics. Finally, I will discuss how mobility models can be incorporated into statistical methods, such as the spatial scan statistic.

email: david.buckeridge@mcgill.ca

PROTECTING PUBLIC HEALTH THROUGH ADVANCED SPATIO TEMPORAL EPIDEMIOLOGICAL MODELING

James H. Kaufman*, IBM Research Division
Daniel A. Ford, IBM Research Division

In this talk we describe the Spatio Temporal Epidemiological Modeler (STEM), an extensible framework for modeling the propagation of diseases. Diseases evolve, changing diseases that are well tolerated by adapted populations into killers. Some diseases also exist in different species and can mutate by swapping genetic material. An example is the well-publicized Avian Influenza A (H5N1). The STEM framework allows simultaneous modeling of spatial and temporal progression of multiple diseases affecting multiple population types (e.g., Humans and Birds). It also facilitates the rapid exchange of new models and relevant data. It also allows arbitrary computational modeling within the scope of these representations. It is an ideal tool for modeling a complex multi-species disease such as Avian Influenza A. STEM also provides an ability to perform 'what if' experiments designed to test the benefits of public health policy decisions that might be made in response to an emerging pandemic. As such, it could lead to solutions not only for scientists, but also for public health officials who require quantitative feedback in order to develop appropriate policies and response plans.

email: kaufman@almaden.ibm.com

14. STATISTICAL LEADERSHIP UNDER PHRMA CRITICAL PATH INITIATIVES

EFFICIENCY OF LATE-STAGE CLINICAL RESEARCH (ECR)

Walter W. Offen*, Eli Lilly and Company
Joe Camardo, Wyeth Pharmaceuticals

The FDA's Critical Path document emphasizes the need for greater efficiencies that could preserve the rigor and reduce the cost, in all phases of clinical research. The pharmaceutical industry clearly embraces these same goals, and is enthusiastic in its support of the Critical Path Initiative. This presentation will focus on the recommendations from a PhRMA working group which has considered ideas to improve efficiency in late-stage clinical research. There are three key areas relating to efficiency to be discussed. One topic is how sponsors might generate additional interpretable safety data post-approval in a timely way, to lessen the inclination to increase the size, cost, and duration of Phase 3 beyond where they are today. The second topic includes several study design concepts that if accepted by FDA and utilized more broadly can lead to more efficient studies. These include multiple co-primary endpoints, non-inferiority, flexible dosing designs, dichotomization of continuous measures, longitudinal data analysis methods, and crossover designs. Finally, we briefly address technology and process enhancements that would have significant impact on study efficiency.

email: offen_walter_w@lilly.com

NOVEL ADAPTIVE CLINICAL TRIAL DESIGN

Brenda L. Gaydos*, Eli Lilly and Company
Michael Krams, Pfizer, Inc.

A design is adaptive when the option exists to use accumulating data to modify aspects of the trial. When adaptive by design, the adaptations are a design feature and require thorough up-front planning to maintain trial integrity and inferential validity. It is important to establish the appropriate scope for adaptive designs in clinical development. A PhRMA working group on adaptive designs was formed in the spring of 2005. One objective of this group is to move the debate on application from unqualified enthusiasm to a fact based evaluation of where such designs are best deployed. Some potential advantages of design flexibility include a greater likelihood of allocating patients within a trial to effective/safe therapies, improved learning about the research question, a better decision on sample size due to improved information on the design parameters, and more informed and possibly earlier development decisions. These advantages could contribute to a reduction in the late stage attrition rate, currently estimated at greater than 50%. However, adaptive designs provide a number of challenges. They are technically and operationally more complex than fixed designs, and more experience is needed to understand and address regulatory concerns. In this presentation, opportunities, challenges and recommendations for consideration will be provided.

email: blg@lilly.com

ROLLING DOSE STUDIES

José C. Pinheiro*, Novartis Pharmaceuticals
Rick Sax, Astrazeneca

Despite revolutionary improvements in basic biomedical science, the number of new drug applications submitted to the FDA has shown a declining trend over the past several years. Both the FDA and PhRMA have started initiatives to identify and address the main drivers leading to the pharmaceutical industry's current problems. One well-known such driver is poor dose selection resulting from incorrect or incomplete knowledge of the dose response relationship for both efficacy and safety. This talk will discuss an innovative class of dose finding designs, called Rolling Dose Studies (RDS), aimed at striking a balance between additional dose response information and increased costs/timelines. In these studies, the number of doses and/or the allocation of patients to doses are allowed to change during the study, as increasing efficacy and safety information becomes available. PhRMA has formed a working group on RDS whose work will be presented at this talk. Different types of RDS designs and methods, focusing on Phase II trials, will be described together with the results of a comprehensive simulation study evaluating the performance of the methods under a variety of trial scenarios. Recommendations on the practical use and potential gains associated with rolling dose studies will be discussed.

email: jose.pinheiro@novartis.com

15. THE ROLE OF NEW DESIGNS FOR EVALUATING VACCINES AND OTHER PREVENTION PROGRAMMES

STATISTICAL CHALLENGES IN COMBINING HUMAN WITH ANIMAL STUDIES: THE CASE OF ANTHRAX VACCINES

Donald B. Rubin*, Harvard University

There are at least three major statistical challenges in the CDC anthrax vaccine trials. The first is that the human trials have the inevitable missing data due to missed visits and missed measurements of immunogenicity and reactogenicity-- these could be called 'violations of data collection protocol'. The (multiple) imputation of these missing data is the first task, because doing so will allow standard intention-to-treat analyses to take place on the human trials, e.g., of reactogenicity outcomes. The second challenge is that there are treatment protocol violations --e.g., times when vaccinations were supposed to be received but were not. Here, the principled approach to the estimation of the actual causal effect of the vaccination on outcomes in humans is to use principal stratification, where the task is effectively to impute the protocol violations that would have been observed if the people were assigned treatments other than the ones that they were actually assigned -- not an easy problem in general, but not impossible either. The third statistical challenge is to use the animal (macaque) trials to build a model that will allow the imputation in the human trials of survival when challenged using the measurements of immunogenicity. The collection of immunogenicity data on both humans and macaques, coupled with the parallel design of these randomized trials, is planned to allow this imputation through assumptions concerning the ignorable assignment (by nature) of immunogenicity levels given the actual randomized assignment of vaccination levels and covariates. This presentation will outline these three tasks and present some details of the first one.

email: rubin@stat.harvard.edu

ROBUST ANALYSIS OF THERAPEUTIC HIV VACCINATION TRIALS

Devan V. Mehrotra*, Merck Research Laboratories
Robin Mogg, Merck Research Laboratories

For many HIV infected patients, use of antiretroviral therapy (ART) results in a sustained suppression of plasma viral load to undetectable levels. However, due to lack of antigenic stimulation, this may also result in a gradual loss of cell mediated immune (CMI) responses that help control HIV infection. In concept, augmenting ART with periodic administrations of an HIV vaccine that boosts CMI responses could enhance control of viral replication. In clinical trials being designed to test this hypothesis, HIV-infected patients with sustained viral suppression will receive an experimental HIV vaccine or a placebo, and subsequently stop taking their antiretroviral drugs. The goal is to assess whether the plasma viral loads during the ART interruption phase are generally lower in the vaccine group. Assessment of a vaccine effect will be challenging if some subjects resume ART or drop out before the scheduled end of the ART interruption phase. To tackle this "missing" data problem, we propose multiple imputation of the missing viral load values followed by use of the Wei-Lachin method. We use a numerical example and simulations to illustrate the robustness and power advantages of our proposed method relative to other methods like REML, weighted GEE, LOCF, and "worst-rank" methods.

email: devan_mehrotra@merck.com

AUGMENTED DESIGNS TO ASSESS IMMUNE RESPONSE IN VACCINE TRAILS

Dean A. Follmann*, NIAID

This talk introduces methods for use in vaccine clinical trials to help determine if the immune response to a vaccine is actually causing a reduction in the infection rate. This is not easy because immune response to the (say HIV) vaccine is only observed in the HIV vaccine arm. If we knew what the HIV-specific immune response in placebo recipients would have been, had they been vaccinated, this immune response could be treated essentially like a baseline covariate and an interaction with treatment could be evaluated. We introduce two methods for inferring this HIV-specific immune response. The first involves vaccinating everyone before baseline with an irrelevant vaccine, e.g. rabies. Randomization allows us to infer a placebo volunteer's response to the HIV vaccine using the rabies response and a prediction model estimated in the vaccine group. With the second method, all unifected placebo patients at closeout are vaccinated with the HIV vaccine. We pretend this post closeout HIV immune response was observed at baseline. A probit regression model with a treatment by HIV-immune response interaction is postulated and estimated using maximum likelihood. Simulations are used to evaluate the methods and practical issues are addressed.

email: dfollmann@Niaid.nih.gov

16. IMS: DIMENSION REDUCTION

DIMENSION REDUCTION: AN OVERVIEW

Bing Li*, Pennsylvania State University

This talk will give an overview of the recent development in sufficient dimension reduction. It will survey the fundamental ideas and challenging issues in the current research, including: the basic assumptions (such as elliptical symmetry) for dimension reduction and how they can be met or relaxed; examples and applications to illustrate the important role of dimension reduction in data analysis; dimension reduction with categorical predictors; dimension reduction for a parameter of interest; current research on the situations where both the predictor and the response are multidimensional; different methods of order determination for a dimension reduction space such as sequential tests, bootstrap, and BIC; general criteria to evaluate the exhaustiveness and accuracy of a dimension reduction method.

email: bing@stat.psu.edu

USING INTRA-SLICE INFORMATION FOR IMPROVED DIMENSION REDUCTION

Liqiang Ni*, University of Central Florida

Many methods for estimating the central subspace in regression require slicing a continuous response. However, slicing can result in loss of information and in some cases that loss can be substantial. We use intra-slice covariances to construct improved inference methods for the central subspace. These methods are optimal within a class of quadratic inference functions and permit chi-squared tests of conditional independence hypotheses involving the predictors.

email: lni@mail.ucf.edu

SUFFICIENT DIMENSION REDUCTION FOR THE SMALL-N-LARGE-P PROBLEMS

Lexin Li*, North Carolina State University
Dennis Cook, University of Minnesota

There has recently been a surge of interest in the analyses of large-scale, high-dimensional data problems. In particular, it is not uncommon to encounter problems with the number of predictors p greater than the number of observations n. Examples include DNA microarray data and quantitative genetics data. In this talk, we will propose a dimension reduction method that can handle small-n-large-p problems. We demonstrate that the proposed method can both effectively identify the active individual predictors, and extract the linear combinations of predictors that preserve all regression information. The method is based on an inverse regression formulation, and has been developed within the framework of sufficient dimension reduction.

email: li@stat.ncsu.edu

AN INTERACTIVE METHOD FOR SUFFICIENT DIMENSION REDUCTION

Francesca Chiaromonte*, Penn State University
Dennis Cook, University of Minnesota
Bing Li, Penn State University

Large-p-small-n data, in which the number of recorded variables exceeds the number of independent observational units, are becoming the norm in a variety of scientific fields. Sufficient dimension reduction provides a meaningful and theoretically motivated way to handle large-p-small-n regressions, by restricting attention to $d < n$ linear combinations of the original p predictors. However, standard sufficient dimension reduction techniques are themselves designed to work for $n > p$, because they rely on the inversion of the predictor sample covariance matrix. We propose an iterative method that eliminates the need for such inversion, using instead powers of the covariance matrix. We illustrate our method with a genomics application; the discrimination of human regulatory elements from a background of non-functional DNA, based on their alignment patterns with the genomes of other mammalian species. We also demonstrate the excellent performance of the iterative method by simulation. We speculate that powers of the covariance matrix may allow us to effectively exploit available information on the predictor structure in identifying directions relevant to the regression.

email: chiaro@stat.psu.edu

17. ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

IDENTIFYING EFFECT MODIFIERS IN AIR POLLUTION TIME-SERIES STUDIES USING A TWO-STAGE ANALYSIS

Sandrah P. Eckel*, Johns Hopkins University
Thomas A. Louis, Johns Hopkins University

Studies of the health effects of air pollution such as the National Morbidity and Mortality Air Pollution Study (NMMAPS) relate changes in daily pollution to daily deaths in a sample of cities and calendar years. Generally, location-specific estimates are combined over locations using a two-stage model. We build on this approach by "fractionating" the city-specific analysis to produce month/city/year specific estimated air pollution effects (slopes). We identify potential effect modifiers via regression and prediction trees with the estimated slopes as dependent variables and predictors such as dew point temperature, temperature, CO, NO2, O3, SO2, season, year and region. Our analysis relates single-day lagged PM10 to daily mortality in people over 65 from the 50 largest NMMAPS cities. Stage 1 consists of two log-linear Poisson regressions. The first produces an estimated overall city-specific effect of PM10 using city-specific NMMAPS models that include smooth functions of temperature, dew point temperature and time. The second regression estimates city-month-year specific PM10 effects using the daily predicted log-death rate from stage 1 as an offset. These city-month-year specific estimates are analyzed in Stage 2 using weighted linear regression and CART with month-city-year specific predictor variables. We report on our methods and findings.

email: seckel@jhsph.edu

SEMI-PARAMETRIC MODELING OF EFFECTS OF AIR QUALITY ON RESPIRATORY HEALTH IN CHICAGO MEDICAID POPULATION

Chava E. Zibman*, University of Chicago
Vanja Dukic, University of Chicago
Paul Rathouz, University of Chicago

Time-series analyses of the relationship between pollution and morbidity generally rely on Poisson regression models with smooth functions of time to account for unobserved time-varying confounders. This paper presents the analyses of daily time series asthma-prescription data over four summers in Chicago. Choice of the degree of smoothness for such functions is an open problem, usually addressed via model-selection criteria such as AIC or BIC. However, rather than choosing a pre-specified amount of smoothness in the adjustment for the confounding variable, we use Bayesian analysis to adjust for uncertainty due to the choice of smoothing parameter for each of the four years separately. In addition, given that the asthma-related outcomes (daily beta agonist prescription counts) are aggregated at the ZIP-code level, we include the neighborhood-specific effects to account for overdispersion. We find that the effects of ozone and PM_{10} are negligible at the relatively low levels of pollution observed in Chicago, but that there is interesting variation in the estimated year-specific smooth functions.

email: chava@uchicago.edu

BAYESIAN MODELING OF AIR PATTERN FOR TWO-ZONE FIELDS

Yufen Zhang*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Gurumurthy Ramachandran, University of Minnesota
Rui Yang, University of Minnesota

The two-zone model is popular in industrial hygiene community. However, little research has been carried out to systematically assess the performance of this model, particularly with regard to statistical estimation from experimental data. We propose a statistical framework for validating such models from properly designed workplace experiments. We adapt methods proposed for Bayesian non-linear regression and for fitting computer models to achieve statistical estimation. For optimizing computation, different interpolation schemes such as tensor-product interpolation and stochastic interpolators (Gaussian Processes) are employed as representations of the solution surface of the model. Tests on data from both simulation and a real experiment are adopted for evaluating our approach and formally assessing tenability of scientific hypothesis.

email: yufenz@biostat.umn.edu

USE OF GAMS TO ASSESS EFFECTS OF AIR POLLUTION ON HUMAN HEALTH: OUR ACAPS EXPERIENCE

Vincent C. Arena*, University of Pittsburgh
Ya-Hsiu Chuang, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh

The generalized additive model (GAM) is a useful analytical tool to assess the short-term health effects of air pollution using time series data. The modeling endeavor requires various decisions regarding the fitting of the GAM. We present a sensitivity analysis to assess the robustness of the risk estimate of the current day PM10 levels on daily cardio-pulmonary hospital admissions that was seen in the Alleghany County Air Pollution Study (ACAPS). The sensitivity analysis includes different lag models for the PM10, different composite measures that combine PM10 levels from multiple monitoring sites, different smoothing functions, and degrees of smoothing on weather and time covariates. Results from our sensitivity analysis indicate that current day PM10 level is related to an increase in daily cardio-pulmonary hospital admissions in a robust way.

email: yac14@pitt.edu

EXPERIMENTAL DESIGNS FOR EVALUATING THE EFFECTIVENESS OF REHABILITATION ACTIONS IN CREATING FISH HABITAT IN THE TRINITY RIVER

Darcy C. Pickard*, Simon Fraser University & ESSA Technologies

The Trinity River Restoration Project is a major effort at restoring salmonid habitat and populations. It is believed that if management actions (eg. mechanically reshaping the channel, changing flows, adding gravel) can stimulate and appropriately mimic natural river processes, then the river will create and maintain suitable habitat areas. We evaluate several experimental designs to determine which designs are most likely to detect a difference in the effectiveness of different mechanical rehabilitation actions. Each action is unique to the rehabilitation site and so cost is used as a simple measure of the complexity of a single mechanical action. Fry rearing habitat is believed to be the limiting type of habitat. This habitat is lost as the growth of riparian vegetation forms permanent berms along the river edge. We modeled the formation of berms using simple transition state matrices to describe the probability of vegetation surviving to the next year or being washed away. The model allows for different probabilities given different rehabilitation actions, flow regimes over time, effects due to channel form and dependence on upstream conditions. The performance of the alternative designs under different scenarios is compared and presented.

email: dcp@sfu.ca

BAYESIAN HIERARCHICAL MODELS IN NEST SURVIVAL STUDIES

Jing Cao*, Southern Methodist University
Chong He, Virginia Tech University

Recently, logistic nest survival models (Dinsmore, White, and Knopf 2002; Shaffer, 2004) have been developed to incorporate biological covariates with the assumption that the nest age on the first encounter of nest can be decided accurately. Also, the nest curve is assumed to be a parametric function (linear or quadratic) of nest age. In this paper, we propose a Bayesian hierarchical model with nest-specific covariates to estimate age-specific daily survival rates. The model can handle any mixture of irregular visiting schedules, and it allows a broad variety of covariates and competing models to be evaluated. With the least restrictive assumptions, the model does not require the knowledge of exact nest age when nest is first found. The typical features of nest survival data, truncation and censoring, are accounted for by the likelihood function and the latent variables. An intrinsic auto-regressive (IAR(2)) prior is employed for the nest age effect. This nonparametric prior provides a much more flexible and parsimonious alternative to the parametric specification. Last but not least, the Bayesian computation is efficient because the full conditional distributions either have closed forms or are log-concave. Finally, we present a simulation study and a analysis of a Missouri dickcissel dataset to illustrate the performance of the model.

email: jcao@smu.edu

BAYESIAN DISTRIBUTED LAG MODELS: ESTIMATING EFFECTS OF PARTICULATE MATTER AIR POLLUTION ON DAILY MORTALITY

Leah J. Welty*, Northwestern University
Scott L. Zeger, Johns Hopkins Bloomberg School of Public Health
Francesca Dominici, Johns Hopkins Bloomberg School of Public Health

A distributed lag model (DLM) is a regression model that includes lagged exposure variables as covariates; its corresponding distributed lag (DL) function describes the relationship between the lag and the coefficient of the lagged variable. DLMs are commonly used in environmental epidemiology for quantifying the cumulative effects of weather and air pollution on mortality and morbidity. Standard methods for forumlating DLMs include unconstrained, polynomial, and p-spline DLMs. These methods may fail to take full advantage of prior information about the shape of the DL function for environmental exposures, or for any other exposure with effects that are believed to smoothly approach zero as lag increases, and are at risk of producing sub-optimal estimates. We propose a Bayesian DLM (BDLM) that incorporates prior knowledge about the shape of the DL function and allows the degree of smoothness of the DL function to be estimated from the data. In a simulation study, we compare our Bayesian approach with standard methods. We also show that BDLMs encompass p-spline DLMs. We apply our BDLM to data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) to estimate the short term health effects of PM10 on mortality from 1987--2000 for Chicago, Illinois.

email: lwelty@northwestern.edu

SCREENING DESIGNS FOR DRUG DEVELOPMENT

Peter Mueller, M.D. Anderson Cancer Center
Gary Rosner, M.D. Anderson Cancer Center
David Rossell*, Rice University/M.D. Anderson Cancer Center

We propose drug screening designs based on a Bayesian decision theoretic approach. The discussion is motivated by screening designs for phase II studies. The proposed screening designs allow to consider multiple treatments simultaneously. In each period new treatments can arise, and currently considered treatments can be dropped from the active set. Once a treatment is removed from the phase II screening trial, a terminal decision is made about abandoning the treatment or recommending the treatment for a confirmatory future phase III study. The decision about dropping treatments from the active set is a sequential stopping decision. We propose a solution based on decision boundaries in the space of marginal posterior moments for the unknown parameter of interest for each treatment. We show a straightforward Monte Carlo simulation algorithm to implement the proposed approach. We compare our methodology with a two-stage design proposed by Yao et al. in a clinical immunology scenario and show that we can achieve a significant reduction in the average number of patients. We also apply it to the design of microarray experiments.

email: rusi@rice.edu

AN ADAPTIVE PHASE I DESIGN FOR IDENTIFYING A DOSE-OUTCOME REGION FOR TWO DRUG COMBINATIONS

Sumithra J. Mandrekar*, Mayo Clinic
Daniel J. Sargent, Mayo Clinic
Yue Cui, Mayo Clinic

Historically, designs for dose seeking trials using drug combinations have been geared towards finding the maximum tolerated dose, with safety as the primary outcome. With target based agents whose dose-efficacy curves are unknown and dose-toxicity relationships are expected to be minimal, alternative designs are needed. The present approach is a natural extension of the adaptive single agent dose-finding design previously presented (Zhang, Sargent, Mandrekar, SIM, 2005). A generalization of the continuation ratio model allowing separate toxicity and efficacy curves for each agent to generate a dose outcome surface for the combination is used. A continual reassessment method with straightforward dose selection criterion using accumulated data from all patients is employed. Our simulation studies demonstrated that the proposed design has favorable operating characteristics in terms of experimentation and recommendation rates, and the average sample size, under a variety of scenarios. We believe that this approach of incorporating toxicity and efficacy into the identification of an optimal combination region is novel and warrants further consideration.

email: mandrekar.sumithra@mayo.edu

OPTIMAL TWO-STAGE DESIGNS IN PHASE-II CLINICAL TRIALS

Anindita Banerjee*, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

Two-stage designs have been widely used in phase II clinical trials. Such designs are desirable because they allow a decision to be made on whether a treatment is effective or not after the collection of the data at the end of each stage. Optimal fixed two-stage designs, where the sample size at each stage is fixed in advance, were proposed by Simon (1989) when the primary outcome is a binary response. This paper proposes an adaptive two-stage design which allows the sample size at the second stage to depend on the results at the first stage. Using a Bayesian decision theoretic construct, we derive optimal adaptive two-stage designs, the optimality criterion being the minimum expected sample size under the null hypothesis. We also develop optimal designs for response probabilities other than the null. Comparisons are made between Simon's two-stage fixed design and the new design with respect to these optimality criteria.

email: abanerj2@ncsu.edu

DESIGNING COVARIATE ADJUSTED RESPONSE ADAPTIVE RANDOMIZED TRIALS IN THE PRESENCE OF COVARIATE BY TREATMENT INTERACTIONS

Ayanbola O. Ayanlowo*, University of Alabama at Birmingham
David T. Redden, University of Alabama at Birmingham

Adaptive designs in clinical trials try to capitalize on the accruing information throughout the trial by adjusting the probability of future treatment assignments. Two main approaches to adaptive randomized designs have been proposed: response adaptive randomized designs and covariate adaptive designs. Response adaptive randomized designs continually update the probability of assignment to a treatment arm based on data accrued about the treatment effect. Covariate adaptive randomized designs, which seek to ensure a balance between treatment arms with respect to a set of known covariates, randomly assign treatment conditional upon the subject's covariate values. Recently, the concept of covariate adjusted response adaptive randomized designs has been proposed. This method adjusts for both the covariate values and the subjects' responses to determine the treatment assignment probabilities. We propose studying the statistical power of covariate adjusted response adaptive randomized models in the presence of covariate by treatment interactions. We propose a method of selecting the optimal sample size required for the equal allocation scheme that usually precedes the covariate adjusted adaptive randomization scheme; such that adequate power is maintained at the end of the study. Simulations will be presented illustrating the properties of this method under different types of covariate by treatment interactions.

email: Titlade@yahoo.com

AN ADAPTIVE DESIGN IN A DOSE-FINDING STUDY FOR THE ACUTE TREATMENT OF MIGRAINE

Vladimir Dragalin*, GlaxoSmithKline

An adaptive treatment allocation rule was chosen to achieve trial's objectives with a comparatively low number of patients and minimizing exposure of patients to nonefficacious doses. The goal of the study is to identify the minimal effective dose (MED) of a drug that is superior to placebo and establish a dose-response relationship. While the adaptive allocation rule is targeting the MED, a forced randomization to placebo and the highest dose is used to build an early stopping rule for futility and to provide information about the entire dose-response curve. At the end of the study, a four parameter logistic regression is used to fit the data.

email: Vladimir.2.Dragalin@gsk.com

ADAPTIVE TREATMENT ALLOCATION WITH CONTINUOUS COVARIATES: A COMPARISON OF METHODS

Nora J. Graber*, Rho, Inc.

Pocock and Simon (1975) has been widely used as a method of adaptive treatment allocation for sequential clinical trials that require balance over several prognostic factors. The limitation of Pocock and Simon's method, however, lies in the fact that all covariates must be categorical. In contrast to the Pocock and Simon procedure, Frane's method (1998) provides an alternative design that can be used when balance is desired on continuous and/or categorical factors. By use of simulations, Pocock and Simon's and Frane's method of adaptive treatment allocation were compared using both categorical and continuous covariates. Continuous variables were transformed into categorical variables by choosing cut-points to dichotomize the continuous variables, thereby allowing Pocock and Simon's method to be used. The statistical performance of the two methods of allocation were compared through a series of simulation studies. It was concluded that Frane's method is superior to Pocock and Simon's method when there is little knowledge about the data. If Pocock and Simon's method must be used, it is best to choose a cut-point close to the median of the data. The simulation study revealed that, whichever allocation method is chosen, it is recommended to use the continuous variables in the analysis.

email: ngraber@rhoworld.com

BAYESIAN DOSE-FINDING DESIGNS BASED ON A NEW STATISTICAL FRAMEWORK FOR PHASE I CLINICAL TRIALS

Yuan Ji*, University of Texas M.D. Anderson Cancer Center

Phase I clinical trials aim to find the maximum tolerated dose of an experimental drug. We consider dose escalation, de-escalation or staying at the current dose as three different stochastic moves over the lattice of a sequence of prespecified dose levels. Each move is chosen by minimizing an expected penalty that determines the dose level for treating the next cohort of patients. We develop a stopping rule under which the termination of the trial ensures that the posterior probability that the current dose is the maximum tolerated dose is larger than a prespecified value. Under a new class of priors, posterior estimates for the dose toxicity probabilities are obtained using the Markov chain Monte Carlo method. We demonstrate the new designs using a real phase I clinical trial.

email: yuanji@mdanderson.org

19. STATISTICAL AND COMPUTATIONAL METHODS FOR GENETIC DATA

RELATEDNESS ESTIMATION FOR STRUCTURED POPULATIONS

Amanda B. Hepler*, North Carolina State University
Bruce S. Weir, University of Washington

The amount of relatedness between two individuals has been widely studied across many scientific disciplines. Currently, a number of relatedness estimation methods exist, however few appropriately account for structured populations. We present a comparison of three maximum likelihood techniques for the estimation of various multidimensional parameters commonly used to measure pairwise relatedness. Two of the three methods will allow for structured populations. Simulation studies compare the three techniques, demonstrating the bias that can occur for estimators not accounting for structured populations. Empirical data sets are also used to compare the methods, using human samples of both microsatellites and single nucleotide polymorphisms. Three novel methods to classify familial relationships are also presented and compared.

email: abhepler@stat.ncsu.edu

COALESCENT ANALYSIS OF MODELING MUTATION PROGRESSION IN COLORECTAL CANCER

Hui Zhao*, M.D. Anderson Cancer Center
Qingyi Wei, M.D. Anderson Cancer Center
Yun-Xin Fu, University of Texas-Houston Health Science Center, School of Public Health

Colorectal cancer is the 3rd most common diagnosed cancer in the United States. Most of the hereditary nonpolyposis colorectal cancer (HNPCC) and 15% of sporadic colorectal cancer show microsatellite instability (MSI). Colorectal cancer starts from a mutation in a normal colorectal cell and grows into a clone of cells that further accumulate mutations and finally develop into a malignant tumor. In molecular evolution terms, the process of colorectal tumor evolution represents the acquisition of sequential mutations. Clinic studies use biomarkers such as microsatellite or single nucleotide polymorphism (SNP) to estimates the mutation frequencies in colorectal cancer. Microsatellite data obtained from single genome equivalent PCR or small pool PCR can be used to infer tumor evolution. Since tumor evolution is similar to population evolution, we use well-established coalescent theory to analyze this type of data. The purpose of our study is to develop a coalescent framework to simulate the colorectal cancer mutation progression. Our simulation frameworks are based on coalescent theory and the nature history of colorectal tumor progress to trace the evolutionary process. Our result indicates that the cell population growth path and total number of mutations in a genealogy contribute to variation in the tumor phenotype.

email: hzhao2@uth.tmc.edu

CONTEXT DEPENDENT MODELS FOR DISCOVERY OF TRANSCRIPTION FACTOR BINDING SITES

Chuancai Wang*, Penn State College of Medicine
Jun Xie, Purdue University
Bruce A. Craig, Purdue University

Transcription factors play a crucial role in gene regulation, and the identification of transcription factor binding sites helps gain insight into gene regulatory mechanisms. The overall goal of this work is to describe a new method of binding site detection called Motif Discovery via Context Dependent Models (MDCDM). We characterize the motif (i.e., binding sites) by a series of position-dependent first-order Markov models. This model considers both the position-specific features of the motif and the dependence between positions of the motif. In addition, a "step-up" testing procedure is used to automatically determine the best-fitting Markov model for the background (i.e., nonsite regions). We compare our approach with the existing methods using both real and simulated data sets. The results show that the detection of binding sites can be greatly improved by accounting for dependence across positions in a motif and appropriately modeling the background dependence.

email: cwang@hes.hmc.psu.edu

INCORPORATING MEDICAL INTERVENTIONS INTO MENDELIAN MUTATION PREDICTION MODELS

Hormuzd A. Katki*, Johns Hopkins Bloomberg School of Public Health and Division of Cancer Epidemiology and Genetics, NCI, NIH, DHHS

People with family history of disease consult with genetic counselors about having mutations that increase disease risk. To aid them, genetic counselors use models to predict if the patient has mutations given their family history of disease. For example, the BRCAPRO model gives the risk of carrying mutations in the BRCA genes based on family history of breast and ovarian cancer. The carrier probability is used to decide whether to undertake genetic testing or subsequent medical intervention. For example, oophorectomy prevents breast and ovarian cancers. Oophorectomy is becoming common among high-risk women, like BRCA mutation carriers, and more women are reporting family histories with oophorectomies. But no mutation prediction model accounts for medical interventions. Ignoring medical interventions can seriously underestimate the mutation carrier probability. We show how to incorporate medical interventions and describe the data required to compute the extra quantities. We propose two assumptions that minimize the amount of required extra quantities. We apply this to incorporating oophorectomy into BRCAPRO. This update to BRCAPRO will be available to counselors for clinical use.

email: katkih@mail.nih.gov

IDENTIFYING NOVEL NF-KB-REGULATED IMMUNE GENES IN THE HUMAN GENOME USING STRUCTURED SUPPORT VECTOR MACHINE WITH DISCRETE KERNEL

Insuk Sohn*, Korea University, Seoul, Korea
Sujong Kim, Skin Research Institute, AmorePacific Corporation R&D Center
Jae Won Lee, Korea University-Seoul, Korea

Identifying novel NF-kB-responsive immune genes in the human genome is important to understand immune functions and immune diseases. For this purpose, we proposed Structured Support Vector Machine (SSVM) with discrete kernel. We applied the proposed discrete SSVM method to the promoters of 62 known NF-kB-regulated immune genes, to find patterns of transcription factor binding sites in the promoters of these known genes. Using these patterns, we examined the promoters of 6440 additional genes to find matches to the patterns. Our method of predicting gene function, based on characteristic patterns of transcription factor binding sites, would be applicable to finding genes with other functions.

email: sis46@korea.ac.kr

A WEIGHTED REGRESSION MODEL FOR A GLOBAL TEST OF HAPLOTYPE EFFECTS IN CASE-CONTROL SAMPLES

Anbupalam Thalamuthu*, University of Pittsburgh
Daniel E. Weeks, University of Pittsburgh

We propose a method for testing association of haplotypes for case-control samples. In a sample of unrelated individuals haplotypes inferred using statistical methods involves phase uncertainty. Modeling haplotype effects for case-control data should account for this phase uncertainty. Here we describe a weighted regression model for effect of two haplotypes carried by an individual on the trait, the weights being the posterior probabilities of possible haplotypes given the marker genotypes. A Bernoulli likelihood is constructed by modeling disease penetrances through a product of two logit transformations, one for each of the haplotypes carried by an individual. Number of parameters in this model is equal to the number of SNPs used for haplotype construction and hence it uses much less d.f. for a global test of haplotype effects as compared to other existing methods in the literature. Power and type-I error of the proposed test are examined using simulated data sets. The statistical power of the method proposed here is compared with two other existing methods.

email: anbupalam.thalamuthu@hgen.pitt.edu

COMPARING THE JOINT DISTRIBUTION OF MULTIPLE CATEGORICAL VARIABLES BETWEEN TWO GROUPS: WITH APPLICATION TO ANALYSIS OF PRE/POST HIV-1 GENOTYPE SEQUENCES

Greg Di Rienzo*, Harvard School of Public Health

Consider a recent HIV-1 clinical trial where an HIV-1 genotype sequence was recorded for each subject at baseline, and, for those subjects who subsequently failed the study regimen, at the time of virological failure. Consider the scientific problem of identifying changes in HIV-1 genotype sequence that are associated with the risk of virological failure. To provide a statistically sound solution, a two-stage resampling-based algorithm is proposed that, for arbitrary data-generating mechanisms, asymptotically controls the associated false-positive rate at any pre-specified level. The first stage estimates the components of the joint distribution of HIV-1 genotype sequence that differ between pre/post measurements. The second stage attempts to eliminate from redundancies from this set, in the sense that the difference can be completely explained by differences with respect to a smaller number of variable/level combinations. In addition to a detailed analysis of this clinical trial, a simulation study is presented that evaluates the methodology for both the paired- and independent-sample case.

email: dirienzo@hsph.harvard.edu

BIVARIATE RANDOM EFFECT MODEL USING SKEW NORMAL DISTRIBUTION WITH APPLICATION TO HIV-RNA

Pulak Ghosh*, Georgia State University
Marcia D. Branco, University of São Paulo
Hrishikesh Chakraborty, RTI International, North Carolina

Correlated data arise in a longitudinal studies from epidemiological and clinical research. Random effects models are commonly used to model correlated data. Mostly in the longitudinal data setting we assume that the random effects and within subject errors are normally distributed. However, the normality assumption may not always give robust results, particularly if the data exhibit skewness. In this paper, we develop a Bayesian approach to bivariate mixed model and relax the normality assumption by using a multivariate skew-normal distribution. Specifically, we compare various potential models and illustrate the procedure using a real data set from HIV study.

email: pghosh@mathstat.gsu.edu

GENERALIZED MONOTONIC FUNCTIONAL MIXED MODELS FOR THE EFFECTS OF RADIATION DOSE HISTOGRAMS ON NORMAL TISSUE COMPLICATIONS

Matthew J. Schipper*, University of Michigan
Jeremy M.G. Taylor, University of Michigan
Xihong Lin, Harvard University

Normal tissue complications are a common side effect of radiation therapy. While the tumor site generally receives a homogenous dose of radiation, the normal tissue surrounding the tumor does not. Good estimates of the dose distribution to the normal tissue of interest can be obtained, but it is not known what function of the dose distribution drives the presence and severity of the complications. Regarding the density of the dose distribution as a curve, a summary measure is obtained by integrating a weighting function of dose (w(d)) over the dose density. We propose a generalized monotone functional mixed model which relates the dose distribution to the complication using this summary measure. For biological reasons the weight function should be monotonic. In our model w(d) is written as an integral of a smooth positive function of d. We 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.

email: mjschipp@umich.edu

SEMIPARAMETRIC APPROACH FOR THE MISALIGNED MEASUREMENTS IN COLON CARCINOGENESIS STUDY

Zonghui Hu*, National Institute of Health
Naisyin Wang, Texas A&M University

We study a mixed-effects model where the main response and covariate variables are linked through the positions where they are measured. However, they are not measured at the same positions or the same sub-units due to technical limitations. A semiparametric approach is proposed for this problem of misaligned measurements, where the existence of an unobserved latent covariate is assumed. Asymptotic properties of the semiparametric estimator are derived, and the regularity conditions for the asymptotic performance are discussed. In the end, we present its application in a colon carcinogenesis study.

email: huzo@niaid.nih.gov

EMPIRICAL BAYES LINEAR MIXED MODEL ANALYSES FOR TWO-COLOR MICROARRAY EXPERIMENTS

Lan Xiao*, Michigan State University
Robert J. Tempelman, Michigan State University

Empirical Bayes strategies based on combining information across genes has been deemed useful for stable and powerful inference in microarray experiments given the large number of genes and small sample sizes for each gene; however, these strategies have primarily been oriented towards shrinkage estimation of gene-specific residual variances. The analysis of microarray data in efficient incomplete block or split plot designs for two color systems is complicated by multiple sources of variability as suitablye accomodated with a mixed model analysis. We evaluate mixed model shrinkage procedures for two commonly used designs (common reference and loop design) based on two approaches to variance component estimation: REML and ANOVA. Our simulation studies indicated the shrinkage based on ANOVA rather than REML estimation of variance components leads to a superior ROC curve for number of true positives versus number of false positives for differential expression. We further compare the two methods using two public gene expression data sets. Keywords: Empirical Bayes; cDNA microarray; mixed model; ANOVA component; ANOVA; REML; differential expression.

email: xiaolan@msu.edu

STATISTICAL ANALYSIS OF DENDRITIC BRANCHING IN HIPPOCAMPAL NEURONS

Rebecka J. Jornsten*, Rutgers University

The analysis of neuron morphology is geared toward increasing our understanding of dendritic branching and developing mechanisms. Neurological disorders such as autism and Rett's syndrome are associated with disrupted or altered branching patterns. Understanding which factors affect branching and how they do so is the first step toward developing drugs that can target these disruptions directly. We analyze branching patterns of neurons from a series of experiments via hierarchical generalized linear mixed models, and mixtures of hierarhical generalized linear mixed models. The data consists of primary dendrites (stemming from the cell body) and their first level of branching into secondary dendrites. We jointly analyze which factors affect the number of primary dendrites, the proportions of primaries that branch, and how many secondaries result from a branching event. The presence and analysis of mixtures in branching data is novel and has not been analyzed in the literature to date. We determine at what levels of branching a mixture component is defined in order to pinpoint what distinct branching patterns are associated with these sub-populations of neurons. The outcome of these analyses enabled us to formulate novel biological hypotheses that are currently being validated in the lab.

email: rebecka@stat.rutgers.edu

21. SEMIPARAMETRIC AND NONPARAMETRIC MODELING

A NONPARAMETRIC ESTIMATE OF THE CUMULATIVE INCIDENCE FUNCTION UNDER TIME-DEPENDENT TREATMENT ASSIGNMENTS

Chung-Chou H. Chang*, University of Pittsburgh
Wei Tian, Inspire Pharmaceuticals, Inc.

We propose a nonparametric method to estimate the cumulative incidence function (CIF) when the assignment of a particular type of treatment to each patient is time-dependent. Our method is a generalization of the Nelson-Aalen estimator of the CIF. We also incorporate the inverse probability treatment weights into the proposed estimator for a study with unbalanced distribution of confounders. The CIF derived from our method can be used to estimate the survival benefits of a particular type of treatment. Liver transplantation will be used as an example.

email: changj@pitt.edu

EFFICIENT ESTIMATION OF POPULATION-LEVEL SUMMARIES IN GENERAL SEMIPARAMETRIC REGRESSION MODELS WITH MISSING RESPONSE

Arnab Maity*, Texas A&M University
Yanyuan Ma, Texas A&M University
Raymond J. Carroll, Texas A&M University

This paper considers a wide class of semiparametric regression models with responses missing at random. Special cases in this approach include generalized partially linear models, generalized partially linear single index models, structural measurement error models and many others. For these problems, profile likelihood kernel estimation methods are well-established in the literature. Here our focus is on estimating general population-level quantities, e.g., population mean, quantiles, probabilities, etc. We derive the asymptotic distributions of estimates of these population-level quantities, showing that in many cases the estimates are semiparametric efficient. For estimating the population mean with no missing data, we show that the sample mean is semiparametric efficient for canonical exponential families, but not in general. We apply the methods to a problem in nutritional epidemiology, where estimating the distribution of usual intake is of primary interest, and semiparametric methods are not available.

email: amaity@stat.tamu.edu

A NOVEL APPROACH TO TESTING EQUALITY OF SURVIVAL DISTRIBUTIONS WHEN THE POPULATION MEMBERSHIP INFORMATION IS CENSORED

Dipankar Bandyopadhyay*, University of Georgia
Somnath Datta, University of Louisville

This paper introduces a novel nonparametric approach for testing the equality of two or more survival distributions when the population membership information are not available for the right censored individuals. Although such data structures arise in practice very often, this problem has received less than satisfactory treatment in the nonparametric testing literature. Currently there is no nonparametric test for this hypothesis in its full generality in the presence of right censored data. We propose to use the imputed population membership for the censored observations leading to fractional weights that can be used with a two sample censored data test. A class of weighted log-rank tests thus obtained this way is studied through simulation. We also obtain an asymptotic linear representation of our test statistic leading to its asymptotic distribution and propose two resampling alternatives which might be easier to use in practice. Our testing methodology is illustrated using two real data sets.

email: dban@stat.uga.edu

MEASURING LATERAL CONTROL IN DRIVING STUDIES

Jeffrey D. Dawson*, University of Iowa
Joseph E. Cavanaugh, University of Iowa
K.D. Zamba, University of Iowa
Matthew Rizzo, University of Iowa

In driving studies using simulators or instrumented vehicles, information may be recorded at high frequencies (e.g., 10-60 frames/second). High variability in the lateral control of the vehicle may result in an increased likelihood of lane crossings and accidents. Boer (2000) proposed a method of quantifying variability in steering wheel positioning, which can also be applied to the lateral position of a vehicle within the driving lane. In his approach, data from three preceding adjacent intervals are used to predict the value at each time point, and the discrepancies between the predicted and observed values are used to define a baseline distribution of prediction errors within a subject. This distribution is then used as a reference for calculating a summary metric (termed 'entropy') in follow-up epochs of interest, such as when a driver may be distracted when using a cell phone. We illustrate this method with data from a driving simulator known as SIREN, and make analytical comparisons between this 'entropy' metric and other uses of the prediction error distributions. We also perform computer simulation experiments under a threshold time-series model to compare Boer's methods to several modifications thereof, as well as to some simple measures of variability.

email: jeffrey-dawson@uiowa.edu

A GEOMETRIC APPROACH TO ESTIMATION OF THE NUMBER OF SPECIES

Changxuan Mao*, University of California, Riverside

Estimating the number of species in a population from a sample of individuals is investigated in a Poisson mixture model. A sequence of estimators are proposed from a geometric perspective. Simulation is used to assess their performance and a genomic application is studied as an illustration.

email: cmao@stat.ucr.edu

ESTIMATION OF THE MEAN FUNCTION OF PANEL COUNT DATA USING MONOTONE POLYNOMIAL SPLINES

Minggen Lu*, University of Iowa
Ying Zhang, University of Iowa
Jian Huang, University of Iowa

We study the nonparametric pseudo-likelihood and full-likelihood estimators of the mean function of a counting process based on panel count data using monotone polynomial splines. The setting for panel count data is one in which n independent subjects, each with a counting process with common mean function, are observed at several possibly different times during a study. Generalized Rosen algorithm was used to compute the estimators. We show the proposed spline estimators are asymptotically consistent and the rate of convergence is higher than 1/3. The simulation study show the spline-based estimators have smaller variances and mean square errors than nonparametric pseudo and full likelihood estimators proposed in Wellner and Zhang (2000). A real bladder cancer trial example is used to illustrate the method.

email: minggen-lu@uiowa.edu

NONPARAMETRIC ECOLOGICAL INFERENCE: INCORPORATING MARGINAL COVARIATE INFORMATION IN A NONPARAMETRIC REGRESSION MODEL FOR AGGREGATE DATA

Joan G. Staniswalis*, University of Texas at El Paso

A nonparametric regression model is considered for ecological inference or analysis of survey data reported for various locations in aggregate form. A complication is that only marginal information is available on the response and covariates, that is, the subject responses are not linked to the database containing the covariate information. A general form for the regression model is proposed, whereby certain covariates are included as in a varying-coefficient regression model, while others are included as in a functional linear model, depending upon whether the covariate is categorical or continuous. A nonparametric regression is computed by penalized weighted least squares. The pointwise maximum squared bias is derived using O'Sullivan (1986), which when taken together with the pointwise standard errors allows for interpretation beyond that possible in just a graphical exploration of a nonparametric fit. This paper demonstrates that existing databases can be used to explore factors that might be associated with health disparities in disease outcomes among different subgroups of the United States population. Estimates for age-adjusted prevalence rate of End Stage Renal Disease (ESRD) by ethnicity are computed using 2000 US Census and 1998 US ESRD Network data for Texas counties.

email: joan@math.utep.edu

22. SPATIAL MODELING

A COMPOSITE LIKELIHOOD CROSS-VALIDATION APPROACH IN SELECTING BANDWIDTH FOR THE ESTIMATION OF THE PAIR CORRELATION FUNCTION

Yongtao Guan*, University of Miami

A useful tool while analyzing spatial point patterns is the pair correlation function. In practice, this function is often estimated by some nonparametric procedure such as kernel smoothing, where the smoothing parameter (i.e.,bandwidth) is often determined arbitrarily. In this article, a data-driven method for the selection of the bandwidth is proposed. The efficacy of the proposed approach is studied through both simulations and an application to a forest example.

email: yguan@miami.edu

IMPROVED DETECTION OF DIFFERENTIALLY EXPRESSED GENES THROUGH INCORPORATION OF GENE LOCATIONS

Guanghua Xiao*, University of Minnesota
Cavan Reilly, University of Minnesota
Betsy M. Martinez-Vaz, University of Minnesota
Wei Pan, University of Minnesota
Arkady Khodursky, University of Minnesota

In determining differential expression in cDNA microarray experiments, the expression level of an individual gene is usually assumed to be independent of the expression level of other genes, but many recent studies have shown that a gene's expression level tends to be similar to that of its neighbors on a chromosome, and differentially expressed genes are likely to form clusters of similar transcriptional activity along the chromosome. When modelled as a one-dimensional spatial series, the expression levels of genes on the same chromosome frequently are spatially correlated, indicating spatial patterns in transcription. Based on these spatial correlations, we can obtain more adequate estimates of gene expression by utilizing the information about gene location. Using the autocorrelation function, we demonstrated the existence of spatial correlations of transcriptional activity in the Escherichia coli chromosome. We proposed a hierarchical Bayesian model that borrows information from neighboring genes to improve the estimation of the transcription level of a given gene and hence the detection of differentially expressed genes. Both, simulation studies and the analysis of experimental data showed that the proposed method outperforms the SAM t statistic, which is widely used in detecting differentially expressed genes.

email: guanghx@biostat.umn.edu

HIERARCHICAL AND JOINT SITE-EDGE METHODS FOR AREAL BOUNDARY ANALYSIS

Haijun Ma*, University of Minnesota
Bradley P. Carlin, University of Minnesota
Sudipto Banerjee, University of Minnesota

Spatial boundary analysis is an important topic in public health research. Unfortunately, few boundary analysis methods for areal (lattice) data exist, even though most public health data are of this type. In this paper, we offer a variety of novel hierarchical models for areal boundary analysis that parameterize both the areas and the edge segments either hierarchically or jointly, leading to conceptually appealing solution that remains computationally feasible. While our approaches borrow from similar developments in statistical image restoration using Markov random fields, important differences arise due to the irregular nature of our lattices, the occasional existence of important covariate information, and most importantly, our desire for full posterior inference on the boundary. We illustrate our methods using both artificial and real data, the latter relating to the problem of determining the service area of a particular cancer hospice system in northeastern Minnesota based only on Medicare billing records.

email: haijunma@biostat.umn.edu

THE EFFECT OF AGGREGATION ON INFERENCES USING SMALL AREA HEALTH DATA

Sandy Burden*, University of Wollongong
David G. Steel, University of Wollongong

Within the public health system, emphasis on preventative health has focussed attention on the risk factors contributing to the diseases affecting society. Spatial variation in population characteristics and environmental exposures aids in improving this understanding. However, despite its widespread use, small area health data suffers from several limitations including the aggregate nature of much of the data and spatial dependence between data points. Aggregation using arbitrary areal boundaries limits the resolution of analyses, whilst changes in area size and location can alter model estimates obtained (the so called modifiable area unit problem - MAUP) resulting in ecological bias. Hence, results can only legitimately be applied to the particular areal units used. To investigate the scale and zoning effects of the MAUP in the context of epidemiological analysis, simulated data is generated at the individual level and then analysed at successively aggregated levels. The simulations help to understand the relative value of aggregate data analysis for spatial epidemiological investigations.

email: sburden@uow.edu.au

STATISTICAL COMPARISON OF OBSERVED AND MULTI-RESOLUTION CMAQ MODELED OZONE CONCENTRATIONS

Li Chen*, University of Chicago
Michael L. Stein, University of Chicago

Community Multi-scale Air Quality (CMAQ) modeling system has been running at different spatial resolutions, e.g., 36 km, 12 km and 4 km. This paper compares CMAQ modeled ozone concentrations at different spatial resolutions with observations. The result shows that higher resolution CMAQ model output does not necessarily provide better prediction with smaller root mean square error (RMSE) than the lower ones. Aggregation is performed to obtain new versions of lower resolution model output based on the higher resolution model output. By doing so, the aggregated lower resolution model output predicts better than others in terms of RMSE. Variation decomposition is used as a tool to understand the statistical behavior of CMAQ model outputs at different resolutions. It helps us to get more accurate predictions.

email: lichen@uchicago.edu

PARAMETERIZATION OF SPATIAL MODELS AND STABILITY OF ESTIMATES

Petruta C. Caragea*, Iowa State University
Mark S. Kaiser, Iowa State University
KyojiFurukawa, Iowa State University

Gaussian models for discrete index random fields are common in applications, in part because the marginal means appear in the usual parameterization of conditional distributions. This facilitates, among other things, the incorporation of covariates into regression models of mean structure. In models having other distributional forms (e.g., Binomial, Poisson) this property does not occur under traditional parameterizations. It is possible, however, to specify such models in a form such that indiviudal parameters are nearly equal to marginal means. Indications are that these parameterizations also lead to greater stability in estimated means across varying degrees of spatial dependence, although this has not been clearly demonstrated for mean structures that depend on covariate information. We investigate the stability of estimated regression parameters in such situations, and consider implications for applications.

email: pcaragea@iastate.edu

MODELING THE EVOLUTION OF AN AIR-BORNE CONTAMINANT RELEASE IN AN URBAN ENVIRONMENT

Margaret B. Short*, Los Alamos National Laboratory

Imagine a puff release of a particulate contaminant in an urban setting with complicated air flow patterns. In order to predict the evolution of such a release, we place monitors at a few strategic locations and collect concentration data. True concentrations are usually zero; and non-zero concentrations may fall below the threshold for detection of the monitoring equipment. Existing methods for modeling concentration data with a detection threshold make unconfirmable assumptions about the behavior of the contaminant when the concentration is below the threshold, whether by imputing an explicit value for the concentration, or by treating those concentrations as parameters to be estimated. These assumptions are particularly inappropriate in our setting, where data are naturally and effectively taken on the log scale. We propose a method that eliminates these assumptions and provides an answer that's defensible and not prone to computational instabilities. We use a process convolution approach to implement our model, and perform calculations by using Markov chain Monte Carlo. We explore the suitability of the proposed model by applying it to output from a simulator that implements a physics based computer model for the transport of particles. This is joint work with Nick Hengartner and Steve Thompson.

email: mshort.zz01@gmail.com

23. STATISTICAL ISSUES IN GENETIC INVESTIGATIONS

CONTROLLING FALSE DISCOVERIES AND FALSE NON-DISCOVERIES IN MICROARRAY ANALYSIS

Sanat K. Sarkar*, Temple University

Multiple hypothesis testing is an important statistical tool in identifying differentially expressed genes in microarray analysis. Several new approaches for controlling false discoveries and false non-discoveries have been developed since Benjamini and Hochberg (1995, Journal of the Royal Statistical Society, Ser. B) first introduced the concept of False Discovery Rate (FDR). These different concepts, and methods controlling them, will be reviewed in this talk.

email: sanat@temple.edu

CHANGING EXPRESSIONS: THE EVOLUTION OF INFORMATION TECHNOLOGY APPLIED TO GENE EXPRESSION

Daniel J. Holder*, Merck Research Laboratories

Modern technological advances have enabled scientists to query biological systems at a molecular level and produce a great variety and abundance of data. Statistics and other informational technologies are aimed at transforming these data into useful information. Although progress has been made, techniques for analysis and interpretation have found it hard to keep pace with data generation. As an example, we will examine the evolving set of techniques used to analyze gene expression. We will suggest that, despite early enthusiasm for the contrary, traditional statistical concepts such as variance components and experimental design are no less appropriate for microarrays then they are for other biochemical assays. However, the exploratory nature of these experiments, together with their complexity and overwhelming number of variables relative to independent samples force us to modify our thinking and lead to the adoption of an updated set of concepts and techniques. From a practitioner's point of view, we will examine some of the techniques we find useful for quantification, quality assurance, differential expression, prediction, and interpretation in gene expression experiments.

email: dan_holder@merck.com

APPROACHES TO ANALYSIS OF NON-GAUSSIAN CLUSTERED DATA FROM GENETIC ANIMAL STUDIES

Inna Chervoneva*, Thomas Jefferson University

Modern genetic animal studies, especially ones utilizing monitoring or image-processing equipment, yield a fairly large number of continuous repeated measures per animal, while the number of animals is small or moderate due to difficulties in breeding genetically altered strains. Furthermore, often multiple samples from different locations are taken, introducing two or more additional levels of clustering. Such data are conventionally analyzed in the framework of the linear mixed effects models when normality assumptions are appropriate. We consider analysis approaches when conditional on random effects distributions do not support the normality assumption and may vary substantially in distributional shape. We show that such approaches allow accommodating non-Gaussian conditional distributions and can provide additional useful information about characteristics of distributions, other than location parameter, that may be of interest to researchers.

email: i_chervoneva@mail.jci.tju.edu

24. STATISTICAL MODELS IN MICROPARTICLE REMEDIATION/DECONTAMINATION

EXPERIMENTAL VALIDATION OF CONTAMINANT CONCENTRATIONS PREDICTED BY A DETERMINISTIC MODEL

Myron J. Katzoff*, National Center for Health Statistics/CDC
Abera Wouhib, National Center for Health Statistics/CDC
Stanley A. Shulman, National Institute for Occupational Safety and Health
JamesS. Bennett, National Institute for Occupational Safety and Health
William K. Sieber, National Institute for Occupational Safety and Health

This talk will be about a methodology for validating a computational fluid dynamics (CFD) model which is expected to be useful in refining adaptive sampling procedures applicable to microparticle removal. Current field sampling practices are inherently adaptive so that the addition of statistical formalism is expected to enable greater efficiency in the use of field resources and valid statistical inferences. The methodology for CFD model validation that we describe will employ statistical techniques used in the frequency domain analysis of spatio-temporal data. The validation problem is posed in a manner intended to be suggestive of a signal detection problem in which a statistical test for a known signal is to be performed. Suggestions are made for some follow-up analyses for addressing situations if the testing procedure rejects the null hypothesis.

email: mjk5@cdc.gov

ESTIMATING TRACER GAS DISTRIBUTION IN A VENTILATION CHAMBER

JamesS. Bennett, National Institute for Occupational Safety and Health
Stanley A. Shulman*, National Institute for Occupational Safety and Health
W. Karl Sieber, National Institute for Occupational Safety and Health
Myron Katzoff, National Center for Health Statistics
Abera Wouhib, National Center for Health Statistics
Brian Adams, South Dakota School of Mines

Tracer gas concentrations in an empty chamber provide data for comparison with computational fluid dynamic (CFD) model predictions for a relatively simple scenario. Experimental considerations were: a) instruments, limited to two, which recorded average concentrations every second; b) trial duration, with 15-minute trials originally run; c) approximate stationarity of measurements collected at each location; d) locations chosen to have large between-location concentration range; e) day-to-day variation. Early experiments showed some instrumental bias. There was evidence of nonstationarity, but no clear trends, and trial duration could be shortened to 4.5 minutes. Lognormality of concentration measurements allowed estimation of geometric means. Further experiments indicated instrument bias did not vary by location. Locations were selected with relatively wide range in geometric means. Location geometric means were estimated and compared, after adjusting for instrument bias and allowing for day-to-day variation. Confidence intervals for ratios of geometric means by location were calculated.

email: sas2@cdc.gov

A COMPARISON OF ROOM CONTAMINATION FIELDS ESTIMATED VIA KRIGING AND DETERMINISTIC AIR FLOW MODELS

James S. Bennett*, National Institute for Occupational Safety and Health
Sean A. McKenna, Sandia National Laboratories
Patrick D. Finley, Sandia National Laboratories
Stanley A. Shulman, National Institute for Occupational Safety and Health
W. Karl Sieber, National Institute for Occupational Safety and Health
Myron Katzoff, National Center for Health Statistics
Abera Wouhib, National Center for Health Statistics
John E. Brockman, Sandia National Laboratories
Richard O. Griffith, Sandia National Laboratories

Tools for estimating contamination fields include probabilistic, data-driven methods such as kriging and deterministic, physical process based approaches such as computational fluid dynamics (CFD). The work presented compares room contaminant field predictions from both methods. The scenario is an office-sized experimental room in the ventilation laboratory. A CFD model that was validated with tracer gas measurements in room air is used to predict aerosol surface concentrations on the floor, through a Lagrangian particle transport model and the assumption that, when particles impact the floor, they stick. The probabilistic approach then treats this field as the population from which a limited number of samples are selected. Different sample sets are used as input to the kriging algorithm and the resulting prediction of the detailed field is compared with the detailed field determined by CFD. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL850000.

email: jbennett@cdc.gov

25. INFERENCE IN RANDOMIZED MULTI-CENTER CLINICAL TRIALS

CONDITIONING ON THE SAMPLE SPACE: A METHOD TO ADJUST FOR LARGE NUMBERS OF INSTITUTIONS WITHOUT INTRODUCING PARAMETERS

Lu Zheng*, Harvard School of Public Health
Marvin Zelen, Harvard School of Public Health

We distinguish between ”design based” versus ”model based” analyses of planned experiments. A design based analysis incorporates the main features of the planned experiment as the principal basis for making inferences. A model based analysis may ignore some features of the planned experiment and use models such as proportional hazards, logistic and linear regression. Our philosophy is that all inferences should be based on design based analyses. The model based analyses are only appropriate if they are close approximations to the design based analyses. An important class of planned experiments is the multi-center randomized clinical trial. A design based analysis would rely on the permutation distribution generated by the randomization process. Ordinarily the number of patients assigned to each treatment within a center is a random variable, but is also an ancillary statistic. Another feature of multi-center randomized trials is the use of permuted blocks to allocate the treatments. The permuted blocks also generate ancillary statistics. An important principal in frequentist inference is to condition on the ancillary statistics as the conditioning will reduce the sample space ordinarily resulting in greater power. Finding the exact distribution of the appropriate test statistic under these circumstances may be difficult, if not impossible. As a result we have developed an approximation to this distribution. Simulations show that the approximation works well. We have investigated the power when the outcomes are continuous, binary, and censored in the context of multi-center trials with variation between institutions. Our investigations indicate that there is an increase in power, conditioning on the ancillary statistics, compared to ignoring the ancillary statistics for the three types of outcome data. The increase in power may be considerable if there is large variation between institutions with respect to end points. The methods have been extended to group sequential trials with similar increases in power. The analyses described here are distribution free, result in an increase in power and are not difficult to carry out.

email: lzheng@hsph.harvard.edu

DESIGN VS MODEL BASED ANALYSIS

John M. Lachin*, The George Washington University

The process of randomization of treatments to subjects in a clinical trial provides a probabilistic structure for the observed result that in turn can lead to a randomization or permutation test for the equivalence of the treatments. Such tests are rarely used for the analysis of such trials. Rather the analysis is routinely conducted using the more prevalent 'population' model-based tests that are based on the concept of sampling observations from a general population. This presentation will review some of the considerations that enter into the general preference for population model-based tests, considerations that could be viewed as obstacles to the use of a randomization-based test as the basis for an inference. These include difficulties with the construct and interpretation of confidence limits on population parameters, the assessment of sample size and power a priori, the ability to 'adjust' for other covariates, and in some cases, the lack of a known large sample permutational distribution for the test statistic under specific allocation schemes, such as 'minimization' or other covariate adaptive randomization schemes, as well as response adaptive schemes such as variations on 'play the winner'. These and other obstacles will be discussed and suggestions offered.

email: jml@biostat.bsc.gwu.edu

LOCAL VS GLOBAL INFERENCE

Marvin Zelen*, Harvard University

Consider a multi-center randomized clinical trial. A local inference is defined as conclusions that only apply to the patients who entered the clinical trial; i.e. best treatment for the patients in the trial. A global inference is defined as conclusions that apply to the population with disease; i.e. best treatment for those with disease. Under what circumstances do the different inferences apply? Patients and centers may be separately envisioned to be a random sample from a population of patients and centers or if they are not random samples they are each considered to be a collection of patients and centers. Consequently the recruitment process generates four models for making inferences; i.e. patients and centers may either be collections or random samples. In practice, patients and centers are each collections. As a result inferences in a trial, strictly speaking, are local inferences that may even differ for each center. Alternatively if the patients and centers are both random samples, the inference is a global one. The variance for comparing the average of two therapies is smallest for the local inference and largest for the global inference. Consequently the power for making a global inference will be less than making a local inference.

email: zelen@hsph.harvard.edu

26. ILS: INTRODUCTION TO LONGITUDINAL DATA

INTRODUCTION TO LONGITUDINAL DATA

Marie Davidian*, North Carolina State University

Studies in which a response is ascertained intermittently over time (longitudinally) on each of a number of 'individuals' (humans, animals, agricultural plots, etc.) are commonplace in many areas of application. Often, the scientific questions of interest in such longitudinal studies focus on how the pattern of response changes over time or on underlying characteristics of the individuals responsible for changing patterns, and on interrelationships between these phenomena and attributes of the individuals and other factors. In order to address such questions reliably, a statistical model framework is required that allows the questions to be stated formally and that faithfully represents variation among and within individuals. The past several decades have seen fundamental advances in the development of such statistical models and accompanying inferential methods. This session will provide an introduction to popular longitudinal data models and methods and of popular statistical software in SAS and Splus/R for their implementation. It will be assumed that audience is familiar with linear and generalized linear regression models and standard techniques for their implementation at an applied level but has had no prior exposure to longitudinal data models and methods.

email: davidian@stat.ncsu.edu

27. IMS: RECENT ADVANCES IN MIXTURE MODELS

SEMIPARAMETRIC ANALYSIS IN CONDITIONAL INDEPENDENCE LATENT CLASS MODELS

Jing Qin*, National Institute of Allergy and Infectious Disease, NIH
Denis Leung, Singapore Management University

For k (k>= 3) variate data drawn from a mixture of two distributions, each having independent components, Hall and Zhou (2003) developed a nonparametric method for identifying the underlying distributions and mixing proportions. Under the additionl assumption that the components are identically distributed, Cruz-Medina et. al. (2004) suggested a semi-parametric method by reducing the problem into a mixture of of multinomial distributions. However, both of these approaches have very lower efficiencies. In this talk, we consider a semiparametric method for estimation in a trivariate mixture model. We use the exponential tilt model of Anderson (1979), in which the log ratio of probability (density) functions from the trivariate components is assumed to be quadratic in the observations. Apart from the exponential tilt assumption, the method does not require training samples. The empirical likelihood technique is used for estimation and testing problems. We show that the maximum empirical likelihood estimate has an asymptotic normal distribution. Furthermore, the empirical likelihood ratio statistic behaves like a chi-squared variable. We use simulations to study the method's small sample behavior and we apply the method in a set of data in developmental psychology.

email: jingqin@niaid.nih.gov

CLUSTERING BASED ON A MULTI-LAYER MIXTURE MODEL

Jia Li*, The Pennsylvania State University

In model-based clustering, the density of each cluster is usually assumed to be a certain basic parametric distribution, e.g., the normal distribution. In practice, it is often difficult to decide which parametric distribution is suitable to characterize a cluster, especially for multivariate data. Moreover, the densities of individual clusters may be multi-modal themselves, and therefore cannot be accurately modeled by basic parametric distributions. We explore in this paper a clustering approach that models each cluster by a mixture of normals. The resulting overall model is a multi-layer mixture of normals. Algorithms to estimate the model and perform clustering are developed based on the classification maximum likelihood (CML) and mixture maximum likelihood (MML) criteria. BIC and ICL-BIC are examined for choosing the number of normal components per cluster. Experiments on both simulated and real data are presented.

email: jiali@stat.psu.edu

GENERALIZING HODGES-LEHMANN: NONPARAMETRIC INFERENCE FOR LOCATION MIXTURES

David R. Hunter*, The Pennsylvania State University

The well-known Hodges-Lehmann estimator gives a nonparametric estimate of center when observations are assumed to come from a symmetric distribution. We extend this estimator to the case of finite location mixtures of a symmetric distribution. Because the class of symmetric distributions is so broad, identifiability is a major issue in these mixtures. We discuss identifiability, then give a general distance-based estimation method. When the distance involved is L_2-distance between distribution functions, the resulting estimators may be shown to generalize the Hodges-Lehmann estimator. We focus particular attention on the two-component case, where this method provides an interesting complement to more traditional parametric approaches.

email: dhunter@stat.psu.edu

28. COPULA AND COX REGRESSION MODELS

SENSITIVITY ANALYSIS FOR SHARED PARAMETER MODELS USING COPULAS

Dimitris Rizopoulos*, Catholic University of Leuven
Geert Verbeke, Catholic University of Leuven
Emmanuel Lesaffre, Catholic University of Leuven

One of the major challenges statisticians face in the analysis of longitudinal data is the problem of missing data. Even though in most longitudinal studies individuals are scheduled to be assessed at a common set of prespecified visit times after enrollment, they often selectively miss their visits or appear at nonscheduled points in time. This situation gives rise to two competing processes; a longitudinal process that measures the quantity of main scientific interest, and a survival process that describes either the time to dropout or the time elapsed between adjacent visits. The form of the association between these two processes is very important in the analysis of longitudinal data. In particular, if the survival process is related to the unobserved longitudinal outcomes, then biased inferences might result from an analysis that ignores this association. Several methods for the joint modelling of longitudinal and time to event data have been proposed in the literature. These models, known as shared parameter models, rely mainly on an extensive use of Gaussian random-effects and therefore induce elliptical types of dependence between the two processes. However, the resulting inferences can be sensitive to the distributional assumptions for the random-effects, in the sense that these assumptions cannot be easily checked from the available data. To assess the effect of the assumed dependence structure between the two processes, we propose a flexible representation of the random-effects distribution using copulas. This formulation enables the consideration of various types of association structure between the two processes by only changing the copula function for the underlying random-effects, while keeping all the other aspects of the model fixed. Thus, the proposed method can be considered as a flexible tool for sensitivity analysis in shared parameter models. Our approach is applied in a longitudinal study, with 24.2% dropout, for the comparison of two oral treatments for dermatophyte toe onychomycosis.

email: dimitris.rizopoulos@med.kuleuven.be

MODELING BIVARIATE SURVIVAL TIMES BY COPULAS

Rui Qin*, The University of Iowa
Michael P. Jones, The University of Iowa

Copula models with relative risk margins are proposed for bivariate survival time data. They are extensions of the classical marginal approach of using Cox regression to model the marginal survival functions by the addition of an association parameter.The parameter is treated as a measure of correlation between the bivariate survival times. The comprehensive family of Archimedean copulas provides flexibility in modeling different correlations. Estimation occurs in two stages. First, Cox regression is used to estimate the marginal survival functions. Second, a pseudo-likelihood for the association parameter is constructed by plugging in the marginal estimators and is then maximized over the association parameter. Empirical process theory is applied to establish consistency and asymptotic normality of the two-stage estimator. Simulation is used to study the behavior of the estimator and the asymptotic conclusions in small to medium size samples. The Frank copula exhibits some advantages over the popular Clayton copula. An example with a real data set is given.

email: rui-qin@uiowa.edu

AN ANALOG PARAMETER ESTIMATOR FOR COPULA MODELS

Antai Wang*, Georgetown University
David Oakes, University of Rochester

We propose an analogue parameter estimator of the dependence parameter in archimedean copula models for bivariate censored data. Our new method is based on solving a self-consistency equation of the unknown parameter. We derive the variance formula for our estimation. A simulation study has shown that our estimators are easy to compute and quite robust.

email: aw94@georgetown.edu

MULTI-DIMENSIONAL COPULA REGRESSION MODELS FOR CORRELATED MIXED/CENSORED OUTCOMES

Mingyao Li*, University of Pennsylvania
Peter X.K. Song, University of Waterloo-Canada

We present a class of multi-dimensional regression models for correlated mixed or censored outcomes that arise often from many practical studies. As a special case, the proposed models reduce to a class of multi-dimensional generalized linear models when the marginal distributions are of the same type. The pivotal technique employed in our method is the Gaussian copula, which enables us to assemble different marginal regression models into a joint modeling framework. We obtain parameter estimation and make statistical inferences using maximum likelihood. Our method provides a unified likelihood framework for regression analysis of multivariate correlated data. It also brings much ease into numerical implementation and hence great potential for efficiently handling of complex data structures.

email: myli@umich.edu

ESTIMATION OF SURVIVAL FUNCTIONS AND COVARIATE EFFECTS BASED ON AN ASSUMED COPULA ACCOUNTING FOR DEPENDENT AND INDEPENDENT CENSORING

Xuelin Huang*, The University of Texas-MD Anderson Cancer Center
Gretchen A. Fix, Rice University
Katherine B. Ensor, Rice University

In many survival analysis settings, the event of interest is subject to multiple types of censoring. While some of them, such as administrative censoring, can be reasonably assumed to be independent censoring, others are clearly associated with the event of interest. Oftentimes knowledge about the direction and degree of this association is available. We consider the estimation in this situation of the marginal survival function of the event of interest and the effects of covariates. Our method deals with trivariate data. This data structure allows for an event of interest and both dependent and independent censoring events. An assumed copula and inverse probability of censoring weighted methods are used. Simulation studies, including sensitivity analyses, are conducted to evaluate the proposed method. This method is motivated by and applied to a finance study of dividend initiation behavior for publicly traded firms. The negative association between bankruptcy and dividend initiation is accounted for.

email: xlhuang@mdanderson.org

NEW APPROACH TO DIRECTIONAL DEPENDENCE USING COPULA FUNCTION

Yoonsung Jung*, Kansas State University
Jong-Min Kim, University of Minnesota-Morris
Engin A. Sungur, University of Minnesota-Morris

Directional dependence can be defined and studied from different perspectives. Recently, Sungur (2005) proposed the concept of directional dependence in bivariate regression setting by using copulas. In this research, we incorporate financial data to the proposed directional dependence method proposed by Sungur (2005). Furthermore, we develop the copula directional dependence method in light of the financial perspective.

email: ysjung72@hotmail.com

INFERENCES OF CHANGE POINT IN PIECEWISE COX MODEL

Zhiying Xu*, Case Western Reserve University
Pingfu Fu, Case Western Reserve University

Cox's (1972) proportional hazards regression model has become a popular tool in survival analysis. However, non-proportional hazards are very common results in clinical studies. In such cases, the proportional hazards models are not applicable. A piecewise Cox model, which assumes proportional hazards in a series of consecutive time intervals, may be one of good alternative choices for non-proportional hazards models. While using the piecewise Cox model, we are facing the difficulty on how to determine the most suitable change points. In this paper we present a computational method to estimate the change points in piece-wise Cox model efficiently. Jackknife and Bootstrap resampling techniques are introduced to estimate the variance and the confidence interval for the estimated change points. In addition, we applied this method to an autologous hematopoietic stem cell transplant study.

email: zhiying_xu@hotmail.com

29. BIOMARKERS AND SURROGATE MARKERS

SURROGATE MARKER VALIDATION FROM AN INFORMATION THEORY PERSPECTIVE

Ariel A. Abad*, Hasselt University
Geert Molenberghs, Hasselt University

The last twenty years have seen a large amount of work in the area of surrogate marker validation. Part of this recent work proposes to undertake the validation exercise in a multi-trial framework which leads to a definition of validity in terms of the quality of both trial-level and individual-level association between a potential surrogate and a true endpoint (Buyse 2000). However, a drawback of this methodology is that different settings have led to different definitions to quantify the association at the individual-level. In the present work, we use an information-theoretic method to create a unified philosophical approach to the surrogate marker evaluation problem. Based on concepts of information theory we propose a new definition of surrogacy with an appealing intuitive interpretation. This approach offers interpretational advantages and due to its generality can be applied in a wide range of situations. It also provides a better insight in the chances of finding a good surrogate endpoint in a given situation. Additionally, we show that some of the previous proposals in the literature to study surrogacy in different settings, can be seen as special cases of this general information-theoretic approach. We illustrate the use of our methodology using data from a clinical study in ophthalmology.

email: ariel.alonso@uhasselt.be

COMBINING LOGISTIC REGRESSION MODELS FOR MULTIPLE BIOMARKERS

Zheng Yuan*, University of Michigan
Debashis Ghosh, University of Michigan

In medical practice, there is great interest in developing methods for combining biomarkers in order to predict future clinical outcome or optimize classification accuracy. We argue that consideration of selection of markers should also be considered in the process. In this talk, we propose a novel model combining algorithm for classification in biomarker studies. The algorithm works by considering weighted combinations of various logistic regression models; six different weighting schemes are considered. A decision-theoretic framework that justifies certain weights is also developed. Simulation studies are performed for the proposed model combining method and compared to standard approaches. The method is also illustrated with application to data from a tissue microarray study in prostate cancer.

email: yuanz@umich.edu

JOINT ANALYSIS OF MULTIPLE LONGITUDINAL BIOMARKERS AND TUMOR COUNT DATA

Yulin Zhang*, University of Wisconsin-Madison
KyungMann Kim, University of Wisconsin-Madison

In cancer chemoprevention clinical trials, the primary objective is to assess the efficacy of the chemopreventive agent. With the clinical endpoint being cancer incidence, which requires large sample size and long time period and thus high cost, intermediate endpoints are under investigation for their potential as surrogate markers. Joint models of longitudinal biomarkers and cancer incidence are expected to improve inference on treatment effect. Joint models of longitudinal biomarkers and the time to event have been well developed in the setting of HIV/AIDS clinical trials (Wulfsohn and Tsiatis, 1997). Lin et al. (2002) proposed latent class models for jointly modeling a longitudinal biomarker and a censored survival outcome. We will present latent class models for joint analysis of multiple longitudinal biomarkers and tumor count data. EM algorithms are used to perform parametric and semi-parametric maximum likelihood estimation of parameters. The models are applied to data from a Skin Cancer Chemoprevention Trial with alpha-difluoromethylornithine (DFMO).

email: yulin@stat.wisc.edu

ON COMBINING DIAGNOSTIC MARKERS

Ruth Pfeiffer, National Cancer Institute
Efstathia Bura*, George Washington University

A popular summary measure of the discriminatory ability of a single continuous diagnostic marker for binary disease outcomes is the receiver-operator characteristics curve (ROC). For most diseases however, single biomarkers do not have adequate sensitivity or specificity for practical purposes. We present an approach to combine several markers into a composite diagnostic test without assuming a model for the conditional distribution predictors given the response. Using sufficient dimension reduction techniques, we replace the predictor vector with a lower-dimensional version, obtained through linear transformations of biomarkers, without loss of information. We show how to combine the linear transformations into a scalar diagnostic score whose performance can be assessed by the ROC curve. In the special case that a single linear combination of the markers contains sufficient information for the outcome, this approach results in the same marker combination obtained by Su and Liu (1993) that maximises the area under the ROC curve. An asymptotic chi-squared test for assessing individual biomarker contribution to the diagnostic score function is derived.

email: ebura@gwu.edu

THE USE OF SURROGATE MARKERS ON EARLY TREATMENT COMPARISON IN A META-ANALYSIS FRAMEWORK

Yun Li*, University of Michigan
Jeremy M.G. Taylor, University of Michigan

key words: Clinical trials; Meta-analysis; Prediction; Surrogate markers. Surrogate markers are an attractive option in clinical trials to predict the treatment effect on the true clinical endpoints, particularly when the true endpoint is expensive to collect. We examine the use of surrogate markers in a meta-analytic framework, in which the true and surrogate endpoints are available for previous trials and the true endpoints are observed for a fraction of subjects in the new trial. We study the efficiency gains from using the surrogate, of the estimated treatment effect in the new trial, as a function of the proportion of true endpoints observed, the correlation between two endpoints at both individual level and trial level, and other parameters of interest. In contrast to the situation where no true endpoints are measured we find there is considerable potential for gain in efficiency with high within-trial correlation with only a small fraction of the true endpoints observed.

email: yunlisph@umich.edu

MODIFICATION OF CONVENTIONAL EDGE DETECTORS FOR SEGMENTATION OF SPOTTED MICROARRAY IMAGES

Jingran Sun*, Amgen Inc.
Peihua Qiu, University of Minnesota

Segmentation of spotted microarray images is crucial in generating gene expression data. It can be regarded as a special case of edge detection in image processing. However, for generating gene expression data, segmentation methods are required to be able to classify pixels as either foreground or background pixels; and, most conventional edge detectors in the image processing literature do not have this property. In this paper, we propose a general procedure for modifying these edge detectors, such that the modified edge detectors have the property of pixel classification, as mentioned above. Numerical studies show that they perform well in applications of spotted microarray image segmentation.

email: jingrans@amgen.com

MICE: MULTIPLE-PEAK IDENTIFICATION, CHARACTERIZATION AND ESTIMATION

Nicoleta Serban*, Georgia Institute of Technology

MICE -- Multiple-peak Identification, Characterization and Estimation -- is a general procedure for estimating a lower bound for the number of components and for estimating the component parameters in a additive regression model. The method consists of a series of steps: a preliminary step for separating the signal from the background, identification of local maxima up to a noise level-dependent threshold, parameter estimation using an iterative algorithm, and detection of mixtures of components using hypothesis testing. The leading example is a nuclear magnetic resonance (NMR) experiment for protein structure determination. After fourier transform of NMR signals, NMR frequency data are multiple-peak data, where each peak corresponds to one component in the additive regression model. In this example, the primary objective is accurate estimation of the location parameters.

email: nserban@isye.gatech.edu

MIXED MEMBERSHIP STOCHASTIC BLOCK MODELS FOR RELATIONAL DATA WITH APPLICATION TO PROTEIN-PROTEIN INTERACTIONS

Edoardo M. Airoldi*, Carnegie Mellon University
David M. Blei, Carnegie Mellon University
Stephen E. Fienberg, Carnegie Mellon University
Eric P. Xing, Carnegie Mellon University

Modeling relational data is an important problem for modern data analysis and machine learning. In this paper we propose a Bayesian model that uses a hierarchy of probabilistic assumptions about the way objects interact with one another in order to learn latent groups, their typical interaction patterns, and the degree of membership of objects to groups. Our model explains the data using a small set of parameters that can be reliably estimated with an efficient inference algorithm. In our approach, the set of probabilistic assumptions may be tailored to a specific application domain in order to incorporate intuitions and/or semantics of interest. We demonstrate our methods on simulated data, where they outperform spectral clustering techniques, and we apply our model to a data set of protein-to-protein interactions, to reveal proteins' diverse functional roles.

email: eairoldi@cs.cmu.edu

A FRAMEWORK FOR KERNEL REGULARIZATION WITH APPLICATION TO PROTEIN CLUSTERING

Fan Lu*, University of Wisconsin-Madison
Sunduz Keles, University of Wisconsin-Madison
Stephen J. Wright, University of Wisconsin-Madison
Grace Wahba, University of Wisconsin-Madison

We develop and apply a previously undescribed framework that is designed to extract information in the form of a positive definite kernel matrix from possibly crude, noisy, incomplete, inconsistent dissimilarity information between pairs of objects, obtainable in a variety of contexts. Any positive definite kernel defines a consistent set of distances, and the fitted kernel provides a set of coordinates in Euclidean space that attempts to respect the information available while controlling for complexity of the kernel. The resulting set of coordinates is highly appropriate for visualization and as input to classification and clustering algorithms. The framework is formulated in terms of a class of optimization problems that can be solved efficiently by using modern convex cone programming software. The power of the method is illustrated not only via simulation study but also in the context of protein clustering based on primary sequence data. An application to the globin family of proteins resulted in a readily visualizable 3D sequence space of globins, where several subfamilies and subgroupings consistent with the literature were easily identifiable.

email: flu@wisc.edu

COMPARISON OF CLASSIFICATION METHODS TO PREDICT COMPLICATIONS TO LIVER SURGERY

Leah Ben-Porat*, Memorial Sloan Kettering Cancer Center
Mithat Gonen, Memorial Sloan Kettering Cancer Center
William Jarnigan, Memorial Sloan Kettering Cancer Center

Often in medical practice, it is of interest to identify patients that are at high risk for morbidity or mortality. As a result, prediction models are useful tools in medical decision making. There are many statistical methods available to build prediction models. The aim of this study is to compare several prediction methods using a large institutional database of patients undergoing liver surgery. Graphical representations of the models were constructed so as to make the models user friendly for clinicians. The database under study includes 2002 consecutive patients who underwent liver resection at Memorial Sloan Kettering Cancer Center from 1991 to 2002 and captures information on more than thirty preoperative risk variables. The classification models were built to predict high grade complications following surgery. The strategies employed were logistic regression, generalized additive models, and classification trees. All models were developed on a training set and evaluated on the test set. The performance of the models was compared by analyzing the ROC curve from the test set. The stepwise logistic regression model had similar predictive accuracy to the generalized additive model and the classification trees. Other non-linear machine learning methods such as neural networks will be applied to this dataset to determine their predictive power.

email: benporal@mskcc.org

MODEL-BASED PROJECTION PURSUIT CLUSTERING

Jie Ding*, GlaxoSmithKline

Model-based clustering is a comprehensive clustering method that assumes the data is sampled from finite mixture probability distribution models and clusters the data by maximizing the likelihood function from (mainly Gaussian) mixture model. It has been shown that many heuristic clustering methods are approximate estimation methods for certain mixture models. But model-based clustering performs less well when the dimension of the data becomes large due to the increase in the number of parameters to estimate. We introduce a model-based projection pursuit clustering method, which combines model-based clustering with dimension reduction technique - projection pursuit. The projection pursuit indices are defined as the function of the projection of the data and the clustering on the data. The analytical and numerical optimization of the index functions are obtained. The author applied this new method to various difficult clustering problems including the clustering of gene expression data when the number of cases N is usually less than the dimension p. The model-based projection pursuit clustering shows consistent improved results when compared to projection pursuit clustering or model-based clustering, individually. In addition, it also provides a low-dimensional (2-d or 3-d) pictorial representation of the clustering.

email: dingjienk@yahoo.com

MULTINOMIAL GROUP TESTING MODEL WITH SMALL-SIZED POOLS AND APPLICATION TO CALIFORNIA HIV DATA: BAYESIAN AND BOOTSTRAP APPROACHES

Jong-Min Kim*, University of Minnesota-Morris
Tae-Young Heo, Electronics & Telecommunications Research Institute-South Korea

This research consider multinomial group testing which is concerned with classification each of N given units into one of k disjoint categoires. In this research, we propose exact Bayesian, approximate Bayesian, bootstrap methods for estimating individual category proportions using the multinomial group testing model proposed by Bar-Lev et al (2005). By the comparison of Mean Squre Error (MSE), it is shown that the exact Bayesian method has a better efficiency and consistency than maximum likelihood method. We suggest an approximate Bayesian approach using Markov Chain Monte Carlo (MCMC) for posterior computation. We derive exact credible intervals based on the exact Bayesian estimators and present confidence intervals using the bootstrap and MCMC. These intervals are shown to often have better coverage properties and similar mean lengths to maximum likelihood method already available. Furthermore,the proposed models are illustrated using data from a HIV blooding test study throughout California, 2000.

email: jongmink@morris.umn.edu

31. SURVEY DATA AND SAMPLING METHODS

THE JOB OUTLOOK FOR BIOSTATISTICS GRADUATES

Joseph L. Hagan*, Biostatistics Program LSU HSC School of Public Health
Stephen W. Looney, Biostatistics Program LSU HSC School of Public Health

In this presentation, we consider the problem of estimating the annual number of masters and doctoral biostatistics graduates in the United States. We also consider estimation of the number of job opportunities for these graduates. We describe and compare several methodologies used to produce these estimates. We discuss the inherent difficulties in estimating the number of entry level jobs available. We also compare our methodology to that of a similar study conducted in 1994. We conclude that the job outlook is good for biostatistics graduates, especially for those obtaining a Ph.D..

email: joe.hagan@cox.net

COMPUTING INCLUSION PROBABILITIES FOR CONSTRUCTING HORVITZ-THOMPSON ESTIMATORS OF SAMPLING PLANS EXCLUDING NEIGHBORING UNITS

Kyoungah See*, Miami University
Robert Noble, Miami University
A. John Bailer, Miami University

Following the relative efficiency comparison of Sampling Plans Excluding Certain Neighboring Units (See, Noble, Bailer, 2005), we now compute the inclusion probabilities for Horvitz-Thompson estimators of the population mean, and their variances for various sampling plans that exclude certain neighboring units(SENU: See, Stufken, Song, Bailer, 2000). Sampling Plans Excluding Neighboring Units(SENU) type sampling plans can be thought of as a two-dimensional case of balanced sampling excluding contiguous units (BSEC: Hedaya, Rao, and Stuken, 1988). For the SENU-type sampling plans we consider, we provide a strategy for (i) estimating the first-order and second-order inclusion probabilities, and (ii) an approximation to the Horvitz-Thompson estimators of the mean and their variances. Illustrations are provided for 5 by 5 and 10 by 10 grids. For the 5 by 5 grid, an exact solution was obtained from an exhaustive search for the first and second order probabilities and therefore, the Horvitz-Thompson estimators and their variances are complete. For the 10 by 10 grid, we obtain the inclusion probabilities based on the simulations and estimate the Horvitz-Thompson estimators and their variances.

email: seek@muohio.edu

ESTIMATING THE DISTRIBUTION FUNCTION USING K-TUPLE RANKED SET SAMPLES

Kaushik Ghosh*, George Washington University
Ram C. Tiwari, National Cancer Institute

In this article, we consider k-tuple ranked set sampling and investigate estimation of the distribution function based on such data. The optimal choice of k is determined and it is shown that the procedure results in improved estimators over ordinary ranked set sampling (RSS) when n >= 3. The method is further generalized to the unbalanced case. A special case of the latter is extreme ranked set sampling and is seen to be more efficient over simple random sample of the same size when set size is less than 4. Results of simulation studies as well as an application to a real data set are presented to illustrate some of the theoretical findings.

email: ghosh@gwu.edu

COVARIATE ADJUSTMENT MAY NOT BE BETTER: THRESHOLDS OF RELATIVE RISK FOR REDUCTIONS IN MSE

Wenjun Li*, University of Massachusetts
Edward J. Stanek III, University of Massachusetts

In theory, covariate-adjusted rates have smaller mean squared errors (MSE) with known variance components. In practice, variance components are unknown and their sample estimators are used instead. We develop guidelines to determine when adjusted rates have smaller MSEs than crude rates using sample estimates of cigarette smoking rates as an example. We simulated a series of populations and samples, with male-to-female ratios from 0.25 to 4, male smoking rates from 40 to 68%, and relative risk of smoking (RR = male rate / female rate) ranging from 1 to 5.67. From each population and sampling plan, sampling was simulated 10,000 times. Adjusted rates were estimated using the sample covariance between smoking and gender, and population variance of gender. The ratios of MSEs of the adjusted rates to MSEs of the crude rates were computed, and the thresholds of RRs, above which adjusted rates have smaller expected MSEs, were estimated graphically. The MSE reductions due to covariate adjustment depend on sample sizes, gender ratios and RRs. In populations with balanced gender ratios, the adjusted rates had smaller MSEs when RRs were above 1.6, 1.4, 1.3 and 1.2 for sample sizes of 25, 50, 100 and 200, respectively. In populations with unbalanced gender ratios, the RR thresholds were higher. In sum, adjusted rates should not be used in all settings, and in particular, not when both RRs and sample sizes are small.

email: wenjun.li@umassmed.edu

MODELLING RARE EVENTS USING GENERALIZED INVERSE SAMPLING SCHEME

Soumi Lahiri*, New Jersey Institute of Technology
Sunil K. Dhar, New Jersey Institute of Technology

Generalized inverse sampling scheme is used when the investigator is interested in the rare events of the population. Samples are drawn until a predetermined number of rare events are observed. Log-linear model can be used for multi-way contingency tables under generalized inverse sampling scheme. It has a broad application to the research of biological and environmental sciences. The frequency counts in this case follow an extended negative multinomial distribution. The model parameters are estimated by maximum likelihood method. A procedure to test linear constraints of the regression parameters for the log-linear model is also derived. The results are illustrated by an example. Other stopping rules for the sampling procedures along with their applications will also be discussed.

email: sl28@njit.edu

WEIGHTED PROPORTIONAL HAZARDS MODELS FOR BIASED SAMPLES WITH ESTIMATED WEIGHTS

Qing Pan*, University of Michigan
Douglas E. Schaubel, University of Michigan

In biomedical studies using the Cox proportional hazards model, the observed data often constitute a biased sample of the underlying target population. The bias can be removed by weighting included subjects by the inverse of their selection probabilities, as proposed by Horvitz and Thompson (1952) and extended to the proportional hazards setting for use in surveys by Binder (1992) and Lin (2000). The weights can be treated as fixed in the cases where they are known or based on voluminous data(e.g. large-scale survey). However, in many practical applications, the weights are estimated and must be treated as such in order for the resulting inference to be accurate. We propose a weighted proportional hazards model in which weights are first estimated through a logistic model fitted to a representative sample from the target population. A weighted Cox model is then fitted to the biased sample. We propose estimators for the regression parameters and baseline hazard. Asymptotic distributions of the parameter estimators are derived, accounting for the additional variance introduced by the randomness of the weights. Our method is illustrated in an analysis of renal transplant patients from the Scientific Registry of Transplant Recipients (SRTR).

email: qingpan@umich.edu

PREDICTING REALIZED CLUSTER MEANS IN UNEQUALLY SIZED CLUSTER POPULATIONS

Edward J. Stanek III*, University of Massachusetts
Julio M. Singer, University of Sao Paulo-Brazil

Best linear unbiased predictors (BLUPs) are commonly added to estimates of fixed effects to predict the expected response of sampled clusters, such as schools or clinics. The predictors are usually based on mixed models for conceptual populations or superpopulations. Such methods account neither for an unbalanced finite population structure nor for unequal probability sampling. We develop design-based predictors for unequally sized clustered finite populations based on two stage sampling. The predictors are non-parametric and require no restrictive assumptions. They are developed using a prediction approach applied to a random permutation model. There are several unique features of the development. First, identifiability of clusters requires expanding the representation of such random variables in the random permutation model induced by the design. Second, collapsing the high dimensional random variables into sample and remainder totals (or means) enables tractable closed form predictors. Third, unconditionally unbiased predictors are only possible for sampling designs in which second stage sampling is conducted with probability proportional to size. However, a weaker unbiased constraint enables predictors to be developed for any two stage design. Finally, such predictors can be closely related to those in balanced designs and outperform their mixed model counterparts.

email: stanek@schoolph.umass.edu

ASCERTAINMENT ADJUSTMENT IN GENETIC STUDIES OF ORDINAL TRAITS

Rui Feng*, University of Alabama at Birmingham
Heping Zhang, Yale University

Most genetic studies recruit high risk families and the discoveries are based on non-random selected groups. We must consider the sequences of this ascertainment process in order to extend the results of genetic research to the general population. In previous papers, we developed a latent variable model to assess the familial aggregation and inheritability of ordinal-scaled diseases, and concluded a major gene component of alcoholism after applying the model to the data from the Yale Family Study of Comorbidity of Alcoholism and Anxiety (YFSCAA). In this report, we examine the ascertainment effects on parameter estimates and correct potential bias in the latent variable model. The simulation studies for various ascertainment schemes suggest the ascertainment adjustment is necessary and effective. We also find that the estimated effects are relatively unbiased for the particular ascertainment scheme used in the YFSCAA, which assures the validity of our earlier conclusion.

email: rfeng@ms.soph.uab.edu

INTERVAL MAPPING FOR EXPRESSION QUANTITATIVE TRAIT LOCI

Meng Chen*, University of Wisconsin-Madison
Christina Kendziorski, University of Wisconsin-Madison
Alan Attie, University of Wisconsin-Madison

The field of quantitative trait loci (QTL) mapping was reignited in the 1980's by advances that facilitated marker genotyping. Today, with major developments in high throughput technologies, an advance of comparable significance has been made in phenotyping. Expression measurements via microarrays are particularly informative when considered as phenotypes for QTL mapping, and much excitement now exists for the field of expression quantitative trait loci (eQTL) mapping. The statistical methods currently available for eQTL mapping are limited. Some allow for interval mapping but do not properly account for multiplicities, while others account for multiplicities but do not allow for interval mapping. We here present an interval mapping approach that accounts for relevant multiplicities and thereby controls experiment wide error rates. Results are demonstrated using both simulated data and data from a study of diabetes in mouse.

email: mengchen@wisc.edu

LINKAGE TESTS FOR AFFECTED RELATIVE-PAIRS WITH INCOMPLETE IBD AND KNOWN IBS

Dennis W. Buckman*, IMS Inc.
Zhaohai Li, George Washington University

Linkage tests for complex diseases often utilize the number of marker alleles that sampled relative-pairs share identical-by-descent (IBD) or the number of marker alleles shared identical-by-state (IBS). In this presentation, linkage statistics are proposed for situations where the marker IBD is incomplete and the marker IBS is known. Assuming the marker IBD is either missing at random or observed unambiguously, the likelihood function is derived for various types of affected relative-pairs. The linkage statistics for affected sib-pairs, aunt(uncle)-nephew(niece)-pairs, half-sib-pairs, and first-cousin-pairs are derived by considering the Taylor series expansion of the corresponding log likelihood about the null value of the recombination fraction. This derivation yields linkage statistics that are proportional to the second derivative of the log likelihood evaluated at the null hypothesis value of the recombination fraction, one-half. For affected grandparent-grandchild-pairs, the first derivative of the log likelihood is not fixed at zero when the recombination fraction is one-half; therefore, the usual approach for developing an efficient score test is utilized. Asymptotic distributions, required sample sizes, and simulation results are presented. In order to address the issue of ambiguous marker IBD, incorporation of marker IBD estimates is considered.

email: buckmand@imsweb.com

EVALUATING ADMIXTURE ESTIMATION AND MAPPING TECHNIQUES THROUGH PLASMODES

Laura K. Vaughan*, University of Alabama at Birmingham
Jasmin Divers, University of Alabama at Birmingham
Miguel Padilla, University of Alabama at Birmingham
Hemant K. Tiwari, University of Alabama at Birmingham
David T. Redden, University of Alabama at Birmingham
David B. Allison, University of Alabama at Birmingham

Admixture mapping has recently stepped into the spotlight as a potentially usefull method for gene mapping. It is particularly promising in populations, such as African Americans, that are formed as a result of recent mixing of distinct populations. This mixing results in long regions of linkage disequilibrium (LD). There has been a significant amount of interest in using this extended LD, combined with the ancestry proportion of individuals being studied, to identify loci that are associated with disease phenotypes. Many methods have been proposed, but few have been thoroughly tested. An attractive alternate to simulated data in the evaluation of these methodologies is the use of plasmodes, where the individuals have a known ancestry. We have gathered a plasmode consisting of a collection of mouse crosses from distinct founding populations, which therefore have known individual ancestry, and have been scored for various obesity related phenotypes and assayed with a collection of genetic markers. Using this plasmode, we can create datasets with varying ancestry to determine how effective different algorithms are at estimating individual admixture, which is a proxy for individual ancestry. We then test their effectiveness at detecting linkage and association of obesity related phenotypes when controlling for individual ancestry.

email: LVaughan@ms.soph.uab.edu

EXTENSION OF VARIANCE COMPONENT LINKAGE ANALYSIS TO INCORPORATE REPEATED MEASUREMENTS

Wei-Min Chen*, University of Michigan
Liming Liang, University of Michigan
Pak C. Sham, University of London and University of Hong Kong
Gonçalo R. Abecasis, University of Michigan

When subjects are measured multiple times, it is important for a linkage analysis to appropriately take into account these repeated measures. In this study, we extend the variance component approach to model repeated measures in a quantitative trait linkage study. We show for the case of a balanced design where the same number of measurements is taken for each subject, a standard linkage test that takes the average of measures as the trait of interest is identical to the linkage test based on our extension of the variance component model. We give the general formulas of optimal sample size and number of repeated measures for a given power or cost. We perform simulations to compare power for different sample sizes and number of repeated measures across several scenarios. We find that repeated measures provide substantial power improvements across genetic models. The proportional increase in expected LOD score depends mostly on measurement error and total heritability but not much on marker map or number of alleles per marker. Finally, we give recommendations on whether to take repeated measures or to recruit additional subjects for different levels of measurement errors and ratios of genotyping, subject recruitment and phenotyping costs.

email: wechen@umich.edu

CONFIDENCE SET INFERENCE ON MAXIMUM LOAD SCORE STATISTIC IN LINKAGE ANALYSIS: SOLVING THE PROBLEM OF MULTIPLE TESTING

Ritwik Sinha*, Case Western Reserve University
Yuqun Luo, Case Western Reserve University

A recent approach to gene mapping based on confidence set inference (CSI) (Lin et al., 2001) promises several advantages. CSI replaces the traditional hypotheses of linkage analysis (H_0: no linkage) by a new set of hypotheses with the null being that the marker is in tight linkage to a trait locus. This reformulation eliminates the problem of multiple testing plaguing traditional linkage analysis. Further advantages include readily available confidence sets with known statistical properties and sufficient localization of disease genes. Currently, CSI has been applied to the Mean and Proportion test statistics (Lin, 2002; Papachristou and Lin, 2005). A better statistic, Maximum Lod Score (MLS, Risch, 1990), makes maximum use of the information available from allelic identity by descent from ASPs, in addition to handling markers with incomplete polymorphism. We propose to test the new set of hypotheses under CSI using the MLS statistic. We conduct the analysis at all locations (not just markers). The tests are multipoint in that the IBD probabilities are calculated based on all the marker data. Our method is tested on data simulated under a wide range of disease models, as well as real data.

email: rsinha@darwin.cwru.edu

SEMIPARAMETRIC TRANSFORMATION MODELS FOR MAPPING QUANTITATIVE TRAIT LOCI WITH CENSORED DATA

Guoqing Diao*, University of North Carolina at Chapel Hill
Danyu Lin, University of North Carolina at Chapel Hill

Variance-component (VC) methods are widely used in the linkage and association analysis of quantitative traits in general human pedigrees. The standard 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. When the trait pertains to the age at onset of a disease, censoring is inevitable because of loss to follow-up and limited study duration. Censoring also arises if the assay cannot detect values smaller (or larger) than some threshold. Applying the standard VC methods to such censored trait data would result in inflated type I error and reduced power. We develop valid and powerful VC methods for censored trait data based on a novel class of semiparametric linear transformation models. Under the proposed models, the latent trait values follow a specific distribution, such as the normal distribution, after a completely unknown transformation. We construct appropriate likelihood functions for the observed data, which may contain left or right censored observations. We develop efficient algorithms to implement the corresponding estimation and testing procedures. Our methods can be used for both linkage and association analysis. Extensive simulation studies demonstrate that the proposed methods outperform the existing methods in practical situations. Applications to a real data are provided.

email: gdiao@bios.unc.edu

33. IMAGING METHODS

HIGH DIMENSION, LOW SAMPLE SIZE PRINCIPAL COMPONENTS

Keith E. Muller, University of North Carolina-Chapel Hill
Yueh-Yun Chi*, University of Washington
Jeongyoun Ahn, University of North Carolina-Chapel Hill
Steve Marron, University of North Carolina-Chapel Hill

Medical images and genetic assays typically generate High Dimension, Low Sample Size (HDLSS) data, namely more variables than independent sampling units. Currently, scientists often use Principal Components Analysis (PCA) of sample covariance matrices to work around the limitations of HDLSS. We provide analytic and Monte Carlo simulations for Gaussian data which strongly discourage the practice. An exception may occur if a few components thoroughly dominate the population covariance pattern, and the number of dominant components falls well below the number of independent sampling units. A nonsingular decomposition of the data implicitly defines a nonsingular matrix with the same nonzero eigenvalues as the singular sample covariance matrix. The equivalence gives a variety of exact and approximate moments and distributions for PCA of HDLSS and Gaussian data.

email: yychi@u.washington.edu

ANALYZING DIFFUSION TENSOR IMAGING DATA

Meagan E. Clement*, Rho, Inc.

A recent protocol innovation with magnetic resonance imaging (MRI) results in diffusion tensor imaging (DTI). The approach holds tremendous promise for improving our understanding of neural pathways, especially in the brain. Unfortunately, little has been done to define metrics or describe credible statistical methods for analyzing DTI data. The present work proceeds in a simple sequence. First, a review of DTI summary measures allows concluding that they all can be interpreted as statistical estimators of population properties for Gaussian stochastic processes. Second, a statistical perspective gives a clear preference among the measures, once some actual data have been considered. Third, one-to-one transformations of the most useful measures lead to accurate representations of their observed distributions in terms of only two estimated parameters each. Fourth, and most importantly, statistical analysis of the four estimated parameters suffices to capture all of the information. In turn, using the estimated parameters as outcomes in statistical models avoids the "curse of dimensionality" (from having far more variables than independent sampling units).

email: clement@email.unc.edu

ESTIMATION EFFICIENCY AND STATISTICAL POWER IN ARTERIAL SPIN LABELING FMRI

Jeanette A. Mumford*, University of Michigan
Luis Hernandez-Garcia, University of Michigan
Gregory R. Lee, University of Michigan
Thomas E. Nichols, University of Michigan

Arterial Spin Labeling (ASL) is a method for measuring blood perfusion quantitatively in the brain with Magnetic Resonance Imaging (MRI). An ASL experiment collects a sequence of images in alternating label (L) and control (C) states, and perfusion information is contained in the difference between L and C images. The standard analysis approach for a sequence of N images is to subtract C/L pairs, yielding a dataset of N/2 difference images. This has also been reported to 'whiten' or decorrelate the data, which is convenient as the full time series has substantial temporal autocorrelation. We show that the simple differencing strategy is suboptimal. Our proposed method is to model the full length-N data, embedding the C/L effect in the model, we explicitly model the temporal autocorrelation, and use Generalized Least Squares (GLS) to obtain fully efficient estimates of the experimental effects. Analytically, we show that differencing methods are less efficient, with standard errors up to 10% larger relative to modeling all the data. Using real data, we find that our method produces increased sensitivity relative to differencing. Since autocorrelation estimation and whitening with GLS is a standard feature of fMRI software, our method is easy to implement.

email: jmumford@umich.edu

FUNCTIONAL PRINCIPAL COMPONENT REGRESSION AND FUNCTIONAL PARTIAL LEAST SQUARES

Philip T. Reiss*, Columbia University
R. Todd Ogden, Columbia University

Regression of a scalar response on signal predictors, such as near-infrared (NIR) spectra of chemical samples, requires reducing the dimension of the predictors. This is usually done either by regressing on components--e.g. principal component regression (PCR) and partial least squares (PLS)--or by spline smoothing methods. We introduce functional versions of PCR and PLS, which combine both of the above dimension reduction approaches. Two versions of functional PCR are developed, both employing B-splines and roughness penalties. The regularized-components version applies such a penalty to the construction of the principal components, while the regularized-regression version incorporates a penalty in the regression. Proceeding similarly, we develop two versions of functional PLS. Simulation and split-sample validation studies with NIR spectroscopy data indicate that functional PCR and functional PLS offer advantages over existing methods in terms of both estimation of the coefficient function and prediction of future observations. Functional PLS is also applied to a neuroimaging data set

email: ptr2003@columbia.edu

SPATIOTEMPORAL MODELING OF FUNCTIONAL MAGNETIC RESONANCE IMAGING DATA

Qihua Lin*, Southern Methodist University
Patrick S. Carmack, University of Texas Southwestern Medical Center at Dallas
Richard F. Gunst, Southern Methodist University
William R. Schucany, Southern Methodist University
Jeffrey S. Spence, University of Texas Southwestern Medical Center at Dallas

Functional magnetic resonance imaging (fMRI) is a promising technique used in many fields, such as neuroscience, to study the functions of human brains. In a typical fMRI experiment, three-dimensional brain scans, each consisting of signals from hundreds of thousands volume elements (voxels), are taken repeatedly over a period of time while a brain is alternating between resting and undertaking tasks. The resultant data are of very complex nature with both spatial and temporal correlations, differential spatial and temporal responses to stimuli, and very low signal-to-noise ratios. A class of spatiotemporal models in a generalized linear model framework is proposed to model such data. The time course of the fMRI signal at each voxel is characterized by a baseline signal, a drift component, a hemodynamic response, and a spatially and temporally correlated error process. An algorithm for model identification and parameter estimation is outlined. These models are expected to have an increased power in identifying brain regions activated by the stimuli.

email: qlin@smu.edu

PREDICTION OF POST-TREATMENT BRAIN ACTIVITY USING A BAYESIAN HIERARCHICAL MODEL

Ying Guo*, Emory University
DuBois Bowman, Emory University

In neuroimaging studies, researchers are sometimes interested in predicting post-treatment brain function based on individual-specific factors. An accurate prediction mechanism provides invaluable information for choosing an optimal treatment for each subject by considering current brain status and relevant health characteristics. We present a statistical method to predict post-treatment brain function using a Bayesian hierarchical model. The first level of the hierarchy models the activation effects across within-subject scans through subject-specific parameters and the second level models subject-specific effects in terms of population parameters. Estimation is performed using the EM algorithm. The accuracy of the proposed prediction method based on the Bayesian hierarchical model is evaluated using K-fold cross-validation. We illustrate the application of our method using positron emission tomography (PET) data from a study of working memory in schizophrenia patients.

email: yguo2@sph.emory.edu

34. STATISTICAL ISSUES IN USING EXPOSURE ESTIMATES IN ENVIRONMENTAL EPIDEMIOLOGY

SEMIPARAMETRIC DYNAMIC STRUCTURAL MODELS FOR MULTIVARIATE EXPOSURES

Amy H. Herring*, The University of North Carolina at Chapel Hill
David B. Dunson, National Institute of Environmental Health Sciences-National Institutes of Health

Exposure assessment in environmental epidemiology often involves measurement of multiple indicators that may vary over time. Structural equation models with a latent time-varying exposure are natural in this setting but typically require strong parametric assumptions about the distribution of the latent variables. We propose a semiparametric Bayesian approach that allows the distribution of the latent exposure to be unknown and to vary over time. The method is applied to a study of drinking water disinfection by-products and adverse pregnancy outcomes.

email: aherring@bios.unc.edu

SOURCE APPORTIONMENT: FROM CHARACTERIZATION TO IMPUTATION

Thomas Lumley*, University of Washington

The source apportionment problem is the problem of decomposing time series of chemical compositions of air pollution particles into a sum of contributions from different sources, based on substantial but incomplete knowledge of the composition of each source. This problem is typically not identifiable, but estimates of average source contributions have proved very useful in characterising the burden of air pollution produced by different sources. The same techniques are now being used to impute daily exposures to different sources for use in epidemiological models. The statistical properties of the resulting health effect estimators are not well understood. I will describe some approaches to this problem and simulation results on the bias resulting.

email: tlumley@u.washington.edu

EXPOSURE MEASUREMENT ERROR CAUSED BY SPATIAL MISALIGNMENT IN ENVIRONMENTAL EPIDEMIOLOGY

Alexandros Gryparis*, Harvard University
Christopher Paciorek, Harvard University
Brent A. Coull, Harvard University

In analyzing the health effects of environmental exposures, when the exposure covariate is based on predictions from a spatial model, the epidemiological model involves measurement error caused by spatial misalignment. We provide a framework for spatial measurement error modeling, showing that smoothing induces a Berkson-type measurement error with non-diagonal error structure. From this viewpoint, we review several existing approaches to estimation: direct use of the spatial predictions, multiple imputation, and Bayesian approaches. We also propose two new approaches to estimation, one based on regression calibration and another based on iterative weighted least squares. Based on analytical considerations and simulation results, we compare the performance of all of these approaches, and suggest several methods that improve upon direct use of the spatial predictions in an epidemiological model.

email: al_grip@yahoo.com

MODEL CHOICE IN TIME SERIES STUDIES OF AIR POLLUTION AND MORTALITY

Roger D. Peng*, Johns Hopkins University
Francesca Dominici, Johns Hopkins University
Thomas A. Louis, Johns Hopkins University

Multi-city time series studies of particulate matter (PM) and mortality and morbidity have provided evidence that daily variation in air pollution levels is associated with daily variation in mortality counts. Methodological issues concerning time series analysis of the relation between air pollution and health have attracted the attention of the scientific community. Investigators around the world have used different approaches to adjust for confounding, making it difficult to compare results across studies. To date, the statistical properties of these different approaches have not been comprehensively compared. To address these issues, we quantify and characterize model uncertainty and model choice in adjusting for seasonal and long-term trends in time series models of air pollution and mortality. First, we conduct a simulation study to compare and describe the properties of statistical methods commonly used for confounding adjustment. Second, we apply and compare the modelling approaches to the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) database, containing data on the the largest 100 cities in the United States.

email: rpeng@jhsph.edu

35. SOLUTIONS FOR MISSING DATA IN COMPLEX SAMPLE SURVEYS RELEVANT IN HEALTH POLICY RESEARCH

INTEGRATED DESIGN AND ESTIMATION STRATEGIES TO CORRECT FOR MISSING DATA IN THE MEDICAL EXPENDITURE PANEL SURVEY

Steven B. Cohen*, AHRQ

The Medical Expenditure Panel Survey (MEPS) is an ongoing longitudinal panel survey designed to produce national estimates of health care utilization, expenditures, sources of payment, and insurance coverage of the U.S. civilian non-institutionalized population. As a consequence of its breadth, the data have informed the nation's economic models and their projections of health care expenditures and utilization. Given the longitudinal nature of its design, the survey is subject to missing data at the wave level in addition to both sampled unit and the item level. To improve the accuracy of the resultant estimates derived from the survey, MEPS is characterized by an integrated set of innovative data collection procedures and estimation strategies. This paper provides a summary of the design and estimation strategies implemented in the MEPS to achieve reductions in the bias in estimates attributable to missing data. Attention is also given to the utility of this integrated survey design framework to achieve improvements in data quality through linkages between sampled individuals, their providers, pharmacies and employers.

email: scohen@ahrq.gov

MULTIPLE IMPUTATION FOR NON-NORMAL CONTINUOUS MISSING VARIABLES IN COMPLEX SURVEYS

Yulei He*, Harvard University
Trivellore E. Raghunathan, University of Michigan

A popular method for handling item-missingness is through multiple imputation, where the set of missing values is "filled-in" by several plausible sets of values to create completed data sets. Each completed data set is analyzed separately and the point estimates and standard errors are combined to construct a single inference. Practically, continuous variables in survey data sets are often nonnormally distributed. To accommodate such nonnormality, we propose imputation approaches in which the continuous data are modeled through Tukey's g-and-h family and its extensions. The developed approaches can handle multivariate data under complicated missing data patterns and mechanisms. We evaluate these methods on actual and simulated data sets. In addition, complex survey design features such as stratification and clustering can be incorporated.

email: he@hcp.med.harvard.edu

MISSING DATA IN CLUSTER SAMPLES: DESIGN-BASED AND BAYESIAN PERSPECTIVES

Recai M. Yucel*, University of Massachusetts
Joseph L. Schafer, The Pennsylvania State University

Methods for handling missing data in sample surveys have been increasingly popular in modern statistical practice. In this paper we review and compare solutions proposed under frequentist and Bayesian approaches. The frequentist approach or desing-based approach, often practiced by survey statisticians, includes sampling-weight adjustments for unit-nonresponse and some method of imputation for item nonresponse. The Bayesian approach or model-based approach are often used to solve item-nonresponse problem while incorporating sampling weights. Suprisingly very little has been done to compare these two methods and to investigate questions such as whether inference based on model-based multiple imputation is compatible with design-based procedures (e.g. as implemented in SUDAAN, GEE methods), or how do the missing data adjustments under design-based approaches perform in the context of secondary analyses, which are primary goals of subject-matter investigators. Finally, we compare performance of these methods perform under different non-response mechanisms (MAR, MCAR, MNAR) by simulations studies and demonstrate the different approaches in large-scale complex surveys.

email: yucel@schoolph.umass.edu

36. FUSING BIOMEDICAL / ENVIRONMENTAL DATA WITH NUMERICAL MODELS

ESTIMATING PARAMETERS FOR HUGE SYSTEMS: TUNING THE COMMUNITY ATMOSPHERE MODEL

Douglas W. Nychka*, National Center for Atmospheric Research

The Ensemble Kalman filter (EKF) is a Monte Carlo based algorithm for data assimilation. Technically, the EKF is an approximate solution of the basic Bayesian filtering problem for dynamical systems. It has the potential, with suitably tuned algorithms, to facilitate data assimilation and to solve inverse problems with only a modest amount of investment in software development and with few assumptions on the linearity of the system. In this talk we describe how it can be applied large geophysical problems such as estimating global weather fields. From a statistical point of view, sequential and local ensemble updates are an intriguing practical approximation to solve a well defined Bayes problem. A goal of this talk is to indicate the simple connection between the EKF update and a more conventional spatial statistics analysis (optimal interpolation) of a geophysical field. This connection helps to illustrate the basic operation of the filter as being equivalent to a simple linear regression.

email: nychka@ucar.edu

STATISTICAL METHODS FOR DATA ASSIMILATION APPLIED TO HURRICANE FORECASTING

Montserrat Fuentes*, North Carolina State University
Kristen Foley, North Carolina State University

Estimating the spatial and temporal variation of surface wind stress fields plays an important role in modeling atmospheric and oceanic processes. This is particularly true for hurricane forecasting. According to NOAA Hurricane Research Division more than 85% of storm surge is caused by winds pushing the ocean surface ahead of the storm. Estimated storm surge values using coastal ocean deterministic models are used for assessments of warnings and evacuation notices and provide valuable information for recovery operations and response planning. We propose a Bayesian spatial-temporal statistical framework to obtain more accurate prediction of wind fields by combining wind information from bouys, ships, satellites, and physical models. Our framework captures different sources of uncertainty and biases about the data and the wind physical models and takes into account spatial misalignment and the change of support problem. A spatial-temporal nonstationary linear model of coregionalization is introduced to explain variability in the horizontal and vertical wind components as well as the cross-dependency between these two components. We present the improvement in the estimated storm surge using coastal ocean deterministic models by having as input fields our predicted wind fields.

email: fuentes@stat.ncsu.edu

SPATIO-TEMPORAL DYNAMICS OF THE SPREAD OF RACCOON RABIES

Lance Waller*, Emory University

The particular strain of rabies associated with raccoon hosts has been endemic in Florida and southern Georgia for many years. At some point around 1977, infected raccoons appeared near the Virginia/West Virginia border and the infection has been spreading from this point, most recently crossing into Ohio. We present background on the ongoing outbreak and review various mathematical models of the spatial dynamics of the spread of the disease. In addition to such models of disease spread, we also have access to tissue samples from rabid animals providing sequencing information on both the virus and the hosts. We explore statistical links between the dynamic spread of infection and the geographic and genetic 'distances' observed between host and virus samples, and move from current testing approaches toward more model-based inference.

email: lwaller@sph.emory.edu

37. STATISTICAL METHODS FOR PUBLIC HEALTH STUDIES IN DEVELOPING COUNTRIES

ACCOUNTING FOR VARIABILITY IN SAMPLE SIZE ESTIMATION WITH APPLICATION TO A MALARIA VACCINE PHASE 2 TRIAL

Michael P. Fay*, National Institute of Allergy and Infectious Diseases
M.E. Halloran, Emory University
Dean A. Follmann, National Institute of Allergy and Infectious Diseases

We use data from a longitudinal study in Mali to study several possible endpoints for designing a Phase 2 vaccine trial for a Malaria vaccine. We assume that the differences in endpoints between 4 year old control subjects and 4 year old vaccinated subjects are similar to observed differences between non-vaccinated 4 and 8 year olds. Because there are less than 50 children within one year of each age group in the preliminary data, there is considerable variability in the resulting sample size estimates. We develop a general method for adjusting the sample size estimates to account for this variability, so that the final sample size estimates have the proper power assuming only that the preminary data and the planned study follow the same probability model.

email: mfay@niaid.nih.gov

DESIGN AND ANALYSIS OF CLUSTER-RANDOMIZED PHASED IMPLEMENTATION STUDIES

Lawrence H. Moulton*, Johns Hopkins Bloomberg School of Public Health

In many settings in less industrialized countries, community-randomized studies can be useful to estimate the total effect of an intervention. This is especially the case when the outcome is a manifestation of an infectious disease. Yet a standard parallel design may not be logistically possible, or politically or culturally acceptable, especially when one arm of a trial involves placebo or standard of care. An alternative is to begin the study in one area, then another, and so on until all areas have received the intervention. 'Stepped wedge' and 'one-way crossover' are other terms for such designs. This presentation describes methods for randomizing these studies, sample size considerations, and possible analytic approaches. An example of a tuberculosis intervention study in Brazil will illustrate the concepts.

email: lmoulton@jhsph.edu

ANALYSIS OF TREATMENT EFFECTS WHEN THE TREATMENT AND OUTCOME ARE SPATIALLY CORRELATED WITH APPLICATION TO A GOVERNMENT FOOD RELIEF PROGRAM IN BANGLADESH

Dylan S. Small*, The Wharton School, University of Pennsylvania

During a major flood in Bangladesh in 1998, the government of Bangladesh conducted a food relief program which provided food aid to certain households. We analyze the effect of this program on various health outcomes such as per capita calorie consumption and the anthropometry of children. Receipt of the food relief was spatially correlated and the outcomes that would have been observed without the food relief are likely also spatially correlated, in part due to flood exposure being spatially correlated. Consequently, in analyzing the effect of the food relief program under an assumption of ignorable treatment assignment conditional on covariates, it is important to account for the spatial correlation. We use hierarchical spatial models to account for the spatial correlation and show the importance of doing so for making appropriate inferences. We also discuss a sensitivity analysis for the assumption of ignorable treatment assignment.

e-mail: dsmall@wharton.upenn.edu

MEASUREMENT ISSUES RELATED TO QUANTIFYING OLIGOSACCHARIDES IN HUMAN MILK TO DETERMINE THEIR ASSOCIATION WITH INFANT DIARRHEA

Mekibib Altaye*, Cincinnati Children's Hospital Medical Center and University of Cincinnati

Diarrhea is a leading cause of death in developing countries, especially in children. Breastfeeding provides significant protection against diarrhea in infancy, yet some breast-fed infants experience multiple episodes of diarrhea. The risk of diarrhea in breast-fed infants can be partly explained by lack of exclusive breast feeding and by some environments that produce high-dose exposure to enteric pathogens both of which may be more prevalent in developing countries. However variations in the composition of protective factors in human milk could also account for variation in risk of diarrhea among such infants. One of the goals in our NIH funded study in Mexico City, Mexico is to explore whether the rates of diarrhea are inversely related with one or more major oligosaccharides of human milk. However the quantification and distribution of each oligosaccharide in human milk may follow a different trajectory throughout infancy. Therefore a single sample obtained early in infancy may or may not be a representative of such trajectory. In this presentation we will discuss some of the measurement issues related to oligosaccharide measurements and more generally, some of the analytic challenges in collaborating on a study that is being conducted in Mexico.

e-mail: altq56@cchmc.org

38. IMS: SPATIOTEMPORAL STATISTICS

A MODIFICATION OF THE EM ALGORITHM WITH APPLICATION TO SPATIO-TEMPORAL MODELS

Stanislav Kolenikov*, University of Missouri

This presentation outlines computational problems encountered in the spatio-temporal modeling, and proposes a modification of the EM algorithm that only uses expectations based on the marginal distributions of the incomplete data. The corrections required for the resulting estimating equations are derived, and asymptotic properties of the resulting estimates are discussed. An empirical example based on a study of PM2.5 is provided.

email: kolenikovs@missouri.edu

ESTIMATING DEFORMATIONS OF ISOTROPIC GAUSSIAN RANDOM FIELDS

Ethan B. Anderes*, University of California-Berkeley

The first part of this talk will present a new approach to the estimation of the deformation of an isotropic Gaussian random field on $\mathbb R ^ 2$ based on dense observations of a single realization of the deformed random field. The use of deformations for modeling non-stationary processes has been applied in diverse fields from geophysics to image analysis. These models are natural extensions of stationary processes that are simple to understand but give rise to a diverse range of behavior. Even though these models seem a good choice when modeling non-stationary random fields they are generally difficult to work with because of the complex restrictions on the deformations like invertablility. This work establishes methodology using a general nonparametric representation of deformations that makes these models tractable. We present a complete methodological package---from model assumptions to algorithmic recovery of the deformation---for the class of non-stationary processes obtained by deforming isotropic Gaussian random fields. The remainder of this talk will discuss applications of these methods to modeling space-time processes using temporally evolving deformations of space-time processes.

email: anderes@stat.berkeley.edu

BAYESIAN MODELING OF EXTREME PRECIPITATION RETURN LEVELS

Daniel S. Cooley*, NCAR/Colorado State University
Douglas Nychka, NCAR
Philippe Naveau, University of Colorado/LSCE-CNRS-IPSL

Quantification of extreme values is important for planning purposes. To aid with the understanding of potential flooding along Colorado's Front Range, we have developed a procedure for producing a map of 24-hour precipitation return levels and uncertainty estimates for the region. We build a three-layer Bayesian hierarchical model . The first layer models daily precipitation above a threshold with the generalized Pareto distribution (GPD) from extreme value theory. The second layer models the spatial latent process that drives the climatological extreme precipitation. We use standard geophysical methods to model the GPD parameters spatially. The third layer designates prior distributions for the spatial parameters. To obtain convergence of the spatial parameters and a map that agrees with the region's geography, the spatial analysis is done in a space where locations are given by climatological quantities rather than by latitude and longitude. The flexibility of the general model allows various sub-models to be compared. We are working to extend the procedure to account for precipitation events of different duration periods. Planners need information about short-term (e.g. 2-hour) and long-term (2-day) precipitation events. Typically, these analyses are performed separately. We are working to produce a model which would account for all duration periods at once. Such a model would borrow strength from the different duration periods as well as from the different stations and would provide insight into how the nature of extreme precipi-tation changes with duration length.

email: cooleyd@ucar.edu

39. COMPUTATIONAL, CLASSIFICATION, AND MODEL SELECTION METHODS

MODEL COMPLEXITY AND THE AIC STATISTIC FOR NEURAL NETWORKS

Doug Landsittel*, Duquesne University
Dustin Ferris, Duquesne University

This study investigates the underlying complexity of neural networks (with single binary outcomes) through simulations and use of the generalized degrees of freedom. Results are then applied to using the AIC statistic for model selection. Specifically, we investigate how different measures of complexity affect determination of the optimal model; subsequent recommendations are outlined for variable selection and specifying other aspects of the network architecture.

email: landsitteld@duq.edu

ON CREATING MODEL ASSESSMENT TOOLS INDEPENDENT OF SAMPLE SIZE

Jiawei Liu*, Georgia State University
Bruce G. Lindsay, Penn State University

A standard goal of model evaluation and selection is to find a model that approximates the truth well while at the same time is as parsimonious as possible. In this paper we emphasize the point of view that the models under consideration are almost always false, if viewed realistically, and so we should analyze model adequacy from that point of view. We investigate this issue in large samples by looking at two types of sample size indices, which are designed to serve as one-number summary measures of model adequacy. We define these indices to be the maximum sample size at which samples from the model and those from the true data generating mechanism are nearly indistinguishable. We show that these definitions lead us to a new way of viewing models as flawed but useful. These concepts are an extension of some important work of Davies (1995).

email: jliu@mathstat.gsu.edu

CLASSIFICATION OF PSYCHOTROPIC DRUGS BASED ON SLEEP - WAKING BEHAVIOR IN RATS

Kristien Wouters, Hasselt University-Diepenbeek, Belgium
José Cortiñas Abrahantes*, Hasselt University-Diepenbeek, Belgium
Abdellah Ahnaou, J&J PRD Janssen Pharmaceutica-Beerse, Belgium
Helena Geys, J&J PRD Janssen Pharmaceutica-Beerse, Belgium
Geert Molenberghs, Hasselt University-Diepenbeek, Belgium
Pim Drinkenburg, J&J PRD Janssen Pharmaceutica-Beerse, Belgium

For thousands of years, humans have used psychoactive substances. Despite the wide variety of effects that such substances can exert on the central nervous system, attempts have been made to categorize drugs into psychoactive classes based on therapeutic efficacy, such as antidepressants, antipsychotics, anxiolytics, hypnotics and stimulants. Pharmaco-ElectroEncephaloGraphy (pEEG) can be reliably used in humans for the discrimination of clinically active, psychotropic drugs, which has fostered the development of corresponding animal pharmaco-EEG models. A crucial problem for pharmaco-EEG studies is that the pharmacological effects on the EEG are easily confounded by marked EEG alterations associated with spontaneous changes in behavior or vigilance. In our approach rats were left undisturbed in order to allow our analyses to separate out six different spontaneous sleep-wake stages: Active Wake, Passive Wake, Light, Deep, Intermediate Stage and REM Sleep. Analyzing EEG data poses an important challenge, because of the high-dimensionality and the longitudinal character of the data. Therefore, a fractional polynomial mixed model, coupled with hierarchical discriminant analysis, has been proposed. In each step of the hierarchical discriminant analysis, one class is differentiated from the rest. The results obtained thus far show high discriminant properties for antipsychotics, anxiolytics and stimulants, while antidepressants and hypnotics are more difficult to differentiate from control.

email: jose.cortinas@uhasselt.be

ESTIMATION OF STOCHASTICALLY ORDERED SURVIVAL FUNCTIONS BY GEOMETRIC PROGRAMMING

Johan Lim, Texas A&M University
Xinlei Wang*, Southern Methodist University
Seung Jean Kim, Stanford University

In the literature, much attention has been given to computing the nonparametric maximum likelihood estimate (NPMLE) of survival functions under various stochastic order constraints. However, every existing procedure is only applicable to a specific type of order constraints or a specific setting; and it often requires effort to implement. In this paper, we propose a new method to compute the NPMLE of survival functions, which is applicable to all the existing order constraints related to survival functions that we are aware of. Our approach is to reformulate the estimation problem as a geometric program (GP). A GP is a standard convex optimization problem characterized by objective and constraint functions that have a special form. Recently developed methods can solve even large-scale GPs extremely efficiently and reliably. Hence, by reformulating the problem to a GP, we get an effective way to compute the NPMLEs. We apply the proposed method to two data sets to illustrate its flexibility in and efficiency in finding the NPMLE.

email: swang@smu.edu

RANOVA: A NEW METHOD OF DETECTING DIFFERENTIALLY EXPRESSED GENES THROUGH PROBE LEVEL DATA FROM OLIGONUCLEOTIDE ARRAYS

Jin Xu*, Ambion Services, Ambion Inc.
Timothy S. Davison, Ambion Services, Ambion Inc.
Charles D. Johnson, Ambion Services, Ambion Inc.

Oligonucleotide arrays such as Affymetrix GeneChips use multiple probes, or a probe set, to measure the abundance of mRNA of every gene of interest. Some analysis methods attempt to summarize the multiple observations into one single score before conducting further analysis such as detecting differentially expressed genes (DEG), clustering and classification. However, there is a risk of losing a significant amount of information and consequently reaching inaccurate or even incorrect conclusions during this data reduction. We developed a novel statistical method called robust analysis of variation (RANOVA) to detect DEG for both two-group and k-group cases. It utilizes probe level data and requires no assumptions about the distribution of the dataset. The method was tested on benchmark datasets and compared with existing summarization methods (RMA, GCRMA, MAS5, PLIER, dChip). The results show that our method successfully detects DEG with positive predictive value of 94% while maintaining a low false discovery rate and consistently out performs the existing methods.

email: jin.xu@ambion.com

40. SURVIVAL ANALYSIS I

ESTIMATING TIME TO EVENT FROM LONGITUDINAL CATEGORICAL DATA: AN ANALYSIS OF MULTIPLE SCLEROSIS PROGRESSION

Micha Mandel*, Harvard School of Public Health
Susan A. Gauthier, Brigham and Women's Hospital
Charles R.G. Guttmann, Brigham and Women's Hospital
Howard L. Weiner, Brigham and Women's Hospital
Rebecca A. Betensky, Harvard School of Public Health

The extended disability status scale (EDSS) is an ordinal score that measures progression in multiple sclerosis (MS). Progression is defined as reaching EDSS of a certain level (absolute progression) or increasing of one point of EDSS (relative progression). Survival methods for time to progression are not adequate since they do not exploit the EDSS level of censored observations. Instead, we suggest a Markov transitional model applicable for repeated categorical or ordinal data. This approach enables derivation of covariate-specific survival curves, obtained after estimation of the regression coefficients and manipulations of the resulting transition matrix. Large sample theory and resampling methods are employed to derive pointwise confidence intervals, which perform well in simulation. The regression models described are easily implemented using standard software packages. Survival curves are obtained from the regression results using packages that support simple matrix calculation. We present and demonstrate our method on longitudinal data collected at the Partners MS center in Boston, MA. We apply our approach to progression defined by time to two consecutive visits with EDSS greater than three, and calculate crude (without covariates) and covariate-specific curves.

email: mmandel@hsph.harvard.edu

METHODS FOR THE ACCELERATED FAILURE TIME MODEL

Zhezhen Jin*, Columbia University

In this talk, I will review and present various estimationand inference methods for the accelerated failure time models developed recently, which includes rank-estimation method, least squares method and M-estimation method and their generalizations to multivariate case.

email: zj7@columbia.edu

A DYNAMIC MODEL FOR SURVIVAL DATA WITH LONGITUDINAL COVARIATES

Gary L. Rosner, The University of Texas-M.D. Anderson Cancer Center
Krzysztof J. Rudnicki*, Rice University

Analyses involving both longitudinal and time-to-event data are quite common in medical research. We propose a flexible semiparametric Bayesian hierarchical modeling approach to analyzing these two types of data jointly by use of dynamic models in both survival and longitudinal aspects of the model. The shapes of the trajectories of the time-dependent parameters are determined by the data rather than the assumed model. The longitudinal marker trajectories are patient specific, and the link between them is provided by the hierarchical structure of the model. A combination of various MCMC techniques, such as Gibbs sampling and Metropolis-Hastings algorithm, is used to obtain a sample from the joint posterior distribution of all the model parameters. Simulation studies provide the measure of the quality of the newly developed method and a real data example follows.

email: k_rudnicki@yahoo.com

CHECKING THE CENSORED TWO-SAMPLE ACCELERATED LIFE MODEL USING INTEGRATED CUMULATIVE HAZARD DIFFERENCE

Seung-Hwan Lee*, Illinois Wesleyan University

Some new statistical tests for the censored two-sample accelerated life model are discussed. Using some estimating functions, certain stochastic processes are introduced. They can be described by martingale residuals, and, given the data, conditional distributions can be approximated by zero mean Gaussian processes. Unlike usual methods regarding the model checking for the two-sample accelerated life model, the new methods 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 well, especially whn the censoring is heavy. The proposed procedures are illustrated with two real data sets.

email: slee2@iwu.edu

SEMIPARAMETRIC SURVIVAL MODELS WITH CENSORED COVARIATES

Gina M. D'Angelo*, University of Pittsburgh
Lisa Weissfeld, University of Pittsburgh

Biomarker data are often censored due to the inability of assays to accurately detect levels of the marker below a given threshold. While methods are readily available for handling censored outcome data, fewer methods are available when both the covariate and outcome are censored. Methods developed for regression models with a censored covariate include imputation, likelihood approaches, and semiparametric approaches. Rigobon et al. (2004) developed a semiparametric approach utilizing single-index models for a censored covariate in a regression model. We propose to develop a semiparametric survival model with censored covariates by extending methodology from Rigobon and Lu et al. (2005). Lu developed partially linear single-index survival models that allow for functional and linear effects of the covariates. We propose profile quasi-likelihood estimation using single-index survival models for the censored data and the Cox proportional hazards model for complete data. In the single-index survival model, the linear effect of the censored covariates will be replaced with a function of the linear effect of the fully observed covariates. Our extension will be compared to a method of filling in the censored values with the lower threshold value and an inverse probability weighted estimating equation. We apply this method to a sepsis study.

email: gmdst17@pitt.edu

NONPARAMETRIC REGRESSION USING KERNEL ESTIMATING EQUATIONS FOR CORRELATED FAILURE TIME DATA

Zhangsheng Yu*, University of Michigan
Xihong Lin, Harvard School of Public Health

We study nonparametric regression for the correlated failure time data under the marginal proportional hazard model. Kernel regression estimating equations are used to estimate nonparametric covariate effects. Independent and weighted working kernel estimating equations(EE) derived from local partial likelihood are studied. The derivative of the covariate function is first estimated and the covariate function estimator is obtained by integrating the derivative estimator. The Trapezoidal rule is used for integration approximation. We show that the nonparametric estimator of the covariate function~{!/~}s derivative is consistent for any arbitrary working correlation matrix and the asymptotic variance is minimized by assuming working independence. We evaluate the performance of the proposed kernel estimator using simulation studies, and apply the proposed method to western Kenya parasitaemia. The semiparametric hazard regression is also considered.

email: zyuz@umich.edu

ON SAMPLE SIZE SELECTION IN CLINICAL TRIALS WITH BOTH ACCRUAL AND FOLLOW-UP PERIODS FOR SEVERAL TREATMENT GROUPS

Susan Halabi*, Duke University
Bahadur Singh, Cancer Center Biostatistics, Duke University and Linberger Cancer Center, UNC

In this talk, we generalize the results of Rubinstein et al. to allow for testing the equality of hazards for several treatment groups assuming Poisson accrual rate. With a pre-specified type-I error rate, power, accrual rate, follow-up period, and hazard ratios, formula for the accrual period of the trial and total sample size are provided. An example illustrating the sample size selection is included. Simulations are performed to evaluate the performance of the power of the MLE test. Tables will be provided which give the required accrual or enrollment period given the continuation rate, entry rate per year, type I error rate of 0.05, power of the MLE test=0.80 or 0.90 and various values of ratio of the median survival in groups j (j>=2) to group 1.

email: susan.halabi@duke.edu

41. METHODS IN EPIDEMIOLOGY

CONSTRUCTING BETTER BINOMIAL CONFIDENCE INTERVALS BY REMEMBERING TWO TECHNIQUES FOR NORMAL CONFIDENCE INTERVALS

Craig B. Borkowf*, US Centers for Disease Control and Prevention

When we construct confidence intervals for normal or conceptually continuous data, we employ two standard techniques to improve the accuracy of those intervals. First, when we estimate the variance from the data, we use the t-distribution instead of the standard normal distribution. Second, we use the maximum likelihood estimator of the variance, but with (n – 1) instead of n in the denominator, where n denotes the sample size. By contrast, we seem to forget these two important techniques when we construct binomial confidence intervals. First, we obey a questionable dogma that dictates that one may not use the t-distribution for binomial confidence intervals, despite the obvious fact that the proportion parameter and hence the true variance is unknown. Second, and still more puzzling, we use the biased maximum likelihood estimator of the variance, $p(1 – p)/n$, when we could just as easily use the unbiased estimator, $p(1 – p)/(n – 1)$, where p is the sample proportion. We show that with a single additional technique, that of slightly adjusting the sample proportion, we can obtain binomial confidence intervals with at least near nominal coverage for all proportion parameters and sample sizes. We discuss the appropriate adjustments in terms of the Agresti-Coull (1998) and the SAIFS (Borkowf, 2006) methods.

email: uzz3@cdc.gov

SOME THOUGHTS ON THE RELATIVE SURVIVAL RATE

Chris M. Drake*, University of California, Davis
Julie Smith-Gagen, Center for Health Data and Research

The relative survival rate is defined as the ratio of the observed survival rate in a population defined by a disease or other condition to the survival rate in a population that is similar with respect to survival in all factors related to survival but free of the disease or condition. The term survival rate does not refer to a rate as commonly used in the statistical literature but rather to a survival probability. Under the assumption of independent causes of death the relative survival rate is equal to the net survival as defined in the competing risk model of Chiang. Thus, one would expect the relative survival to be a non-increasing function of time. In practice, however, the relative survival rate from time $t=0$ to time $t=t^*$ can be observed to be non-increasing at first and then start to rise again. Reasons are related to the age distribution of the study population or the fact that the study population and comparison population may not have the same mortality risk from other causes. In this talk we will discuss these issues from the perspective of comparability of the study group and the control group. We will also present an application of relative survival to a study of rectal cancer.

email: cmdrake@ucdavis.edu

A LIKELIHOOD-BASED APPROACH TO QUANTIFICATION OF THE SPREAD OF AN INFECTIOUS DISEASE

Laura F. White*, Harvard School of Public Health
Marcello Pagano, Harvard School of Public Health

Increased interest in monitoring for bioterrorist events and outbreaks of novel infectious diseases has expanded the amount of data readily available for the monitoring of disease patterns. Current research focuses on exploiting this data for the purpose of rapidly detecting disease outbreaks. We propose to further exploit this data to quantify the severity of an outbreak using common epidemiological measures, such as the basic reproductive number, R0 and the serial interval. We show that information on the time of presentation with symptoms is sufficient to obtain initial estimates of these parameters using a likelihood based approach, leading to a more rapid and effective public health response.

email: lforsber@hsph.harvard.edu

MORE REALISTIC ASSUMPTIONS FOR CONTROLLING CONFOUNDING IN OBSERVATIONAL STUDIES OF TIME VARYING EXPOSURES

Marshall M. Joffe*, University of Pennsylvania

In observational assumptions, analysts typically assume that treatment assignment is ignorable. Robins (1986) has developed a sequential version of this assumption for longitudinal studies with time-varying treatments and exposures. Unfortunately, the assumption is unrealistic for most observational studies in which treatment and covariates are measured at regular intervals but treatment is not under control of the investigator. We develop a modification to the standard sequential ignorability assumption that is more appropriate for these settings. Further, we develop a method for estimation of parameters in structural nested distribution models that is valid under the revised assumption. We apply the methods to data from the Lipid Research Clinics Coronary Primary Prevention Trial.

email: mjoffe@cceb.upenn.edu

42. BAYESIAN METHODS IN GENOMICS DATA ANALYSIS

BAYESIAN ANALYSIS OF LOSS OF HETEROZYGOSITY BY MODELING OF FREQUENCY OF ALLELIC LOSS DATA

Hanwen Huang*, University of North Carolina at Chapel Hill
Fei Zou, University of North Carolina at Chapel Hill
Fred A. Wright, University of North Carolina at Chapel Hill

One objective of allelic-loss studies is to identify which, if any, chromosome arms that harbor tumor suppressor genes. Instability-selection model has been developed for allelic-loss data where allelotypes are available. For many allelic-loss experiments, only the frequency of allelic loss (FAL) rather than the allelotypes is available. Recently, an extended likelihood based instability-selection model has been proposed for FAL data. Due to the complexity of the likelihood function, it is difficult to obtain the maximum likelihood estimates. How to access the significance of likelihood ratio test is also unclear since the asymptotic distribution of the likelihood ratio statistic is not known. In this paper, an alternative Bayesian version of the extended instability-selection model approach was proposed. Adavantages of the Bayesian approach includes: 1)computations are simplified since missing data are imputed; 2)Bayes factors based on Savage's density ratio are readily available for hypothesis testing problem. Application of our Bayesian approach to four cancer studies yields similar tumor suppressor gene location estimates previously reported.

email: hhuang@bios.unc.edu

BAYESIAN HIERARCHICAL MODEL TO DETECT QTL

Susan J. Simmons*, University of North Carolina Wilmington
Edward Boone, University of North Carolina Wilmington
Ann E. Stapleton, University of North Carolina Wilmington

Most QTL algorithms and analysis assume that there is only one observation within each genotype. However, plant researchers tend to have a number of clones, or observations that are genetically the same within each genotype. Historically, plant researchers combine the information within each genotype to a single value (for example a mean or median), and then continue with the more conventional approaches. However, in doing so, the information about the variability within each genotype is lost. We developed a hierarchical model that incorporates the variability within each genotype and between the genotypes to identify QTLs. We illustrate our model on a simulated data set and on a real Arabidopsis data set.

email: simmonssj@uncw.edu

SEMIPARAMETRIC BAYESIAN INFERENCE FOR SAGE DATA-MODEL BASED CLUSTERING FOR COUNT DATA

Michele Guindani*, The University of Texas M.D. Anderson Cancer Center
Peter Mueller, The University of Texas M.D. Anderson Cancer Center

SAGE (serial analysis of gene expression) experiments record abundance of distinct mRNA tags in a given sample. Typically, more than 80% of the tags record low frequencies and many rare tags in the probe might not be actually observed in the sample. We propose a semi-parametric hierarchical Bayes approach to estimate the true abundance of each of the observed tags. In particular, we assume independent Poisson sampling with tag-specific Poisson rates, each arising as a sample from a Dirichlet process (DP) with a conjugate base measure. The a.s. discrete nature of the DP random measure implies a clustering of the tag indices according to their relative abundance. We also assume a random number of distinct tags in the probe, which allows a precise assessment of the true abundancies compared to the usual empirical estimators. We report inference obtained via MCMC simulation for the true tag frequencies, including shrinkage for tags with low empirical frequencies, honest accounting for uncertainties, and dependencies. Also, we provide posterior inference for the unknown number of distinct tags in the probe. We only analyze data under one biologic condition. Extension to a probability model for multi sample data is straightforward. We comment on appropriate model extensions and inference.

email: mguindani@mdanderson.org

A COMPOSITIONAL RETROSPECTIVE ANALYSIS OF MICROARRAY DATA

Jingqin Luo*, Duke University
Edwin S. Iversen, Duke University
Merlise A. Clyde, Duke University

In many studies, survival data involve several types of failure. When the effects of covariates of interest differ markedly for the different types, a type-specific analysis is important. Different approaches have High-dimensional Microarray data are ususally incorporated in cancer study to explain disease incidence etc. While modeling feature space of high dimensionality has always been a challenging issue, most statistical methods proceed analysis under various simplified assumptions. A practically successful method, naive Bayes classifier(NB) assumes that features are conditionally independent given class label. Besides accurate classification, exploration of feature dependency network and discovery of important features are also of great interest. Our compositional retrospective method achieves this goal by slightly relaxing NB's underlying assumption. Given a permutation order of all features, the method models the joint conditional distribution of the feature space as sequential product of a series of conditional probability distributions. In each conditional probability, a feature can rely on k other features (class label included), with k a user-specified number. Choice of such a feature subset can be decided via classic variable selection method if k is small and more aggressive variable shrinkage methods otherwise. A spectral ordering algorithm can be used to find an optimal permutation. The method will be illustrated by some simulated examples and real cancer datasets.

email: rosy@duke.edu

USING CLUSTERING TO ENHANCE HYPOTHESIS TESTING

David B. Dahl*, Texas A&M University
Michael A. Newton, University of Wisconsin-Madison

Both multiple hypothesis testing and clustering have been the subject of extensive research in high dimensional inference, yet these problems have traditionally been treated separately. We propose a hybrid statistical methodology that uses clustering information to increase testing sensitivity. A test for object i that uses data from all objects clustered with it will be more sensitive than one that uses data from object i in isolation. While the true clustering is unknown, there is increased power if the clustering can be estimated relatively well. We first consider a simplified setting which compares the power of the standard Z-test to the power of a test using an estimated cluster. Theoretical results show that if the cluster is estimated sufficiently well, the new procedure is more powerful. In the setting of gene expression data, we develop a model-based analysis using a carefully formulated conjugate Dirichlet process mixture model. The model is able to borrow strength from objects likely to be clustered. Simulations reveal this new method performs substantially better than its peers. The proposed model is illustrated on a large microarray dataset.

email: dahl@stat.tamu.edu

43. LONGITUDINAL DATA ANALYSIS

ROLE OF ALTERNATIVE STATISTICAL METHODS OF CD4 CELL COUNT EXTRAPOLATION IN QUANTIFYING HIV RELATED MORBIDITY

Bingxia Wang*, Boston University

Background: HIV morbidity (opportunistic infections (OI)) stratified by CD4 cell count is an important marker of HIV disease progression. CD4 cell counts are not measured at the time of OIs. The impact of misspecification of CD4 trajectory on the accuracy of estimation of OI incidence within different CD4 strata has not been studied. Objective: To estimate CD4-specific OI incidence rates (IR) using alternative methods of CD4 cell extrapolation, and to examine how misspecification of CD4 trajectories influences the accuracy of OI IR estimation. Methods: We performed a simulation study varying the "true" CD4 trajectories, models used to estimate trajectories and CD4 strata. We assumed several "true" CD4 cell trajectories: straight line, step functions, and elliptical quarter curves. We used random-effects model, step-function method, linear interpolation method, linear and natural cubic splines to extrapolate CD4 cell count at the time of OI. We calculated relative bias as a means of evaluating model accuracy. Results: When CD4 cell counts increase, IR increases and variability due to alternative model use increase. The bias ranged from 150% for random effect model to -22% for step-function method. Conclusions: The lower the CD4 cell counts the greater was the impact of CD4 misspecification. Among all considered, natural cubic spline performed the best.

email: bxwang@bu.edu

SIGNAL INTENSITY PROCESSING BASED ON NON-LINEAR MIXED MODELING TO STUDY CHANGES IN NEURONAL ACTIVITY

Jan Serroyen*, Hasselt University-Diepenbeek, Belgium
Geert Molenberghs, Hasselt University-Diepenbeek, Belgium
Marleen Verhoye, University of Antwerp-Antwerp, Belgium
Vincent Van Meir, University of Antwerp-Antwerp, Belgium
Annemie Van der Linden, University of Antwerp-Antwerp, Belgium

We analyze data on the impact of testosterone on the dynamics of Mn2+ accumulation measured by magnetic resonance imaging in three songbird brain areas: the nucleus robustus arcopallii (RA), area X, and the high vocal center (HVC). Birds with and without testosterone were included in the experiment, and repeated measurements were available in both a pre and post drug administration period. We formulate a non-linear modeling strategy, allowing for the incorporation of (1) within-bird correlation, (2) the non-linearity of the profiles, and (3) the effect of treatment. For two of the outcomes (RA and area X), biological theory suggests a parametric form, while for HVC this is not the case. Since the HVC outcome bears some resemblance with the two-compartment model known from pharmacokinetics, this model was considered a sensible choice. We use a different model, based on fractional polynomials, as a sensitivity analysis for the latter. All methods used provide good fits to the data, confirm results from previous, simple analyses undertaken in the literature, but were able to detect additional effects of treatment that had so far gone undetected. The fractional polynomial and two-compartment models provide similar substantive conclusions, the two together can be seen as a form of sensitivity analysis.

email: Jan.Serroyen@uhasselt.be

PREDICTION OF RENAL GRAFT FAILURE USING MULTIVARIATE LONGITUDINAL PROFILES

Steffen Fieuws*, K.U.Leuven, Belgium
Geert Verbeke, K.U.Leuven, Belgium

Patients who underwent renal transplantation are monitored longitudinally at irregular time intervals during 10 years. Each visit yields a set of biochemical and physiological markers containing valuable information to anticipate a failure of the graft. General linear, generalised linear or nonlinear mixed models are used to describe the longitudinal profiles of each marker. These univariate mixed models are joined into a multivariate mixed model by specifying a joint distribution for the random effects. Due to the high number of markers, a pairwise modelling strategy, where all possible pairs of bivariate mixed models are fitted, has been used to obtain parameter estimates for the multivariate mixed model. These estimates are used in a Bayes rule to calculate at each point in time the risk that the graft will fail in the remaining time of the 10-year period since the moment of transplantation.

email: steffen.fieuws@med.kuleuven.be

ROBUSTNESS IN JOINT MODELING OF A PRIMARY REGRESSION MODEL AND A LONGITUDINAL PROCESS

Xianzheng Huang*, North Carolina State University
Leonard Stefanski, North Carolina State University
Marie Davidian, North Carolina State University

Joint model is often used to link a primary regression model and a longitudinal process via shared random effects to characterize the association between a primary endpoint and a longitudinal profile. Parametric joint modeling can gain efficiency and provide insight into underlying feature of the longitudinal process. But these gains from joint modeling often rely on correct specification of the random-effect model. We present methods to diagnose model misspecification of the random effects when jointly modeling a primary regression model and a longitudinal process. The methods are illustrated via application to simulated data and data from a study of bone mineral density in peri-menopausal women.

email: doudouxz@yahoo.com

MULTIVARIATE HIDDEN MARKOV PROCESSES: APPLICATION TO CORONARY VASCULAR DISEASE PROGRESSION

Melanie M. Wall*, University of Minnesota
Judith Rousseau, Université Paris
Chantal Guihenneuc-Jouyaux, Université Paris 5

In the modeling of the relationships between several longitudinal processes, it is often of interest to consider changes across time of each variable and the influence that each process has on the others across time. If the processes themselves are best modelled as discrete stages and the measurements of these stages are not done directly, but instead via some observable variable or variables expected to measure the stage with error, then a hidden markov model may be appropriate where the time process and relationship among the variables is modeled directly on the underlying latent variables or latent stages. For modeling relationships among processes, a multivariate hidden markov model is developed. In the study of coronary heart disease, there are several different but related processes which can be considered to be changing across time within individuals. This current study examines the relationship among 3 different such longitudinal processes (heart function, physical impact, and quality of life) all hypothesized as discrete latent variables underlying observed clinical and self-report measurements.

email: melanie@biostat.umn.edu

FAILED CLINICAL TRIALS AND FAILED ANALYSES: NEW HELP FROM TRAJECTORY-BASED ANALYSES

Ralitza Gueorguieva*, Yale University
Ran Wu, Yale University
Brian Pittman, Yale University
Stephanie O'Malley, Yale University
John Krystal, Yale University

Clinical trials are often designed to repeatedly measure variables yet often analysis is restricted to one or more summary measures. For example, in alcohol research, drinking data are collected daily but only composite variables such as time to drinking episode and percent drinking days are considered as outcomes. By analyzing all available data, additional information can be obtained about the pattern of change over time and power for detecting treatment differences can be improved. Herein, we revisit two randomized clinical trials of naltrexone for alcohol dependence with negative results on the summary measures considered in the a priori specified analyses. We analyze daily drinking data using a semi-parametric group-based approach (Nagin, 1999). We identify distinct trajectories of response over time and estimate the effect of treatment on a subject's chance to belong to a particular trajectory. We observe that subjects on naltrexone are less likely than subjects on placebo to belong to a "heavy drinker" trajectory class. We will present simulations to investigate the performance of the trajectory method. [Supported by the Department of Veterans Affairs Cooperative Study Program, the Center for Translational Neuroscience of Alcoholism (P50 AA012870-05), and grants RO1-AA10225 and K05-AA014715.]

email: ralitza.gueorguieva@yale.edu

44. MEASUREMENT ERROR

AN ESTIMATING EQUATIONS APPROACH TO FIT LATENT EXPOSURE MODELS

Brisa N. Sánchez*, Harvard School of Public Health
Louise M. Ryan, Harvard School of Public Health

Classical approaches for structural equation or latent variable models typically require distributional assumptions on both observed and latent variables. While least square approaches relax distributional assumptions, strict variance structures are imposed on the marginal distribution of the data to ensure unbiased mean parameters. We propose an estimating equations approach for latent exposure models. Our approach is similar to a two-step regression on factor scores, but we jointly estimate all model parameters such that correct variance estimates can be produced. We show that parameter estimates from our method are unbiased; that compared to maximum likelihood, the loss of efficiency of our method is relatively small under correct model specification; and that in theory our method is robust to misspecification of the variance in the outcome. Further, we show that our method is a generalization of the regression calibration approach to measurement error modelling. We apply our methods to a study of the effects of in-utero lead exposure on child development.

email: bsanchez@hsph.harvard.edu

LOCALLY EFFICIENT ESTIMATORS FOR SEMIPARAMETRIC MODELS WITH MEASUREMENT ERROR

Yanyuan Ma*, Texas A&M University
Raymond J. Carroll, Texas A&M University

We derive constructive locally efficient estimators in semiparametric measurement error models. The setting is one where the likelihood function depends on variables measured with and without error, where one of the variables measured without error is modelled nonparametrically. The algorithm is based on backfitting. This problem includes the partially linear measurement error model as a special case. We show that if one assumes a parametric model for the latent variable measured with error and if this model is correct, then our estimators are semiparametric efficient. In contrast to standard problems, our methods enjoy the property that even if the latent variable model is misspecified, our methods still lead to consistent and asymptotically normal estimators. We illustrate the methods with a logistic regression problem where the latent variable measured with error is modelled by a quadratic regression, while a variable measured without error is modelled nonparametrically. We also apply the methodology to a data example where the putative latent variable distribution is a shifted lognormal, but concerns about the effects of misspecification of this assumption suggest the need for a more model-robust approach.

email: ma@stat.tamu.edu

STATISTICAL METHODS FOR MEASUREMENT COMPARISON IN CLINICAL STUDIES

Jing Han*, St. Francis Hospital

It is difficult to get the clinical data, such as cardiac stroke volume, by direct measurement without adverse effects. The true values remain unknown. Instead indirect methods are used, and a new method has to be evaluated by comparison with an established technique rather than with the true quantity. Therefore, it is common to study the agreement of measurements between two methods in clinical studies. But the correct statistical approach is not obvious. Methods of analysis used in the comparison of two methods of measurement are reviewed, including the graphical analysis, such as Bland Altman plot and mountain plot; and statistical analysis, such as, ordinary regression, Deming regression, Passing Bablok regression, and Kappa statistics.

email: hanjlst@gmail.com

OPERATING CHARACTERISTICS OF GROUP TESTING ALGORITHMS FOR CASE IDENTIFICATION 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
Jonathan Dreyfuss, University of North Carolina at Chapel Hill
Daniel J. Westreich, University of North Carolina at Chapel Hill
Christopher D. Pilcher, University of North Carolina at Chapel Hill

We derive and compare the operating characteristics of hierarchal and square array based testing algorithms for case identification in the presence of testing error. The operating characteristics investigated include efficiency (i.e., expected number of tests per specimen) and error rates (i.e., sensitivity, specificity, positive and negative predictive values, per-family error rate, and per-comparison error rate). The methodology is illustrated by comparing different pooling algorithms for the detection of individuals recently infected with HIV in North Carolina and Malawi.

email: hkim@bios.unc.edu

METHODS FOR COX REGRESSION WITH NON-CLASSICAL MEASUREMENT ERROR IN THE COVARIATES

Pamela A. Shaw*, University of Washington
Ross L. Prentice, Fred Hutchinson Cancer Research Center

We develop methods for the analysis of censored failure time data using Cox regression when the covariate of interest is measured with both subject-specific and systematic measurement error. Current methods addressing covariate measurement error in Cox regression have dealt largely with classical (unbiased) error without allowing for a systematic error component. Three parameter estimation techniques developed for the classical error model, riskset regression calibration (Xie, Wang, and Prentice JRSS Series B, 2001), conditional score (Tsiatis and Davidian, Biometrika, 2001), and non-parametric corrected score (Huang and Wang, JASA, 2000), will be extended to the case where the covariate available on the whole cohort follows a generalized measurement error model and a covariate measured with unbiased error is available on a subset of subjects. Diet assessment data from nutritional epidemiology are used to motivate this more flexible model. Asymptotic theory and variance estimators are provided for each of the new methods. Performance of the methods is then compared under different scenarios for the true error structure using a simulation study. Similar to the classical measurement error case, the relative success for these methods depends on the magnitude of the relative risk and on the distribution of the true error.

email: pshaw@u.washington.edu

ESTIMATION FOR GENERALIZED STRUCTURAL EQUATION MODELS WITHOUT NORMALITY ASSUMPTIONS ON THE CONTINUOUS FACTORS

Jia Guo*, University of Minnesota
Melanie M. Wall, University of Minnesota
Yasuo Amemiya, IBM-T.J. Watson Research Center

In the research of public health, psychology, and social sciences, many research questions involve continuous predictor variables which can not be observed directly, such as self-esteem, depression and body satisfaction. Instead, these latent variables (factors) are usually measured by a set of observable variables indirectly. A generalized structural equation model can be built to examine the relationship between these latent variables and certain outcomes either observed or unobserved. Typically, the continuous latent variables are assumed to follow a normal distribution and estimation procedes using full likelihood via the EM algorithm or using a fully Bayesian approach via MCMC methods. The focus of this talk is to examine estimation methods for this model that do not rely on strong normality assumptions for the distribution of the unobservable latent predictors. One is based on functional modelling the predictors by treating them as fixed. The other one is based on structural modelling by using mixtures of normals approximation.

email: jiaguo@biostat.umn.edu

MEASUREMENT ERROR IN POPULATION DYNAMIC MODELS

John P. Buonaccorsi*, University of Massachusetts
John Staudenmayer, University of Massachusetts

Measurement or observation error is common in measurements of animal abundance. In this talk we examine the effects of measurement error and how to correct for it in the random walk with drift model. This model which has been used to describe the behavior of a variety of species and is used in Population Viability Analysis. Analytical expressions are given for the bias in naive estimators (which ignore measurement error) under a very broad class of measurement error models, including allowance for changing measurement error variances over time. The impact of these biases on estimates of various parameters, including the growth rate and probability of extinction are demonstrated numerically. We then present moment and likelihood based methods to correct for measurement error, with and without the use of estimated measurement error variances. The methods are applied to existing population datasets and evaluated further via simulations.

email: johnpb@math.umass.edu

DETECTING LINKAGE DISEQUILIBRIUM IN THE PRESENCE TO LOCUS HETERGENEITY

Jian Huang*, University of Iowa
Deli Wang, University of Alabama at Birmingham

Locus heterogeneity is a common phenomenon in complex diseases and is one of the most important factors that affect the power of either linkage or linkage disequilibrium (LD) analysis. In linkage analysis, the heterogeneity LOD score (HLOD) rather than LOD itself is often used. However, the existing methods for detecting linkage disequilibrium, such as the TDT and many of its variants do not take into account locus heterogeneity. We propose two novel likelihood-based methods, an LD-Het likelihood and an LD-multinomial likelihood, to test linkage disequilibrium (LD) that explicitly incorporate locus heterogeneity in the analysis. The LD-Het is applicable to general nuclear family data but requires a working penetrance model. The LD-multinomial is only applicable to affected sib-pair data but does not require specification of a trait model. For affected sib-pair data, both methods have similar power to detect LD under the recessive model, but the LD-multinomial model has greater power when the underlying model is dominant or additive.

email: jian@stat.uiowa.edu

ANALYSIS OF COMPLEX TRAITS WITH ORDINAL OUTCOME DATA

Heping Zhang*, Yale University
Xueqin Wang, Yale University
Yuanqing Ye, Yale University

There is growing interest in genome-wide association analysis using single-nucleotide polymorphisms (SNPs), because traditional linkage studies are not as powerful in identifying genes for common, complex diseases. A variety of tests for linkage disequilibrium have been developed and examined for binary and quantitative traits. However, since many human conditions and diseases are measured in an ordinal scale, methods need to be developed to investigate the association of genes and ordinal traits. Thus, in the current study we propose and derive a score test statistic that identifies genes that are associated with ordinal traits when gametic disequilibrium between a marker and trait loci exist. Through simulation, the performance of this new test is examined for both ordinal traits as well as quantitative traits. The proposed statistic not only accommodates ordinal traits and have superior power for ordinal traits, but also has similar power of existing tests when the trait is quantitative. Therefore, our proposed statistic has the potential to serve as a unified approach to identifying genes that are associated with any trait, regardless of how the trait is measured.

email: heping.zhang@yale.edu

NONPARAMETRIC PATHWAY BASED REGRESSION METHODS FOR ASSESSING GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS

Hongzhe Li*, University of Pennsylvania

For many complex diseases, especially for cancers, there are many types of metadata available which are related to biological pathways. Currently, information derived from metadata such as known biological knowledge has hardly been utilized in the modelling step. We propose to develop and evaluate novel gradient descent boosting procedures for nonparametric pathways-based regression (NPR) analysis to efficiently integrate genomic data and metadata. The boosting methods for the NPR models can be used for studying interactions between continuous variables and binary variables by using different base procedures. The methods also provide an alternative of mediating the problem of a large number of potential interactions by limiting analysis to biologically plausible interactions between genes in related pathways. Simulation results and real data analysis will be presented.

email: hli@cceb.upenn.edu

FAMILY-BASED HAPLOTYPE STUDIES

Glen A. Satten*, Centers for Disease Control and Prevention
Andrew S. Allen, Duke University
Anastasios A. Tsiatis, North Carolina State University

Modeling human genetic variation is critical to understanding the genetic basis of complex disease. The Human Genome Project has discovered millions of binary sequence variants, called single nucleotide polymorphisms, and millions more may exist. As coding for proteins takes place along chromosomes, organization of polymorphisms along each chromosome (the haplotype phase structure) may prove to be important in discovering genetic variants associated with disease. As haplotype phase is often uncertain, procedures that model the distribution of parental haplotypes can, if this distribution is misspecified, lead to substantial bias in parameter estimates even when complete genotype information is available. Using a geometric approach to estimation in the presence of nuisance parameters, we address this problem and develop locally-efficient estimators of the effect of haplotypes on disease that are robust to incorrect estimates of haplotype frequencies. The methods are demonstrated with a simulation study of a case-parent design. We also consider estimation of haplotype x environment interaction effects and models based on identifying genomic regions containing risk alleles by comparing differences in haplotype sharing among transmitted and untransmitted haplotypes.

email: gas0@cdc.gov

46. DEALING WITH MISSING DATA: ARE WE THERE YET?

MISSING DATA IN INFECTIOUS DISEASE RESEARCH

Barbra A. Richardson*, University of Washington

Missing data in infectious disease research can occur for several reasons. These reasons include laboratory assay results not being available (due to failure of an assay or samples not being available to be tested), and low sensitivity/specificity of diagnostic tests used in field research settings. Examples of problems encountered with missing data in infectious disease research will be presented. These examples include problems estimating the rate of mother-to-child transmission of HIV-1, and problems estimating the incidence of STDs using a diagnostic test with $<100\%$ sensitivity. The effects of such missing data on the interpretation of analyses using standard statistical methodology will be discussed, and more recent developments of statistical methodology to address biases induced by the missing data in these particular situations will be reviewed.

email: barbrar@u.washington.edu

STRATEGIES FOR MISSING PATIENT REPORTED OUTCOMES

Diane L. Fairclough*, University of Colorado Health Sciences Center

Missing data are inevitable when studying individuals with significant morbidity and mortality. In many cases, the missing data is believed to be related to negative aspects of treatment (toxicity) or the disease (progression) and thus likely to be non-ignorable. As the specific dropout mechanism can not be determined, sensitivity analyses are recommended. Among the methods most commonly used are simple forms of imputation such as last value or worst value carried forward, multiple imputation, pattern mixture models and shared parameter models. Three examples will be used to show that some of these strategies may not be feasible or adequately address the expected non-ignorable process. Without a good understanding of the likely patterns of change over time and dropout, developing detailed analysis plans priori to the study will be difficult. Recommendations include the identification auxiliary or surrogate data that can be used to convert the problem to ignorable conditional on the observed outcomes and the supplementary information and the identification of alternative restrictions for pattern mixture models.

email: diane.fairclough@uchsc.edu

SENSITIVITY ANALYSIS FOR INCOMPLETE CLINICAL TRIAL DATA

Geert Molenberghs*, Universiteit Hasselt-Diepenbeek, Belgium

Clinical trial data, especially when longitudinally collected, is often incomplete due to dropout. Whereas incompleteness was traditionally dealt with through such methods as last observation carried forward or complete case analysis, currently a paradigm shift towards missing at random (MAR) based methods is taking place. Thanks to direct likelihood, multiple imputation, Bayesian methods, and weighted estimating equations, for example, and the availability of flexible software, such methodology is within reach. Of course, methods within this paradigm do pose non-trivial issues since not all results, known for complete, balanced data, carry over to the incomplete setting. This has implications for model selection, model criticism, etc. Furthermore, missing not at random (MNAR) mechanisms are difficult to rule out on purely statistical grounds and hence appropriate forms of sensitivity analysis should be conducted to strengthen one's confidence in the conclusions reached with the main model.

email: geert.molenberghs@luc.ac.be

THE COMPLEXITIES, COMPLICATIONS AND CONTRIBUTIONS OF DEALING WITH MISSING DATA IN CLINICAL TRIALS

Ralph B. D'Agostino, Sr.*, Boston University
Joseph M. Massaro, Boston University
Lisa Sullivan, Boston University

Clincal trials are beset with the problem of missing data. Many techniques have been suggested for dealing with missing data ranging from ignoring them to using last observation carried forward to multiple imputation to performing sensitivity analyses. In this talk we discuss, by way of examples, the complexities and complications that generate missing data and trying to understand how they arise, how to deal with them and the interpretation of results when they are considered. We also review a number of contributions that have been produced to deal with them. In addition we review some of our experiences and procedures that we have found effective for dealing with them.

email: ralph@bu.edu

47. STATISTICAL ISSUES IN ENVIRONMENTAL MONITORING

STATISTICS ISSUES IN DESIGNING AN OPTIMAL DETECTION SYSTEM WITH MULTIPLE SENSORS

Carol Y. Lin*, Emory University
Lance A. Waller, Emory University
Robert H. Lyles, Emory University
Barry P. Ryan, Emory University

Combining individual tests or sensors in a detection system often improves diagnostic performance. Optimal decision-theoretic combinations exist. However, when designing a detection system, cost effectiveness is important, particularly when combining a set of sensors with heterogeneous individual cost and performance. We consider the expected cost of correct and incorrect decisions of the system, constrained by the budget available to select an optimal system from systems with various combinations of different types of sensors. We further examine system performance with various numbers of sensors and allow correlations between individual sensors. The results quantify how increasing the number of sensors increases the combined probability of detection (sensitivity) and decreases the false alarm rates (1-specificity), and how increasing correlation of individual sensors decreases the combined probabilities of detection increases the false alarm rate. Furthermore, we illustrate the magnitude of decline associated with correlation increases with the number of individual sensors. To illustrate, we consider a hypothetical network of air pollution monitors in Boston consisting of both expensive and accurate sensors as well as inexpensive and less accurate sensors.

email: cylin@emory.edu

OPTIMAL NETWORK DESIGN FOR SPATIAL PREDICTION, COVARIANCE PARAMETER ESTIMATION, AND EMPIRICAL PREDICTION

Dale L. Zimmerman*, University of Iowa

Inferences for spatial data are affected substantially by the spatial configuration of the network of sites where measurements are taken. In this talk, criteria for network design that emphasize the utility of the network for prediction of unobserved responses assuming known spatial covariance parameters are contrasted with criteria that emphasize the estimation of the covariance parameters themselves. It is shown, via a series of related esamples, that these two main design objectives are largely antithetical and thus lead to quite different 'optimal' designs. Furthermore, a hybrid design criterion that accounts for the effect that the sampling variation of covariance parameter estimates has on prediction is described and illustrated.

email: dzimmer@stat.uiowa.edu

ESTIMATION FOR LONITUDINAL SURVEYS WITH REPEATED PANELS OF OBSERVATIONS

Jason C. Legg*, Iowa State University
Wayne A. Fuller, Iowa State University
Sarah M. Nusser, Iowa State University

We consider a longitudinal study composed of a first-phase sample with multiple subsets of these units selected for observation over time. One such design is used for the National Resource Inventory, where a core panel of segments is observed yearly and annual supplements are selected using a rotation design. Since observations are taken over time, there is a dependency in the data that can be exploited in estimation. We use an estimated generalized least squares (EGLS) approach that utilizes the estimated time-dependency to improve estimation of level and change relative to direct survey estimators. We discuss estimating an appropriate variance matrix for two-phase longitudinal data for use in an EGLS estimator. Since longitudinal studies often involve a large number of variables and the output of such studies is a dataset with weights for end users, we provide a consistent jackknife replication variance method for our EGLS estimator. This approach relies on having a consistent jackknife variance estimator for the first-phase sample. The National Resource Inventory will serve as the motivating example for this work.

email: ‘jlegg@iastate.edu

SPATIAL LASSO WITH APPLICATION TO GIS MODEL SELECTION

Jay Breidt*, Colorado State University
Nan-Jung Hsu, National Tsing-Hua University
Hsin-Cheng Huang, Academica Sinica
Dave Theobald, Colorado StateUniversity

Geographic information systems (GIS) organize spatial data in multiple two-dimensional arrays called layers. In many applications, a response of interest is observed on a set of sites in the landscape, and it is of interest to build a regression model from the GIS layers to predict the response at unsampled sites. Model selection in this context then consists not only of selecting appropriate layers, but also of choosing appropriate neighborhoods within those layers. We formalize this problem and propose the use of Lasso to simultaneously select variables, choose neighborhoods, and estimate parameters. Spatial smoothness in selected coefficients is incorporated through use of a priori spatial covariance structure, and this leads to a modification of the Lasso procedure. The LARS algorithm, which can be used in a fast implementation of Lasso, is also modified to yield a fast implementation of spatial Lasso. The spatial Lasso performs well in numerical examples, including an application to prediction of soil moisture.

email: jbreidt@stat.colostate.edu

48. ILS: INTRODUCTION ON DATA MINING AND ITS RECENT DEVELOPMENT

CLASSIFICATION WITH SUPPORT VECTOR MACHINE AND PSI-LEARNING

Yufeng Liu*, University of North Carolina

Classification for high dimensional low sample size data becomes more and more pertinent in bioinformatics. Among many different classification methods, margin-based techniques are generally expected to yield good performance. In this talk, I will give an overview of margin-based classifiers including methods of SVM and Psi-learning. Different aspects of these methods including multicategory classification, variable selection, implementation, and applications will be discussed.

email: yfliu@email.unc.edu

FEATURE SELECTION AND CLUSTERING FOR HIGH DIMENSIONAL DATA

Xiaodong Lin*, University of Cincinnati

High dimensional data are regularly generated from various sources. Thus subsets of data are usually concentrated in different feature subspaces. In this talk, I will give an overview of techniques that achieve simultaneous feature selection/ dimension reduction and clustering. Methods based on finite mixture models, support vector machine and principle component analysis will be discussed. I will also present a new approach based on a constrained version of mixture of factor analyzers to achieve this goal. The performance of these methods will be illustrated using both synthetic and real datasets.

email: linxd@math.uc.edu

HIGH DIMENSIONAL, LOW SAMPLE SIZE DATA ANALYSIS: DATA PILING AND GEOMETRIC REPRESENTATION

Jeongyoun Ahn*, University of North Carolina at Chapel Hill
Steve Marron, University of North Carolina at Chapel Hill

In classification with High Dimension Low Sample Size data, there exists a one-dimensional direction in the data space such that the projected data have only two distinct values. I proved that this direction is uniquely determined in the data space and exists within the affine set of the data, but is orthogonal to the affine sets of each class. It is also seen that (1) the direction maximizes not only the amount of i°data pilingi± but also the distance between the piling sites; (2) it has a similar formula to the Fisheri¯s linear discrimination (FLD) direction and is shown to be equivalent to FLD in non-HDLSS cases; (3) it is heuristically a desirable classification direction when the data are in the aforementioned HDLSS geometrical limit. This geometric representation says that the randomness of the data with an extremely high dimensionality and small sample size only lies in random rotations of a regular simplex.

email: jyahn@email.unc.edu

VISUALIZATION CHALLENGES IN INTERNET TRAFFIC RESEARCH

Cheolwoo Park*, University of Georgia
Barbara Gonzalez, University of Louisiana at Lafayette
Felix Hernandez-Campos, University of North Carolina at Chapel Hill
Steve Marron, University of North Carolina at Chapel Hill

This is an overview of some recent research, and of some open problems, in the visualization of large data sets with Internet traffic applications. One challenge comes from the sheer scale of the data, where millions (and far more if desired) of observations are frequently available. Another challenge comes from ubiquitous heavy tail distributions, which render standard ideas such as i°random sampling will give a representative samplei± obsolete. An alternate sampling approach is suggested and studied.

email: cpark@stat.uga.edu

SEMIPARAMETRIC NORMAL MODELS FOR MULTIVARIATE SURVIVAL DATA IN FAMILY STUDIES

Malka Gorfine*, Technion Institute-Israel
Ross L. Prentice, Fred Hutchinson Cancer Research Center
Li Hsu, Fred Hutchinson Cancer Research Center

In this talk we present computational feasible estimating procedures for jointly estimating the marginal regression coefficients and the dependence parameters under a multivariate normal distribution of the transformed cumulative hazard function. Under this model, the failure times are first transformed into normal random variates marginally and then the transformed variates are assumed to follow a multivariate normal distribution with the covariance matrix accounting for the correlations among cluster members. In contrast to other copula models, the multinormal model allows unrestricted pairwise dependence parameters among the cluster members. Two sampling scheme are considered. A prospective study which requires following disease-free individuals over time in order to observe disease incidence, and a retrospective case-control design consists of a random sample of independent diseased individuals and non-diseased individuals, along with their family members.

email: lih@fhcrc.org

NONPARAMETRIC ASSOCIATION ANALYSIS OF MULTIVARIATE COMPETING RISKS DATA

Yu Cheng*, University of Wisconsin-Madison
Jason P. Fine, University of Wisconsin-Madison
Michael R. Kosorok, University of Wisconsin-Madison

While nonparametric analyses of bivariate failure times have been widely studied, nonparametric analyses of bivariate competing risks data have not been investigated. Such analyses are important in familial association studies, where multiple interacting failure types may invalidate nonparametric analyses for independently censored clustered survival data. We develop nonparametric estimators for the bivariate cause-specific hazards function and the bivariate cumulative incidence function, which are natural extensions of their univariate counterparts and make no assumptions about the dependence of the risks. The estimators are shown to be uniformly consistent and to converge weakly to Gaussian processes. Summary association measures are proposed and yield formal tests of independence in clusters. The estimators and test statistics perform well in simulations with realistic sample sizes. Their practical utility is illustrated in an analysis of dementia in the Cache County Study.

email: yucheng@wisc.edu

NEW DEVELOPMENTS IN THE ANALYSIS OF FAMILIAL AGGREGATION

Rebecca A. Betensky*, Harvard School of Public Health

Family studies are of interest for assessing familial aggregation of disease. As families are typically not randomly sampled, the ascertainment process must be taken into account in the analysis. This talk will propose a joint modeling approach to handle complex ascertainment events. It will also propose an exact test for analysis of familial aggregation, as well as a computationally tractable random effects model. The methods will be applied to data from the NCI sponsored Cancer Genetics Network.

email: betensky@hsph.harvard.edu

50. GRAPHICAL ANALYSIS AND REPORTING OF CLINICAL SAFETY AND EFFICACY DATA

GRAPHICAL APPROACHES TO THE ANALYSIS OF SAFETY DATA FROM CLINICAL TRIALS

Ohad Amit*, GlaxoSmithKline
Lane W. Peter, GlaxoSmithKline
Shi-tao Yeh, GlaxoSmithKline
Richard Heiberger, Temple University

Patient safety has always been a primary focus in the development of new pharmaceutical products. The predominant method for statistical evaluation and interpretation of safety data collected in a clinical trial is based on the use of descriptive statistics generally displayed in tabular form. There are several opportunities to enhance evaluation of drug safety through the use of graphical displays. Graphical displays can convey concisely multiple pieces of information more effectively than can be done in tabular form. They can be used in an exploratory setting to assist in the identification of emerging safety signals or in a confirmatory setting as a tool to explain and elucidate known safety issues. We developed several graphical displays for routine safety data collected during the course of a clinical trial, covering a broad range of graphical techniques. The graphical displays were developed with a focus on key clinical trial safety endpoints including QTc interval , endpoints assessing hepatotoxicity and adverse events of special interest. Statistical and graphical principles underlying the production and interpretation of the displays are discussed in detail.

email: ohad.amit@gsk.com

VISUAL REPRESENTATIONS OF DATA USED DURING THE NDA REVIEW CYCLE

Mat Soukup*, Food and Drug Administration

The Prescription Drug User Fee Act (PDUFA) provides a deadline for the review of a New Drug Application (NDA), 10 months for a standard review and 6 months for a priority review. With such restraints on time, it is important for the statistician to relay pertinent information to the medical review team and be able to do so in a short amount of time. This talk will present a number of graphical techniques that are used within the PDUFA review cycle that enable both the medical and statistical teams to visualize the vast amounts of data in an NDA. Specifically, the presentation will demonstrate a couple of examples where Trellis graphics and S-Plus Graphlets are used over traditionally long tables of output. Such methods are useful in conveying vast amounts of information with relative ease while also reducing the amount of time needed to understand relationships and structures in the data when looking at tabled output.

email: Mat.Soukup@fda.hhs.gov

NOW LOOK AT THIS: CONCEPTS FOR VISUALIZING CLINICAL DATA

Andreas Krause*, Pharsight Corporation

The presentation illustrates concepts and examples of standard clinical data graphs. The aim is to provide an insight into conceptualized approaches alongside with case studies that can easily be implemented in practice. The clinical development environment starts to realize that graphs are analysis tools in their own right, not merely a display of what has been calculated in numbers beforehand. The prime example, so-called Trellis or lattice graphs, is introduced together with clinical examples that highlight the analytical capabilities of the methodology of conditional graphs. The case studies demonstrate that it is indeed important to think about the question at hand and the data structure to select the appropriate type of display. Colors will be used to enhance the understanding of the data structure. Finally, examples are provided to illustrate how graphs can actually obfuscate the information in the data.

email: andreas@elmo.ch

THE STATE OF DATA VISUALIZATION IN THE REPORTING OF CLINICAL RESULTS

Matthew D. Austin*, Amgen, Inc

Statistics has evolved rapidly because of improved computational resources. Along with the strides in computational statistics, great improvements have been realized in visualization techniques that were fostered by the recent technological revolution. However, the reporting of data from clinical trials has remained almost unchanged over this time period. Graphical presentation of clinical results in publications and in professional meetings is still dominated by a simple point estimate for the mean value or percentage of responders and associated error bars to compare treatment groups in a clinical trial. The reviewer of this information must believe that the given summary statistic encompasses all the information that is needed to make an informed decision about the outcome of the clinical trial. However, with very little work more informative displays can be utilized that will better characterize results of clinical research. This presentation will focus on alternative, easily interpreted displays for some of the most common visualization techniques used in clinical research. As well as simply presenting the alternative techniques, focus will be given to teaching non-statisticians how to interpret the techniques and explaining how and why they present a better overall picture of the data.

email: maustin@amgen.com

STATISTICAL GRAPHICS IN DRUG DISCOVERY AND DEVELOPMENT

Michael A. O'Connell*, Insightful Corporation

With current trends of diminishing drug pipelines, safety concerns, and lengthy drug development processes; there is a critical need for efficiencies in drug discovery and development. The use of concise, compelling and standardized graphical analyses can help create such efficiencies. Statistical graphics play a crucial role in the interpretation of drug discovery and clinical development data. Applications include analyzing high-throughput drug discovery data, reporting clinical trial efficacy and/or safety data, evaluating mixtures of ingredients in product formulation, examining defect rates in manufacturing, or reviewing monthly sales data. In all of these situations a concise statistical graphic provides a far more efficient means to making scientific and business decisions than paging through realms of tables and listings. This session includes presentations from key areas in drug discovery and clinical development by thoughtleaders from the FDA, the pharmaceutical industry and the statistical graphics community. The format of the session is five speakers and one discussant; my talk will provide the discussion.

email: moconnell@insightful.com

MAKING CLEAR, CONCISE BUSINESS DECISIONS!

Tom Filloon*, Procter & Gamble

The need for rapid, clear, concise reporting of data in a portable display has never been more crucial. Effective statistical graphics (i.e, scientific visualization) are quintessential for keeping up in today's competitive business and scientific/research environments. A concise statistical data display is a far more efficient means of making business decisions than trying to have people come to the same conclusion after reviewing data in a tabular format. Various case examples will be shown from the areas of formulation, manufacturing, sales and clinical safety & efficacy.

email: filloon.tg@pg.com

51. RANDOM EFFECTS AND FRAILTY MODELS

ASSESSING THE EFFECTIVENESS OF POTENTIAL LONGITUDINAL BIOMARKERS IN MULTIVARIATE SURVIVAL ANALYSIS

Feng-shou Ko*, University of Pittsburgh Graduate School of Public Health
Stewart J. Anderson, University of Pittsburgh Graduate School of Public Health

We develop an extension of a method proposed earlier by Henderson et al. (2002) which combines the analysis of longitudinal and time to event information. In our development of the joint likelihood function, we incorporated a frailty parameter into a semi-parametric survival model. To compare our method to the proposed by Henderson et al., we performed simulations to assess the power for detecting longitudinal biomarkers. In our simulations, three latent processes were generated similar to those used in their work. Our results were: 1) higher correlations between the longitudinal biomarker values and survival time functions were associated with higher values for the power of score test; 2) for equal sample sizes, the power of a score test for relatively large numbers of subjects and small numbers of time points was higher than relatively small numbers of subjects and large numbers of observed time points; and 3) the power associated with our method was somewhat higher than that in Henderson, et al. To further compare our method to that of Henderson et al., we analyzed the liver cirrhosis data set presented in their paper. Our results were similar to theirs. Both methods showed that prothrombin is an effective surrogate for survival in liver cirrhosis patients.

email: fek4@pitt.edu

SEMIPARAMETRIC ANALYSIS OF CORRELATED RECURRENT AND TERMINAL EVENTS

Yining Ye*, University of Michigan
John D. Kalbfleisch, University of Michigan
Douglas E. Schaubel, University of Michigan

In clinical studies and observational studies, recurrent event data (e.g. hospitalization) with a terminal event (e.g. death) are often encountered. In many instances, the terminal event is strongly correlated with the recurrent event process. In this article, we propose a correlated Cox model to jointly study 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 and a Poisson model is used to estimate the dispersion of the frailty variable. A sandwich estimator is used to achieve additional robustness. An analysis of hospitalization data for patients in the peritoneal dialysis study is presented to illustrate the proposed method.

email: yey@umich.edu

APPLICATION OF RECURRENT EVENT DATA ANALYSIS METHODOLOGIES IN CLINICAL TRIAL STUDIES

Xiaohong Zhang*, Iowa State University
Matt Austin, Amgen, Inc.
Li Chen, Amgen, Inc.

Several methods have been proposed that are extensions of Cox's proportional hazard model to handle recurrent event data. In addition to the original papers, several papers have presented comparisons of the different methods. This presentation will contain a comparison of stratified and non-stratified extensions as well as extensions for mixed effects (frailty) models. Specific methods include the methodology proposed by Anderson-Gill (AG), the stratified methods of Prentice-Williams-Peterson (PWP) and mixed model extensions of the AG and PWP methodology. Results from a simulation study will be presented comparing the methods power, bias and variance. The comparison will be made for a simulated clinical trial where the incidence rate of the event of interest is 30% with 50% of subjects who experience one event having multiple events. The original sample size for the hypothetical trial is based on a 40% reduction in the incidence of the event of interest, and the question of interest will be if the treatment reduction is less than expected (30% reduction) which of the methods will provide the highest power to detect a treatment difference. Focus will be on why certain methods provide more power (bias, reduction in variability) in this situation.

email: xhzhang@iastate.edu

SHARED FRAILTY MODELS FOR GROUPED MULTIVARIATE SURVIVAL DATA

Denise A. Esserman*, Columbia University
Andrea B. Troxel, University of Pennsylvania

We expand frailty models to the grouped multivariate survival setting, with the specific goal of application to quality of life data. We use the univariate normal distribution to model the random frailty parameters. Via simulations, we study the effects of estimating covariates when the frailty variance is both fixed and estimated using maximum likelihood and residual maximum likelihood methods. In addition, we explore the impact of mis-specifying the frailty distribution.

email: dae2001@columbia.edu

A MODEL-BASED MEASURE OF INTER-RATER AGREEMENT

Kerrie P. Nelson*, Max Planck Institute for Demographic Research
Don Edwards, University of South Carolina

The issue of inter-rater agreement arises in many different situations, whenever two or more raters subjectively classify one or more items according to a classification scale. For example, skeletons of unknown age can be classified by an osteologist into an age-group based upon his/her examination of various features of the skeleton. A number of simple summary statistics are available to describe the agreement present in such a dataset, including Cohen's kappa and intra-class correlation coefficients. Despite a loss of information, and various biases that can occur, Cohen's kappa remains a very popular measure of agreement by many professionals. In this talk, we propose a model-based kappa-like statistic which is based upon the generalized linear mixed model with crossed random effects, with a binary classification. The statistic adjusts for important covariates that may influence agreement, and also chance-agreement, while allowing for a flexible number of raters and items. General inference regarding the underlying populations of raters and items is also provided. Applications of the use of this statistic and its corresponding interpretation will be given.

email: kerrie@stat.sc.edu

LONGITUDINAL, MULTIVARIATE TRAJECTORY MODELS FOR ESTIMATING AMERICAN DISABILITY

Jason T. Connor*, Carnegie Mellon University
Stephen A. Fienberg, Carnegie Mellon University
Daniel S. Nagin, Carnegie Mellon University

We compare two forms of multivariate mixture models to describe latent subclasses of frailty in a longitudinal, population based sample of American senior citizens. Data: The National Long Term Care Survey longitudinally measures disability in American seniors in 1982, 1984, 1989, 1994, and 1999. We use 3447 subjects, a subset of subjects with data in at least 4 waves. Methods/Results: We marginally fit mixture models for each of 7 disability measures (3 mixtures per disability), then of the 2,187 (3 ^ 7) possible combinations of latent strata, determine that 70% of seniors' multivariate patterns can be described by just 20 latent groups. We illustrate these common frailty patterns. Next we estimate latent disability groups simultaneously across disabilities and report the predominate disability groups in the American population. Finally we compare these two models for identifying latent patterns of disability in seniors. Conclusions: Trajectory models are useful for data reduction in multivariate, longitudinal datasets with a large number of latent subpopulations, and for clearly illustrating common multivariate disability patterns over time.

email: jconnor@stat.cmu.edu

MULTIPLE COMPARISONS WITH SEVERAL METHODS

Jixiang Wu*, Mississippi State University
Johnie N. Jenkins, USDA-ARS-Mississippi State
Jack C.. McCarty, USDA-ARS-Mississippi State

Multiple-comparison among different treatments or parameters is an important issue in applied statistics. Usually, this is conducted by least significant difference (LSD) method, Turkey's method, and Duncan's method, etc. However, these methods require the assumption of homogeneousity of variances for different treatments or estimated parameters. With the heterogeneous variances the confidence interval (CI) method and Paterson's method (1939) may be applied. In this presentation, we will compare different methods for multiple comparisons in terms of testing powers under various assumptions. Such results should be helpful for comparing predicted genetic effects or treatment effects in conjunction to mixed linear model approaches and resampling methods. An example of cotton data will be used to show the comparisons using different methods.

email: jw7@ra.msstate.edu

52. ANALYZING HIGH DIMENSIONAL GENOMICS DATA

CHARACTERIZING THE GENETIC STRUCTURE FROM SINGLE-NUCLEOTIDE POLYMORPHISM DATA

Xi Chen*, North Carolina State University
Bruce S. Weir, University of Washington

The need to characterize the genetic structure of human populations has increased with recent large-scale disease association studies. The accurate population specific estimation of Wright's Fst is necessary,especially when pupulations are dependent. A new approach based on GEE method can provide more robust estimtes comparing with moment method. The results of simulation will be shown on different population genealogies, along with some estimates for very dense SNP maps.

email: xchen@stat.ncsu.edu

ANALYSIS METHODS FOR ILLUMINA DASL DATA

Karla V. Ballman*, Mayo Clinic College of Medicine

Illumina's cDNA-mediated Annealing, Selection, extension and Ligation (DASL) assay is designed to generate RNA profiles from degraded tissue samples such as formalin-fixed paraffin-embedded (FFPE) tissues. The DASL assay uses a set of three gene-specific oligos, each probing a unique sequence, per gene. These query oligos span about 50 bases, allowing partially degraded RNAs to be used in the assay. Illumina has software that combines the expression intensities of the three probes to produce one expression value per gene. Using data from paired frozen tissue samples and FFPE samples (i.e. frozen and FFPE samples from the same tissue), both run with the DASL assay, the expression values produced from the Illumina software are compared to expression values produced by other potential methods. The goal is to determine which method produces expression values for the FFPE tissues closest to those from the frozen tissues. In addition, the tissue comes from two different groups (tumor versus normal) allowing us to compare the methods in terms of which one produces a list of differentially expressed genes (based on the FFPE tissue) most like the one produced using frozen tissue.

email: ballman@mayo.edu

A PSEUDOLIKELIHOOD APPROACH FOR SIMULTANEOUS ANALYSIS OF ARRA COMPARATIVE GENOMIC HYBRIDIZATIONS (ACGH)

David A. Engler*, Harvard University
Gayatry Mohapatra, Massachusetts General Hospital
David N. Louis, Massachusetts General Hospital
Rebecca A. Betensky, Harvard University

DNA sequence copy number has been shown to be associated with cancer development and progression. Array-based Comparative Genomic Hybridization (aCGH) is a recent development that seeks to identify the copy number ratio at large numbers of markers across the genome. Due to experimental and biological variations across chromosomes and across hybridizations, current methods are limited to analyses of single chromosomes. Moreover, it is often difficult to accurately identify regions of copy number gain and loss from the results of many current methods. We propose a more powerful approach that borrows strength across chromosomes and across hybridizations. We assume a Gaussian mixture model, with a Markovian dependence structure, and with random effects to allow for intertumoral variation, as well as intratumoral clonal variation. For ease of computation, we base estimation on a pseudolikelihood function. The method produces quantitative assessments of the probability of genetic alterations at each clone, along with a graphical display for simple visual interpretation. We assess the characteristics of the method through simulation studies and through analysis of a brain tumor aCGH data set. We show that the pseudolikelihood approach is superior to existing methods in detecting regions of copy number alteration.

email: engler@fas.harvard.edu

HIGH DIMENSIONAL PHENOTYPE SCORING OF HAND OSTEOARTHRITIS DATA USING EXPLORATORY MULTIVARIATE ANALYSIS

Sergio Eslava*, GlaxoSmithKline
Kwan R. Lee, GlaxoSmithKline
Keith Crowland, GlaxoSmithKline
Uzma Atif, GlaxoSmithKline

Objective. To effectively summarize/score a set of X-Ray measurements related to Hand Osteoarthritis phenotypes from the GOGO (Genetics of Generalized Osteoarthritis) Project, so the scores can be used as phenotype summary statistics for linkage analysis. Methods. Data was adjusted for covariates (Age, Sex and BMI) and then normalized. Several multivariate methods were used in the dimensionality reduction, including 3 Unsupervised (PCA, FA and VARCLUS) and 2 Supervised (PLS and PLS-DA) methods. The results were compared to select the best method. Variable Clustering (VARCLUS) was selected because it provided the best interpretability. Results. VARCLUS models revealed a clear separation of the distinct joint groups in the hand into different components, confirming the need to assess each joint as a unique phenotype. The measures that showed to better explain the variation in the model were KL grade, JSN, osteophytes, sclerosis and cysts. Conclusion. Multivariate Analysis methods are a reliable way to reduce the dimensionality of large and complex phenotypic datasets by generating scores that summarize the multidimensional space into a more manageable number of variables. Performing linkage analysis using the results from MVA can lead to a better understanding of the disease and its underlying genotype.

email: sergio.eslava@gsk.com

A WAVELET-BASED APPROACH TO CLUSTERING TIME-DEPENDENT GENE EXPRESSION PROFILES

Bong-Rae Kim*, University of Florida
Ramon C. Littell, University of Florida
Rongling Wu, University of Florida

Currently available analysis of time-dependent gene expression data has been limited to the characterization of genes and arrays with similar expression patterns by using clustering approaches, with no consideration of the developmental mechanisms underlying gene expression. In this talk, we will present a general mixture model for cataloguing time-dependent gene expression profiles in which the temporal pattern of gene expression is modeled by Fourier series approximations and the time-dependent covariance matrix structured by autoregressive or antedependence models. We implement the idea of wavelet dimension reduction into the mixture model for gene clustering, aimed to de-noise the data by transforming an inherently high-dimensional biological problem to its tractable low-dimensional representation. As a first attempt of its kind, we capitalize on the simplest Haar wavelet shrinkage technique to break an original signal down into spectrum by taking its averages and differences and, subsequently, to detect gene clusters that differ in the smooth coefficients extracting from noisy time series gene expression data. This wavelet-based model will have many implications for addressing biologically meaningful hypotheses at the interplay between gene actions/interactions and developmental pathways in various complex biological processes or networks.

email: bkim@stat.ufl.edu

STATISTICAL PERFORMANCE OF CLADISTIC STRATEGIES FOR HAPLOTYPE GROUPING IN PHARMACOGENETICS

Jared K. Lunceford*, Merck Research Laboratories
Nancy Liu, Merck Research Laboratories

Haplotypes at multiple single nucleotide polymorphisms (SNPs) are increasingly popular covariates for capturing the key genetic variation present over a region of interest in the DNA sequence. Though haplotypes can provide a clearer assessment of genetic variation in a region than their component SNPs considered individually, the multi-allelic nature of haplotypes increases the complexity of their role as dependent variables in statistical models intended to discover association with outcomes of interest. Haplotype grouping (cladistic) methods that cluster extant haplotypes according to estimates of genealogical closeness have been proposed recently as approaches for reducing model complexity and increasing power. Two such approaches to using closeness among sample haplotypes to direct and simplify testing are methods proposed by Seltmann et al. (2003) and Durrant et al. (2004). Using genotype data gathered from diabetic patients participating in clinical trials as a guide, we have conducted a simulation-based investigation of the performance of these two methods in the context of testing for pharmacogenetic interactions in candidate genes for which high-density SNP data have been gathered.

email: jared_lunceford@merck.com

53. DIAGNOSTIC AND SCREENING TESTS

ACCURACY OF BIOMETRIC AUTHENTICATION/IDENTIFICATION SYSTEMS

Peter B. Imrey*, The Cleveland Clinic Foundation

"Biometric" authentication/identification couples sensing instrumentation with algorithms for image or other signal comparisons, in semi-automated systems for i) authentication: confirming physical identity by a current signal's similarity to one obtained from the claimed person previously, or ii) identification: linking an unknown person's signal to previously stored signals with linked social identifiers (e.g., name). Popular modalities include fingerprints, iris scans, hand geometry, and dynamic signatures. These systems are semi-automated since humans deal with signal acquisition problems and contested decisions. Biometric systems are being implemented, ostensibly for security protection, in access control settings both large (passports and drivers licenses) and small (fingerprint unlocking of cellular phones). Instrumentation vendors, government agencies, commercial third-party testers, and academics all promulgate accuracy assessments. How are these assessments arrived at? How informative are they in policy formulation, such as when biometrics are advocated for counter-terrorism? The talk will discuss design and analysis of biometric accuracy studies, emphasizing inclusion of major sources of variation, heterogeneity in both the presenting population and associated error penalties, statistical issues in summarizing results, and the need for field experimentation within operational systems. Analogy with signal detection theory and medical screening tests is initially helpful, but evaluating biometric authentication/identification is more challenging.

email: pimrey@bio.ri.ccf.org

BAYESIAN ADAPTATION OF THE SUMMARY ROC CURVE MODEL FOR META-ANALYSIS OF DIAGNOSTIC TEST PERFORMANCE

Scott W. Miller*, Medical University of South Carolina
Debajyoti Sinha, Medical University of South Carolina
Elizabeth Slate, Medical University of South Carolina
Don Garrow, Medical University of South Carolina
Joseph Romagnuolo, Medical University of South Carolina

Meta-analytic methods for diagnostic test performance, Bayesian methods in particular, have not been well developed. Here we present a novel Bayesian method for meta-analysis of diagnostic test performance using the Summary Receiver Operator Characteristic (SROC) curve method of Moses, Shapiro and Littenberg (1993, Statistics in Medicine,12,1293-1316). Our method retains the simplicity of the original SROC model while more accurately incorporating uncertainty in the parameters, and can be readily extended to incorporate the effect of covariates. The method is then demonstrated by analyzing the diagnostic performance of endoscopic ultrasound (EUS) in the detection of biliary obstructions relative to the current gold standard of endoscopic retrograde cholangiopancreatography (ERCP).

email: millersw@musc.edu

A UNIFIED FAMILY OF NONPARAMETRIC ROC AREA ESTIMATORS IN GROUP SEQUENTIAL DESIGNS

Liansheng Tang*, University of Washington
Xiao-Hua Zhou, University of Washington
Scott S. Emerson, University of Washington

We consider a unified nonparametric ROC area estimator and its implementation in group sequential design. The asymptotic variance of the estimator is obtained in favor of its implementation in group sequential tests. In the simulation study we describe how the sample sizes can be obtained for comparing the area under the curve and the partial area under the curve for two diagnostic tests by using our approach. We illustrate our method through non-small cell lung cancer trials and calculate sample sizes and sequential boundaries.

email: lstang@u.washington.edu

SEQUENTIAL EVALUATION OF A MEDICAL DIAGNOSTIC TEST WITH BINARY OUTCOMES

Yu Shu*, The George Washington University
Aiyi Liu, National Institute of Child Health & Human Development-Dept. of Health & Human Services Zhaohai Li, The George Washington University, National Cancer Inst, Dept. of Health & Human Services

In a study evaluating a medical diagnostic test, ethical and cost concerns require termination of the study if the test is evidently inefficient (or efficient) in diagnosis of diseases. We propose sequential designs to evaluate the sensitivity and specificity of a diagnostic test. One method proposed in the paper uses error spending approach (Lan and DeMets, 1983) that allows early stopping if the sensitivity and specificity of a new medical test are both within the level of tolerance. Another method, motivated by Simon's (1989) optimal criterion in a phase II clinical trial setting, terminates the study if either sensitivity or specificity is below the minimally acceptable level. Critical values and sample sizes are tabulated for various two-stage designs and compared to the sample sizes of single-stage design.

email: yshu@gwu.edu

ROC ANALYSIS WITH NON-BINARY REFERENCE STANDARD

Shang-Ying Shiu*, Brown University
Constantine Gatsonis, Brown University

Statistical methods for the evaluation of the accuracy of diagnostic tests usually assume a binary true disease status. However, this assumption may not be realistic in practical settings in which i§diseasei¨ is defined by dichotomizing continuous or ordinal categorical measures. In this paper we consider situations in which both the diagnostic test and the reference standard are reported as continuous measures. We propose a semi-parametric model for estimating the sensitivity, specificity and the ROC curve in this setting. When the order restriction is imposed on the mean of the test result variable, the isotonic regression and the monotone smooth splines are two approaches we consider. The model provides the basis to assess the effect of varying reference standard threshold on the performance of a diagnostic test. An example to evaluate the ability of the maximum SUV in predicting axillary node involvement in women diagnosed with breast cancer is presented.

email: shiu@stat.brown.edu

RECENT DEVELOPMENTS IN THE DORFMAN-BERBAUM-METZ (DBM) PROCEDURE FOR MULTIREADER ROC STUDY ANALYSIS

Stephen L. Hillis*, Iowa City V.A. Medical Center
Kevin S. Berbaum, University of Iowa

The Dorfman-Berbaum-Metz (DBM) method is the most frequently used method for analyzing multireader ROC studies. We discuss the following recent developments in the DBM method: (1) the use of normalized pseudovalues allows the method to be based on the original accuracy estimates rather than the jackknife estimates, thus eliminating the problem of jackknife estimates outside the parameter space; (2) the use of less data-based model simplification results in type I errors closer to the nominal level; (3) the use of a new denominator degrees of freedom eliminates the occasional problem of very wide confidence intervals and gives type I errors closer to the nominal level; and (4) sample size computations can be easily made based on pilot data or previous studies. Another recent finding is that the DBM method and the Obuchowski-Rockette method yield identical results when based on the same procedure parameters. This finding means that the DBM procedure only requires the assumptions of the OR method, which are less restrictive and conceptually easier to comprehend; in particular, this relationship makes it clear that the DBM assumptions of normal and independent pseudovalues are not necessary since they are only 'working model' assumptions.

email: steve-hillis@uiowa.edu

54. METHODS FOR MULTIPLE ENDPOINTS

NEW CONFIDENCE BOUNDS FOR QT STUDIES

Dennis D. Boos*, North Carolina State University
David Hoffman, Sanofi-Aventis, New Jersey
Robert Kringle, Sanofi-Aventis, New Jersey
Ji Zhang, Sanofi-Aventis, New Jersey

Current guidelines for acceptable QT interval performance are based on the maximum of a series over time of simple 95% confidence bounds. This procedure is typically very conservative as a procedure for obtaining a 95% bound for the maximum of the population parameters. This paper proposes new bounds for the maximum, both analytical and bootstrap-based, that are lower but still achieve 95% coverage in the context of crossover and parallel designs.

email: boos@stat.ncsu.edu

JOINT MODELS FOR A PRIMARY ENDPOINT AND MULTIVARIATE LONGITUDINAL DATA

Erning Li*, Texas A&M University
Nae-Yuh Wang, The Johns Hopkins University School of Medicine
Naisyin Wang, Texas A&M University

We study the association between a primary endpoint and features of multiple longitudinal processes using joint models, in which a multivariate linear random effects model is proposed for the multiple longitudinal processes whose subject-specific random effects are predictors in a generalized linear model for the primary endpoint. An asymptotic bias analysis indicates that the estimators obtained by Li et al. (2004, Biometrics 60, 1-7), which make no assumption on random effects but assume independent within-subject errors in the longitudinal covariate process, can yield biased inference when these within-subject errors are in fact correlated. To overcome this drawback, we generalize their results to joint models with more flexible multivariate longitudinal covariate processes and develop inferential methods for generalized linear model parameters. Our new methods not only require no assumption on random-effect covariates but also allow general covariance structures of the within-subject measurement errors. We also obtain an estimator for the covariance of the within-subject measurement errors, which is consistent, robust to non-normal random effects and can be obtained using available softwares. This finding enables easy and fast implementation of the proposed methods for joint models. Simulation studies and an application to a hypertension study are used for illustration.

email: eli@stat.tamu.edu

IDENTIFICATION OF RESPONDERS IN AN INTERSTITIAL CYSTITIS CLINICAL TRIAL

Benjamin E. Leiby*, University of Pennsylvania School of Medicine
Mary D. Sammel, University of Pennsylvania School of Medicine
Thomas R. Ten Have, University of Pennsylvania School of Medicine
Kevin G. Lynch, University of Pennsylvania School of Medicine

In this paper, we propose a multivariate growth curve mixture model that groups subjects based on multiple symptoms measured repeatedly over time. Our model synthesizes features of two models. First, we follow Roy and Lin (2000) in relating the multiple symptoms at each time point to a single latent variable. Second, we use the growth mixture model of Muthen and Shedden (1999) to group subjects based on distinctive longitudinal profiles of this latent variable. The mean growth curve for the latent variable in each class defines that class's features. For example, a class of "responders" would have a decline in the latent symptom summary variable over time. A Bayesian approach to estimation is employed where the methods of Elliott et al (2005) are extended to simultaneously estimate the posterior distributions of the parameters from the latent variable and growth curve mixture portions of the model. We apply our model to data from a randomized clinical trial evaluating the efficacy of Bacillus Calmette-Guerin (BCG) in treating symptoms of Interstitial Cystitis. In contrast to conventional approaches using a single subjective Global Response Assessment, we use the multivariate symptom data to identify a class of subjects where treatment is effective.

email: bleiby@cceb.upenn.edu

DELETION DIAGNOSTICS FOR ALTERNATING LOGISTIC REGRESSIONS

John S. Preisser*, University of North Carolina
Jamie Perin, University of North Carolina
Bahjat F. Qaqish, University of North Carolina

Regression diagnostics are introduced for marginal mean and within-cluster association models for correlated binary outcomes estimated with alternating logistic regressions. Computational formulae for one-step deletion diagnostics are introduced that measure the influence of a cluster of observations on the estimated regression parameters and on the estimated linear predictor. The diagnostics are applied to data from a nonrandomized community trial to reduce underage drinking with the aim of assessing the influence of a cluster of observations corresponding to a community on log pairwise odds ratio estimates for intracluster association of any alcohol use in the past 30 days.

email: john_preisser@unc.edu

MULTIVARIATE GAUSSIAN POWER CONFIDENCE INTERVALS DUE TO ESTIMATING COVARIANCE IN ONE OR TWO GROUPS

Sola Park*, University of North Carolina at Chapel Hill
Keith E. Muller, University of North Carolina at Chapel Hill

Uncertainty about the error covariance is often the biggest barrier to accurate power analysis in multivariate and repeated measures models. Using a covariance estimate in a power calculation introduces statistical uncertainty in a computed power value. Taylor and Muller (1995) described easy to compute and exact confidence intervals for power values based on variance estimates of univariate models with Gaussian errors. We extend the known results to multivariate linear models for any hypothesis involving one or two groups, as in typical clinical trials. We describe how to exactly transform any such model to an equivalent univariate model. In turn, univariate results give exact confidence intervals for power of a useful class of repeated measures and multivariate linear models. Keywords: Confidence intervals, multivariate approach, repeated measures, sample size, power

email: spark@email.unc.edu

A GLOBAL TEST TO DETECT REGULATED GENES IN OLIGONUCLEOTIDE GENE CHIP

Dung-Tsa Chen*, University of Alabama at Birmingham
James Chen, NCTR, FDA
ChenAn Tsai, Academia Sinica
Seng-jaw Soong, University of Alabama at Birmingham

The global test has been applied to clinical trial because many clinical trial studies are conducted to compare two treatment groups (e.g., a new treatment and a control or placebo) with respect to several endpoints. Since oligonucleotide gene chip uses a set of probes to interrogate a given gene, treatment effect will be determined by the set of probe expressions. Clearly, the global test can fit well to this question by treating the set of probe expressions as a set of endpoints. We will consider the Lauter's exact test for the hypotheses testing. This global test computes an overall quasi t-statistic by standardizing the expression levels on each probe. Since the global test takes into account of the correlation structure between probes to address the issue of probe heterogeneity, the test can have a better detection power in identifying differentially expressed genes. To address the multiplicity issue, we will adopt a permutation technique to estimate the FDR by calculating the percentage of genes identified by chance.

email: dtchen@uab.edu

REGRESSION ANALYSIS FOR MODELING 16O/18O STABLE-ISOTOPE DISTRIBUTIONS FOR MASS-SPECTROMETRY ANALYSIS

Jeanette E. Eckel-Passow*, Mayo Clinic
Ann L. Oberg, Mayo Clinic
Christopher J. Mason, Mayo Clinic
Douglas W. Mahoney, Mayo Clinic
Robert H. Bergen, Mayo Clinic
Janet E. Olson, Mayo Clinic
Terry M. Therneau, Mayo Clinic

16O/18O stable-isotope labeling is a tool with which to quantify the relative expression differences of proteins between two biological samples. Using stable isotopes, labeled molecules from one sample are pooled with unlabeled molecules from another sample and then subjected to mass-spectral analysis. Stable-isotope methodologies make use of the fact that identical molecules of different stable-isotope compositions are differentiated in a mass spectrometer and are represented in a mass spectrum as distinct isotopic clusters with a known mass shift. We describe a regression analysis procedure for full-scan data that models pairs of isotopic clusters from the same molecule, and likewise quantifies the amount present in each of the two biological samples. Working with full-scan data provides the ability to quantify low-abundance molecules, which is particularly important in cancer-biomarker discovery studies, as well as potentially allowing for information-dependent identification. Additionally, regression analysis is a well characterized and versatile modeling tool that allows intra-sample variability to be incorporated into subsequent analyses. The proposed modeling approach is evaluated with respect to bias and its ability to distinguish molecules with extreme expression difference across the two samples.

email: eckel@mayo.edu

55. CASE-CONTROL STUDIES

MLE METHOD FOR CASE-CONTROL STUDIES WITH LONGITUDINAL COVARIATES

Honghong Zhou*, University of Michigan
Xihong Lin, Harvard University
Bin Nan, University of Michigan

The case-control design is commonly used for studying rare diseases. An emerging problem in case-control studies is the presence of a longitudinal covariate, which is collected longitudinally but retrospectively. We develop a full likelihood approach for case-control studies with longitudinal covariates. We propose a logistic model coupled with a linear mixed model to jointly model the binary outcome and the longitudinal covariate. Specifically, we assume that the longitudinal covariates follow a linear mixed model and the primary binary outcome relates to the longitudinal covariate through latent subject specific random effects, for example, individual baselines and slopes. We derive the true retrospective likelihood for the data and obtain estimates via Maximum Likelihood Estimation (MLE). The retrospective likelihood involves un-estimable population information and can be approximated when the disease is rare. The proposed MLE method is demonstrated under normality of random effects. We apply the proposed method to analyze a case-control breast-cancer dataset where weight is a longitudinal covariate.

email: zhouh@umich.edu

COMPARISONS OF SEQUENTIAL TESTING APPROACHES FOR DETECTION OF ASSOCIATION BETWEEN DISEASE AND CANDIDATE GENES: A SIMULATION STUDY

Andres Azuero*, University of Alabama at Birmingham

A common issue in case-control genetic association studies is that a large number of markers may be tested in the same case-control group, increasing the type I error probability due to multiple comparisons. Sobell et al. (1993) proposed a three-staged sequential testing procedure that allows for a large number of markers to be screened by testing consecutively on three independent case-control subgroups. The authors reported their approach controls for false positive associations, while not seriously affecting power. Sham (1994) argued that the use of independent subgroups at each stage, instead of accumulating data, decreases power. Thus, a sequential scheme was suggested, where at each stage the subsamples of previous stages are combined, maintaining power as the combined sample size increases. It was also suggested that sequential procedures based on Wald (1945) be applied to the problem. This study uses simulated datasets to compare the behavior of the procedures proposed by Sobell, Sham, and Wald's Sequential Probability Ratio Tests, in regards to Type I error rate, false discovery rate, and power.

email: andreo@uab.edu

WEIGHTED ESTIMATING EQUATIONS FOR CASE-CONTROL STUDY WITHIN COHORT WITH CORRELATED FAILURE TIMES

Sangwook Kang*, University of North Carolina at Chapel Hill
Jianwen Cai, University of North Carolina at Chapel Hill

Case-control study design is an efficient and economic method to ascertain a large number of cases in a relatively short period of time. In many studies, the case-control study is conducted within a well-defined cohort. An important assumption for the conventional case-control studies is the statistical independence among subjects. However, in many biomedical studies, this assumption might not hold. In a retrospective dental study, it was of interest to evaluate the degree to which pulpal involvement affects tooth survival. Cases and controls were sampled and a non-pulpally involved tooth was matched to the pulpally involved tooth within each subject. The survival times of the matched teeth within subjects could be correlated and thus the independent assumption might not be valid. We study the marginal proportional hazards regression model for this type of correlated case-control data within cohort. We propose a weighted estimating equation approach or estimating the parameters in the model. The proposed method can also be applied to the studies whose sampling procedure depends on other covariates. Different types of weights are also considered for improving efficiency. Asymptotic properties of the proposed estimators are investigated and their finite sample properties are assessed via simulations studies. The proposed method is applied to the aforementioned dental study.

email: skang@bios.unc.edu

A MINIMUM DISTANCE APPROACH TO LOGISTIC REGRESSION VIA THE CASE-CONTROL FORMULATION

Howard D. Bondell*, North Carolina State University

It is well known that the maximum likelihood fit of the logistic regression parameters can be greatly affected by atypical observations. Several robust alternatives have been proposed in the literature and implemented in standard statistical software packages. However, upon considering the model via the case-control viewpoint, it is clear that current techniques can exhibit poor behavior in many common situations. A new robust class of estimation procedures is introduced. The estimates are constructed via a minimum distance approach after identifying the model with a semiparametric biased sampling model. The approach is developed under the case-control sampling scheme, but is applicable under prospective sampling as well. Estimators resulting from this minimum distance methodology are shown to compare favorably with existing methods used in logistic regression. A particularly useful choice of distance measure is described via a semiparametric empirical characteristic function. These new approaches can be highly efficient if the model is true, while remaining robust to small deviations in the model. Thus they can be used to fit the logistic regression model if it is appropriate for the bulk of the data, even in the presence of atypical observations.

email: bondell@stat.ncsu.edu

SADDLEPOINT APPROXIMATIONS IN MATCHED CASE-CONTROL STUDY

Malay Ghosh, University of Florida
Bhramar Mukherjee, University of Florida
Upasana Santra*, University of Florida

In estimating the odds ratio in a matched case-control study, one uses the conditional likelihood to eliminate stratum specific nuisance parameters.We obtain the saddlepoint approximation to the discrete distribution of the conditional maximum likelihood estimate (CMLE) for 1:1 and 1:2 matched case-control studies and compare them with the standard asymptotic normal approximation to the distributionof the CMLE. For 1:M matched studies, we use saddlepoint approximation to the distribution of the Mantel-Haenszel (MH) estimate of the odds ratio (suitably normalized by variances proposed for the MH estimate). Comparison with the large sample normal approximation for the distribution of this MH statistic shows that the saddlepoint approximation performs much better in general. This illustrates the inadequacy of the normal approximation to the distribution of CMLE as well as MH type statistics, especially when the number of discordant pairs is not large.

email: usantra@stat.ufl.edu

CASE-CONTROL FOLLOW-UP STUDIES: A NEW APPROACH TO SAMPLING FROM A COHORT

Wenguang Sun*, University of Pennsylvania
Marshall M. Joffe, University of Pennsylvania

Case-control follow-up studies provide a new approach to sampling from a cohort and are useful when it is desired to study more than one type of event in the cohort. In this design, all subjects in the cohort who develop either event of interest are sampled, as well as a fraction of the remaining subjects. Typically, it will be desired to study the association between some factors and one of the event types on which sampling is based. We outline three general approaches to analyzing data from these studies, based on the approach of Robins et al. (1994): a simple approach using a weighted estimating function, a more complicated one in which the weighted estimating function is augmented, and the most complicated approach, using the optimal weighted and augmented function. We use simulation to compare the efficiency of these approaches and compare the case-control follow-up approaches to the simpler case-control design.

email: wsun@cceb.upenn.edu

PRESIDENTIAL INVITED ADDRESS

STATISTICAL SCIENCE – KNOWLEDGE FROM INFORMATION

Scott L. Zeger, Frank Hurley and Catharine Dorrier
Professor in Biostatistics and Chair of the Department
of Biostatistics, The Johns Hopkins University
Bloomberg School of Public Health

We are awash in information. This is especially so in the biological and health sciences where bio- and information technologies have produced novel measures of molecules, cells, individuals and populations. But more information does not necessarily produce more understanding.

In this talk, we use three diverse examples ranging from measuring gene expression to quantifying the effects of air pollution on daily mortality, to estimating the number of deaths caused by the invasion of Iraq, to discuss the utility of statistical ideas and methods for achieving new knowledge from complex information.

In each case, technical statistical issues, while important, are not the central ones to obtain or use the results. Rather, a willingness to challenge preconceived notions, the choice of boundaries for the questions asked and the ability to verify and communicate findings and their uncertainty are key to how the statistical analysis influences science and policy. We argue for the importance of "reproducible research" to contend with increasingly complex information and analysis. Finally, we speculate about the role of statistical societies in promulgating statistical ideas and standards that foster careful analysis to achieve knowledge from information.

56. CENSORED DATA IN THE ENVIRONMENTAL, AGRICULTURAL AND MEDICAL SCIENCES

STATISTICAL METHODS FOR CENSORED (NONDETECT) ENVIRONMENTAL DATA

Dennis R. Helsel*, US Geological Survey

Measurements of trace chemicals in water, air, soils, and living organisms frequently result in values reported only as less than the laboratory detection limit ("less-thans" or "nondetects"). A common practice in environmental science for interpreting these data is to substitute one-half the detection limit and perform traditional statistical tests. This overly-simplistic practice results in significant errors in interpretation. Survival analysis (or reliability analysis) methods are standard tools for interpreting right-censored data in medical and industrial statistics, but have rarely been applied to left-censored environmental data. Parametric and nonparametric survival analysis methods, though usually applied to a time variable, can provide summary statistics, hypothesis tests, and regression equations for environmental variables such as concentrations. The results are unequivocal, powerful, and accurate. Applications of survival analysis methods to environmental data are summarized from the author's newly-released textbook Nondetects And Data Analysis: Statistics for Censored Environmental Data, published by John Wiley.

email: dhelsel@usgs.gov

ANALYSIS OF DESIGNED EXPERIMENTS IN THE PRESENCE OF CENSORED DATA

Linda J. Young*, University of Florida
Mary C. Christman, University of Florida
Ramon C. Littell, University of Florida

In designed experiments, the observed data may be left-, right-, or doubly-censored. Ad hoc approaches are generally used in these cases. For example, zeroes may be used for left-censored values, or an arbitrary number at or above the upper threshold may be substituted for right-censored values. Alternatively, the EM-algorithm may be used to impute values for the censored ones. If this is done, the correlation structure is altered, affecting the F-test for treatment effects and tests for differences in treatment means. In this paper, approaches to drawing inferences from studies for which the EM algorithm has been used to impute values for the censored ones will be discussed.

email: LJYoung@ufl.edu

SURVIVAL ANALYSIS IN TWO-STAGE RANDOMIZATION DESIGNS

Abdus S. Wahed*, University of Pittsburgh

Two-stage randomization designs are frequently used in cancer and AIDS clinical trials to study the effectiveness of combinations of therapies. In these designs, patients are initially randomized to an induction treatment followed by another randomization to a maintenance treatment contingent upon their response to the initial treatment. The objective is to compare different combinations of induction and maintenance therapies to find the combination that is most beneficial. In many cases, the outcome of interest is the survival time, and censoring complicates the analysis of data from such designs. It is commonplace to analyze the data separately for the two stages which simplifies the analysis but does not directly address the objective of the study. Recently, a few articles [Lunceford et al. (Biometrics, 2002), Wahed and Tsiatis (Biometrics, 2004)] have suggested consistent estimators for the survival distributions of treatment policies (combinations of treatments). In this talk, we will review available estimators and discuss methods for constructing more efficient estimators. The techniques will be demonstrated through a simulation study and an application to a Leukemia dataset.

email: wahed@pitt.edu

57. STATISTICAL ISSUES IN META-ANALYSIS OF GENOMIC AND TRANSCRIPTIONAL DATA WITH A FOCUS ON ARRAY CGH

METHODS FOR THE JOINT ANALYSIS OF ARRAY CGH AND GENE EXPRESSION DATA

Adam B. Olshen*, Memorial Sloan-Kettering Cancer Center
E. S. Venkatraman, Memorial Sloan-Kettering Cancer Center

Biologists are interested in finding genes that have both abnormal DNA copy number and differential expression. Our talk will focus on the search for these genes when both types of data are collected using microarrays. We will address how often abnormal DNA copy number leads to differential expression. In addition, we will examine what type of copy number changes can be identified using only expression data. In addition to introducing our methodology, we will make comparisons to previous efforts to address these problems. Methods will be demonstrated on data from a germ cell tumor study. Finally, we will provide background information on array CGH and DNA copy number to help motivate this talk and other talks in the session.

email: olshen@yahoo.com

COMBINING COPY NUMBER AND GENE EXPRESSION DATA FOR THE ANALYSIS OF CANCER DATA

Jane Fridlyand*, Dept. of Epidemiology & Biostatistics, Comprehensive Cancer Center, University of California-San Francisco
Ritu Roydasgupta, University of California-San Francisco
Sandy DeVries, University of California-San Francisco
Koei Chin, University of California-San Francisco
Fred Waldman, University of California-San Francisco
Joe Gray, LBNL
Donna Albertson, University of California-San Francisco

The development of solid tumors is associated with acquisition of complex genetic alterations, indicating that failures in the mechanisms that maintain the integrity of the genome contribute to tumor evolution. Thus, one expects that the particular types of genomic derangement seen in tumors to reflect underlying failures in maintenance of genetic stability, as well as selection for changes that provide growth advantage. In order to investigate genomic alterations we are using BAC microarray-based comparative genomic hybridization (array CGH). Transcriptional profiles are measured using HGU133A Affymetrix chips. The computational task is to map and characterize the number and types of copy number alterations present in the tumors, and so define copy number phenotypes as well as to associate them with known biological markers and with gene expression data. We define distinct types of genomic events and identify the groups of genes associated with different instabilities. We conclude that various types of genomic instability is associated with the defects in distinct functional groups as determined by Gene Ontology. This result has implications for potential targeted therapies. The study is conducted using three extensive (between 50 and 100) tumor sets of ovarian and breast cancer patients.

email: janef@cc.ucsf.edu

DETECTION OF THE DNA COPY NUMBER CHANGES USING HIGH DENSITY OLIGONUCLEOTIDE ARRAYS

Jing Huang*, Affymetrix Inc.
Wen Wei, Roche Molecular Systems, Inc.
Joyce Chen, Affymetrix Inc.
Jane Zhang, Affymetrix Inc.
Guoying Liu, Affymetrix Inc.
XiaoJun Di, Affymetrix Inc.
Rui Mei, Affymetrix Inc.
Shumpei Ishikawa, University of Tokyo
Keith W. Jones, Affymetrix Inc.
Michael H. Shapero, Affymetrix Inc.

One of the hallmark features of cancers is DNA copy number alterations. We have developed a method termed whole genome sampling analysis (WGSA) which genotypes over 100,000 single nucleotide polymorphisms (SNPs) from human genomic DNA. Hybridization of target DNA to high density oligonucleotide arrays containing both perfect match (PM) and mismatch (MM) probes allows allele-specific identification of homozygous deletions, gene amplifications, and chromosomal regions of reduced and elevated copy number. The comparison of probe intensity from an experimental sample to intensity distributions derived from a large reference set allows the statistical significance of any DNA copy number change to be determined. The WGSA assay shows a high linear response between changes in genomic copy number and fluorescence intensity. Copy number changes identified in an established human breast cancer cell line were further verified using quantitative PCR. Overall, WGSA results are similar to those of array CGH but have the added advantage of coupling genotype calls to copy number information. This allows for a more detailed analysis of genomic alterations. With mean and median euchromatic inter-SNP distances of 23.6 kb and 8.5 kb respectively, this method affords resolution that is not easily achievable with alternative experimental approaches.

email: jing_huang@affymetrix.com

VISUALIZING AND ANALYZING HIGH DENSITY SNP DATA WITH SNPSCAN

Ingo Ruczinski*, Johns Hopkins University
Rob Scharpf, Johns Hopkins University
Jason Ting, Kennedy Krieger Institute
Jonathan Pevsner, Kennedy Krieger Institute

High density single nucleotide polymorphism microarrays (SNP chips) provide information on a subject's genome, such as the chromosomal copy numbers and the genotype (heterozygosity/homozygosity). In contrast to other existing approaches such as fluorescence in situ hybridization (FISH) and karyotyping, these novel tools provide high-resolution of genome-wide measurements that can be used, for example, to detect chromosomal microdeletions. As a variety of diseases are caused by chromosomal abnormalities such as aneuploidies, microdeletions, microduplications, and uniparental disomies, SNPchips promise new insights for these diseases by aiding in the identification of regions of chromosomal abberations, and can possibly suggest targets for intervention. However, the identification of different classes of anomalies using SNP data is challenging, as the data can be very noisy, and the regions of chromosomal abnormalities can be very small. We introduce the R package SNPscan to visualize and analyze high density SNP data, and show examples of its usefulness.

email: ingo@jhu.edu

58. HEALTH SURVEY DATA

COMBINING INFORMATION FROM MULTIPLE SURVEYS TO ENHANCE ESTIMATION OF MEASURES OF HEALTH

Nathaniel Schenker*, National Center for Health Statistics

Survey estimates are often affected by non-sampling errors due to missing data, coverage error, and measurement or response error. Such non-sampling errors can be difficult to assess, and possibly correct for, using information from a single survey. Thus, combining information from multiple surveys can be beneficial. This talk will discuss four projects undertaken by researchers within and outside the National Center for Health Statistics, in which information from multiple surveys was combined to adjust for non-sampling errors and thereby enhance estimation of various measures of health. The projects can be described briefly as follows: combining estimates from a survey of households and a survey of nursing homes to enhance the coverage of the population; using information from an interview survey to bridge the transition in race reporting in the United States census; combining information from an examination survey and an interview survey to improve on analyses of self-reported data; and combining information from two interview surveys to enhance small-area estimation. Issues that can arise when information is combined from multiple surveys will be discussed as well.

email: nschenker@cdc.gov

USING NATIONAL SURVEYS TO COMPUTE THE NUMBER OF DEATHS ATTRIBUTABLE TO A RISK FACTOR

Barry I. Graubard*, National Cancer Institute
Katherine M. Flegal, National Center for Health Statistics/Centers for Disease Control and Prevention
David F. Williamson, Centers for Disease Control and Prevention
Mitchell H. Gail, National Cancer Institute

Estimates of the attributable number of deaths (AD) from all-causes can be obtained by first estimating population attributable risk (AR) adjusted for confounding covariates, and then multiplying the AR by the number of deaths determined from vital mortality statistics that occurred in the population for a specific time period. Proportional hazard regression estimates of adjusted relative hazards obtained from mortality follow-up data from a cohort is combined with a joint distribution of risk factor and confounding covariates to compute an adjusted AR. Two estimators of adjusted AR are examined, which differ according to the reference population that the joint distribution of risk factor and confounders is obtained. Methods based on influence function theory that are used in survey sampling are applied to obtain expressions for estimating the variance of the AD estimator. These variance estimators can be applied to data that range from simple random samples to (sample) weighted multi-stage stratified cluster samples like those used in national household surveys. The variance estimation of AD is illustrated in an analysis of excess deaths due to having a non-ideal body mass index using data from the second National Health and Examination Survey.

email: graubarb@mail.nih.gov

CORRELATION IN MULTISTAGE HEALTH SURVEY DESIGNS

Jai W. Choi*, National Center for Health Statistics
Balgobin Nandram, Worcester Polytechnic Institute

Situations often arise in a large-scale household survey where in a complex probability sample of clusters rather than of individuals is taken from a large population. Typically, such cluster sampling involves a number of smaller secondary clusters and each secondary cluster consists of a number of elementary units. Two different types of intracluster correlation may arise in such data. One is the intracluster correlation between members in the primary cluster and the other is the intracluster correlation between members in the secondary cluster, both may be defined by terms of a analysis of variance model (ANOVA). This paper discusses estimation strategies for the two types of intracluster correlation using ANOVA components. Other is the method by directly estimating the parameters of the definition of correlation. Data used for demonstration is qualitative data from an unbalanced design. The variance and large sample distribution of direct estimator may also be discussed.

email: jwc7@cdc.gov

59. ILS: INTRODUCTION TO BAYESIAN ANALYSIS AND SOFTWARE

INTRODUCTION TO BAYESIAN ANALYSIS AND SOFTWARE

Bradley P. Carlin*, University of Minnesota

Hierarchical Bayes methods enable the combining of information from similar and independent experiments, yielding improved inference for both individual and shared model characteristics. This ILS (introductory leture session) will introduce hierarchical Bayes methods, demonstrate their usefulness in challenging applied settings, and show how they can be implemented using modern Markov chain Monte Carlo (MCMC) computational methods. The speaker will also provide an introduction to and (time permitting) a live demonstration of WinBUGS, the most general Bayesian software package available to date. Use of the methods will be demonstrated in advanced high-dimensional model settings of interest to biostatistical and biomedical researchers, where the MCMC Bayesian approach often provides the only feasible alternative that incorporates all relevant model features. The ILS is generally aimed at students and practicing statisticians who are intrigued by all the fuss about Bayes and Gibbs, but who may still mistrust the approach as theoretically mysterious and practically cumbersome.

email: brad@biostat.umn.edu

60. IMS: RECENT DEVELOPMENTS IN FALSE DISCOVERY RATE

HIERARCHICAL FDR CONTROLLING PROCEDURES

Daniel Yekutieli*, Tel Aviv University

I will present FDR trees - a newclass of hierarchical FDR controlling procedures. In this new testing approach rather than test all the hypotheses simultaneously, the tested hypotheses are arranged in a tree of disjoint subfamilies and the tree of subfamilies is tested hierarchically. This is a very flexible and powerful testing framework suited to perform complex statistical analysis. I will present the theoretical properties of FDR trees and then demonstrate the use of FDR trees with real data examples: complex statistical analysis of microarray data and Log-linear model selection.

email: yekutiel@post.tau.ac.il

TAIL STRENGTH OF A DATASET

Jonathan Taylor*, Stanford University
Robert Tibshirani, Stanford University

We propose an overall measure of significance for a set of hypothesis tests. The tail strength is a simple function of the p-values computed for each of the tests. This measure is useful, for example, in assessing the overall univariate strength of a large set of features in microarray and other genomic and biomedical studies. It also has a simple relationship to the false discovery rate of the collection of tests. We derive the asymptotic distribution of the tail strength measure, and illustrate its use on a number of real datasets.

email: jonathan.taylor@stanford.edu

SENSITIVITY AND SPECIFICITY OF FDR METHODS IN NEUROIMAGING

Thomas E. Nichols*, University of Michigan
Wei Xie, University of Michigan

False Discovery Rate (FDR) methods are ideally suited for neuroimaging, as the standard FDR method is valid under positive dependence (smoothness), and because users have found that methods with strong control of familywise error are too stringent. Despite the growing use, FDR methods have had relatively little evaluation in the typical, highly smooth neuroimaging data. Further, there is often a misunderstanding that control of the expected FDR, E(V/R), implies control of the realized FDR, V/R; that is, users interpret a level-q thresholded image has having no more than q*R false discoveries. False Discovery Proportion (FDP) methods have been developed to address this problem, where a threshold is found such that P(V/R < q) > C; that is, the number of false discoveries V is less than q*R with confidence level C. In this work we perform extensive simulations of FDR and FDP methods. We assess specificity under a range of smoothness, as some have reported increased conservativeness with increased smoothness. We also assess power for a range of smoothnesses and physiologically-realistic stimuli, and characterize difference in power between FDR and FDP methods.

email: nichols@umich.edu

61. INTERVAL-CENSORED TIME-TO-EVENT DATA

MAXIMUM LIKELIHOOD ANALYSIS OF REPEATED LEFT- AND INTERVAL-CENSORED BIOASSAY DATA

Jonathan S. Hartzel*, Merck & Co., Inc.

In a clinical trial to compare mumps immune responses induced by two mumps-containing vaccines, post-vaccination serum samples were collected from 2 groups of subjects and tested for the presence of neutralizing antibodies against five mumps virus isolates using a 2-fold dilution series from a plaque reduction neutralization assay. Thus, the measured titer is a left-censored or interval-censored version of the true unobserved titer. A goal of the study was to estimate the fold-difference in geometric mean titers across all isolates between groups. The conventional method of analysis assumes the measured titer to be the true titer (or halved if left-censored), and thus is subject to potential bias and to under estimating of the variability. Here, we develop a general linear mixed model incorporating both left- and interval-censored response data and allowing for an unstructured covariance matrix. ML estimation of the model using a Newton-Raphson algorithm requires repeated evaluation of multivariate normal probabilities. These probabilities are estimated via the GHK simulator using quasi-MC samples generated from a randomized lattice rule. Results of the proposed model are compared with those obtained from a model in which the measured titer is assumed to be the true titer.

email: jonathan_hartzel@merck.com

A CONDITIONAL APPROACH FOR REGRESSION ANALYSIS OF CASE 2 INTERVAL-CENSORED FAILURE TIME DATA

Lianming Wang*, University of Missouri-Columbia
Jianguo Sun, University of Missouri-Columbia
Xingwei Tong, University of Missouri-Columbia

Interval-censored failure time data often arise in clinical trials and medical follow-up studies and a few methods have been proposed for their regression analysis using various regression models (Finkelstein, 1986; Huang, 1996; Lin et al., 1998; Sun, 2005). This paper considers the regression analysis using the additive hazards model, for which it seems that there is no inference approach available except for some special cases. To estimate regression parameters of interest, a conditional inference approach is presented that does not involve estimation of the cumulative baseline hazard function. Asymptotic properties of the proposed parameter estimates are established and some simulation results and an illustrated example are provided. KEY WORDS: Additive hazards model; Counting processes; Estimating equation; Proportional hazards model; Regression analysis.

email: lwdzc@mizzou.edu

SURVIVAL CURVE ESTIMATION FOR INFORMATIVELY COARSENED DISCRETE EVENT-TIME DATA

Michelle D. Shardell*, University of Maryland School of Medicine
Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health
Sam A. Bozzette, San Diego Veterans Affairs Medical Center

Interval-censored, or more generally, coarsened event-time data arise when study participants are observed at irregular time periods and experience the event of interest in between study observations. Such data are often analyzed assuming non-informative censoring, which can produce biased results if the assumption is wrong. This paper extends survivor curve estimation to allow informatively interval-censored data by incorporating various assumptions about the censoring mechanism into the model. We include a Bayesian extension in which final estimates are produced by mixing over a distribution of assumed censoring mechanisms. We illustrate these methods with a natural history study of HIV-infected individuals using assumptions elicited from an AIDS expert.

email: mshardel@epi.umaryland.edu

A NONPARAMETRIC TEST FOR INTERVAL-CENSORED FAILURE TIME DATA WITH UNEQUAL CENSORING

Chao Zhu*, University of Missouri-Columbia
Jianguo Sun, University of Missouri-Columbia

This paper considers nonparametric comparison of survival functions, one of the most commonly required tasks in survival studies. For this, several test procedures have been proposed for interval-censored failure time data in which distributions of censoring intervals are identical among different treatment groups (Petroni and Wolfe, 1994; Pan, 2000; Zhang et al., 2001). Sometimes the distributions may not the same and depend on treatments. A class of test statistics is proposed for situations where the distributions may be different for subjects in different treatment groups. The asymptotic normality of the test statistics is established and the test procedure is evaluated by simulations, which suggest that it works well for practical situations. An illustrative example is provided.

email: chaozhu2001@yahoo.com

SENSITIVITY OF KAPLAN-MEIER ESTIMATE TO NONIGNORABLE CENSORING

Tao Liu*, University of Pennsylvania
Daniel F. Heitjan, University of Pennsylvania

Untestable assumptions about the censoring mechanism can affect estimates of survival functions that assume censoring is ignorable. This article explores the sensitivity of the Kaplan-Meier estimator to informative censoring by extending the method of ISNI (index of local sensitivity to nonignorability) (Troxel, Ma, Heitjan, 2004) to the nonparametric case. A coarse-data model (Heitjan and Rubin, 1991) describes the association between failure and censoring processes, upon which the correct likelihood that accounts for the censoring mechanism is constructed. We also discuss a self-consistent estimator of the survivor function which can give consistent estimates for a given informative censoring model. We use it to assess the appropriateness of ISNI as a sensitivity measure for the Kaplan-Meier estimator via a simulation. Using the ISNI of the Kaplan-Meier estimate, we demonstrate it is easy to evaluate the sensitivity of nonparametric estimates of moments and quantiles. As an example, we apply the method to the Stanford heart transplant data.

email: tliu@cceb.upenn.edu

'SMOOTH' INFERENCE FOR SURVIVAL FUNCTIONS WITH ARBITRARILY CENSORED DATA

Kirsten Doehler*, North Carolina State University
Marie Davidian, North Carolina State University

Standard methods for inference on survival functions with right- or interval-censored data are traditionally nonparametric, and hence impose no assumptions on the true survival distribution. We propose a new procedure for estimation of survival functions that allows a unified approach to handling different kinds of censoring that is based on the premise that, if one is willing to make mild smoothness assumptions on the underlying true survival distribution, efficiency gains and computational advantages over nonparametric methods may be possible. The approach assumes that the survival distribution has a 'smooth' density, which is approximated by the so-called seminonparametric (SNP) density. The SNP has a flexible 'parametric' representation that admits a convenient expression for the likelihood and allows it to capture arbitrary shapes through choice of a tuning parameter, which may be carried out based on standard criteria such as AIC and BIC. We describe the approach and its implementation and validate its performance in empirical studies. We also develop a test statistic for comparing two survival distributions and contrast its performance with traditional nonparametric tests.

email: kadoehle@stat.ncsu.edu

THE IMPACT OF CENSORING PATTERNS ON THE ANALYSIS OF INTERVAL-CENSORED DATA

Guozhi Gao*, Amgen Inc.
Xiang Zhang, Amgen Inc.
Steven Snapinn, Amgen Inc.
Qi Jiang, Amgen Inc.

While methods exist for analyzing interval-censored data, due to their complexity and the lack of available software, a common but naïve approach is to use models for standard survival data, where an event is treated to occur at the beginning, the middle, or the end of the interval it belongs to. However, the performance of this naïve approach is not entirely clear, and there are studies where it does or does not show reasonable performance. To our best knowledge, there has not been any successful investigation or satisfactory explanation for this controversial phenomenon. In this presentation, through simulations we investigate the performance of the naïve approach for grouped survival data under various settings. Explanation will be given through a study of the (partial) likelihood and estimation equations. We found that the unstable behavior of the naïve approach could be due to different right censoring patterns between the two treatment groups. This idea is illustrated using real data where the exact event time is observed, and then artificially grouped into intervals. Finally, we study the performance of some existing standard methods for grouped data, applied to the cases where the right censoring patterns between the two treatment groups are different.

email: ggao@amgen.com

62. MODELING METHODS IN EPIDEMIOLOGY

STATISTICAL MODELING AND ITS EVALUATION OF REFERENCE VALUES FOR PULMONARY FUNCTION TEST: A MULTIVARIATE APPROACH

JungBok Lee*, Korea University
Chol Shin, Korea University
Jae Won Lee, Korea University

Spirometry is one of the important tests to assess pulmonary mechanical function. The interpretation of pulmonary function tests is comparing the measured values with average values from a representative sample of healthy nonsmoking populations for which a reference range has been determined. Hence, reference values should be chosen that are appropriate to the person be investigated and also reference value equations which provide a context for evaluating the pulmonary function values of an individual subject, and also should describe relations among pulmonary function measurements. For predicting the equations, simple linear regression approach was the most common model used to describe pulmonary function. However, the prediction of linear regression equations for pulmonary function measurements(such as Forced Vital Capacity(FVC), Forced Expiratory Volume(FEV), etc.) was performed separately, which caused to distort the relation of pulmonary function measurements. In this study, we consider multivariate linear mixed and semi-parametric model with smoothing spline and provide the result of model evaluation. The model construction is conducted with ongoing longitudinal epidemiologic study data, named Korean Genomic Study and NHANES III data.

email: jungboklee@korea.ac.kr

CANCER RISK ASSESSMENT OF ENVIRONMENTAL AGENTS BY STOCHASTIC MODELS OF CARCINOGENESIS

Wai-Yuan Tan*, University of Memphis
Wenyen Zhao, University of Memphis
Chao W. Chen, US EPA
Li-jun Zhang, University of Memphis

In this paper we have developed procedures to assess cancer risks of environmental agents by biologically supported stochastic models of carcinogenesis. The stochastic models used are the stochastic multi-stage models of carcinogenesis involving mutations and stochastic birth and death processes for the proliferation and differentiation of initiated cancer stem cells. We have applied these models to analyze the data of effects of arsenic in drinking water on bladder cancer published in Environmental Health Perspective (Morales et al. 2000, Vol. 108, pp 655-661). Our results showed that the two-stage model fitted a little better than the 3-stage models for the data as evidenced by AIC and BIC. Our results also showed that the carcinogen is basically an initiator; the promotion effect is quite small unless the dose level is very large.

email: waitan@memphis.edu

FLEXIBLE BAYESIAN MULTISTATE MODELS FOR MULTIVARIATE LONGITUDINAL DATA

Bo Cai*, NIEHS
David B. Dunson, NIEHS
Joseph B. Stanford, University of Utah

In longitudinal studies, an individual often progresses through several latent health states, with multivariate data changing in distribution non-linearly according to trajectories dependent on the latent state. For example, in time to pregnancy studies, interest often focuses on identifying predictors of the day-specific probabilities of conception in relation to the timing of intercourse given multiple days of intercourse and mucus observations within each menstrual cycle. It is also biologically of interest to determine different latent phases that hormones and mucus could have in menstrual cycles. In this paper we propose a Bayesian multistate approach to allow different states (e.g. fertile and infertile windows) to flexibly vary across subjects and dependent observations within subjects. Based on multivariate adaptive regression splines, our model allows the distribution non-linearly to vary for dependent trajectories (e.g. multiple dependent mucus characteristics). A Markov chain Monte Carlo algorithm is described for posterior computation. The method is illustrated through application to fertility data from a cohort study of women using the Creighton Model Fertility Care System.

email: cai@niehs.nih.gov

COMPARING SMOOTHING TECHNIQUES FOR MODELING EXPOSURE-RESPONSE CURVES IN COX MODELS

Usha S. Govindarajulu*, Harvard School of Public Health
Donna Spiegelman, Harvard School of Public Health
Sally W. Thurston, University of Rochester Medical Center
Ellen A. Eisen, Harvard School of Public Health

To allow for a non-linear exposure-response relationship, we applied flexible nonparametric smoothing techniques to models of time to lung cancer mortality in an occupational cohort with a skewed exposure distribution. We focused on three different smoothing techniques in Cox models: penalized splines, restricted cubic splines, and fractional polynomials. We compared standard software implementations of these three methods based on the visual representations they produced, criteria for model selection, and degrees of freedom. We proposed a measure of the difference between a pair of curves based on the area between them, standardized by the total area under the curves.. The three dose-response curves were all similar where the exposure data were dense, with the difference between curves at the 90th percentile was only 2-5% of the total difference. Using weighted area based on inverse variance weighting reinforced the overall similarity of the three smoothing methods and located the divergence in the extreme high end of the skewed exposure distribution. In this example, the penalized spline curve was closest to the restricted cubic spline, with a difference less than half that between the other two pairs of curves. Finally, we also simulated data and modeled each smoothing technique in order to better understand how these curves would fit particular exposure distributions and compare.

email: usha@alum.bu.edu

UNIFYING REGRESSION APPROACHES FOR ESTIMATING CHRONIC EFFECTS OF AIR POLLUTION ON HUMAN HEALTH

Sorina E. Eftim*, Johns Hopkins Bloomberg School of Public Health
Francesca Dominici, Johns Hopkins Bloomberg School of Public Health

We investigate under which data structures a Cox proportional hazard model of individual-level hazard of death on day t with random effects is equivalent to a Poisson model of daily numbers of deaths with spatial smoothing for estimating chronic effects of air pollution on human health accounting for spatial confounding. Previous work has shown that, under certain conditions, the Cox and Poisson likelihoods are equivalent. Here we plan to extend this work to account for spatial confounding by either including random effects or spatial smoothing. Specific questions that we aim to address are: 1.Contrast Poisson regression with spatial smoothing for aggregated data with a Cox Proportional Hazard model for individual-level data with spatial smoothing in estimating chronic effects of air pollution on human health in spatially correlated data;2.Conduct simulation study to investigate advantages and/or disadvantages (computational, methodological, substantive) of the two approaches. Compare with previous approaches (Poisson model with random effect); 3. Illustrate approaches using the Medicare cohort.

email: seftim@jhsph.edu

63. CLINICAL TRIALS

POWER APPROXIMATION FOR THE VAN ELTEREN TEST BASED ON LOCATION-SCALE FAMILY OF DISTRIBUTIONS

Yan Zhao*, Eli Lilly and Company
Yongming Qu, Eli Lilly and Company
Dewi Rahardja, University of Indianapolis

The van Elteren test, as a type of stratified Wilcoxon-Mann-Whitney test for comparing two treatments accounting for stratum effects, has been used to replace the analysis of variance when the normality assumption was seriously violated. The sample size estimation methods for the van Elteren test have been proposed and evaluated previously. However, in designing an active-comparator trial where a sample of responses from the new treatment is available but the active comparator is known only up to summary statistics, the existing methods are either inapplicable or poorly behaved. In this paper we develop a new method assuming the responses from both treatments are from the same location-scale family. Theories and simulations have shown that the new method performs well when the location-scale assumption holds and works reasonably when the assumption does not hold. Thus, the new method is preferred when computing sample sizes for the van Elteren test in active-comparator trials.

email: yzhao@lilly.com

WORKING WITH THE DATA SAFETY MONITORING BOARD FOR A CLINICAL TRIAL: A QUESTION OF POWER

Felicity B. Enders*, Mayo Clinic
Jeffrey A. Schmoll, Mayo Clinic
Tanya L. Hoskin, Mayo Clinic

Primary Sclerosing Cholangitis (PSC) is a rare liver disease with few treatment options. A clinical trial is currently underway to estimate the effect of high-dose ursodeoxycholic acid for long term treatment of the disease. Due to low endpoint accrual, the Data Safety Monitoring Board (DSMB) requested an estimated timeline for endpoint accrual. We present a practical method for predicting future endpoint accrual in a clinical trial when patients enroll during multiple years.

email: enders.felicity@mayo.edu

BAYESIAN DESIGN FOR A SECOND CLINICAL TRIAL

Elizabeth A. Johnson*, Bloomberg School of Public Health-Johns Hopkins University
Scott L. Zeger, Bloomberg School of Public Health-Johns Hopkins University
Jay Herson, Bloomberg School of Public Health-Johns Hopkins University

This paper considers the situation where a randomized clinical trial has been completed to compare two treatment groups. A second randomized trial is planned to acquire additional evidence about the relative treatment efficacies. Effect size is measured by the log odds ratio. This paper describes a Bayesian trial design that utilizes the information gathered on the endpoint in the first trial and prior knowledge about the likely heterogeneity in treatment effects across the superpopulation of all possible trials of which the two at hand are thought of as a representative sample. Sample size and power calculations are presented under different scenarios to demonstrate the influence of the data from the first trial and the specification of the prior. Resulting sample sizes and power are compared with their frequentist counterparts to demonstrate the utility of the Bayesian approach. Regulatory considerations for specification of the prior based on the characteristics of the endpoint in the first trial will be given.

email: ejohnson@jhsph.edu

LONGITUDINAL NESTED COMPLIANCE CLASS MODEL IN THE PRESENCE OF TIME-VARYING NONCOMPLIANCE

Julia Y. Lin*, University of Pennsylvania School of Medicine
Thomas R. Ten Have, University of Pennsylvania School of Medicine
Michael R. Elliott, University of Michigan School of Public Health

We propose a nested latent class model for analyzing longitudinal randomized trials when subjects do not always adhere to the treatment to which they are randomized in the context where randomization remains constant through time but compliance may vary over time. Traditional intention-to-treat and as-treated analyses may produce biased causal effect estimates in the presence of subject noncompliance. Utilizing a nested latent class model that uses subject-specific and time-invariant 'superclasses' allows us to summarize longitudinal trends of time-varying compliance patterns, and estimate the causal effect of the intervention controlling for longitudinal compliance behaviors. Analyses of the PROSPECT study of the effect of the intervention on depression outcomes show that subjects with more severe depression are more likely to adhere to treatment randomization, and those that are compliant show improvement in depression. Simulation results show that our estimation procedure produces reasonable parameter estimates under correct model assumptions.

email: jlin@cceb.upenn.edu

COMPOSITE DESIGN FOR DOSE-FINDING UNDER BIVARIATE PROBIT MODEL

Yuehui Wu*, GlaxoSmithKline
Vladimir Dragalin, GlaxoSmithKline
Valerii V. Fedorov, GlaxoSmithKline

The primary goal of a dose-finding study is to establish the dose-response relationship. The optimal experimental design framework provides enough structure to make this goal attainable in a restricted sense. It is assumed that the available doses (the design region) and the response variables have been defined and there exists a known structure for the mathematical model describing the dose-response relationship (the model). The focus is on choosing the dose levels in some optimal way to enhance the process of estimating the unknown parameters of the model. Our major target in this study is the implementation of composite optimal design techniques in dose-finding studies using bivariate Probit model when two endpoints, the first endpoint corresponds to efficacy and the second to toxicity, can be observed simultaneously on each subject.

email: yuehui.2.wu@gsk.com

WEIGHTED LOG RANK SUBTRACTION

John L. Bryant*, University of Pittsburgh

In the setting of operable breast cancer treatment, we consider the comparison of survival in two arms of a randomized clinical trial or in a meta-analysis of trials. Particularly in meta-analyses, it may be difficult to obtain accurate cause-of-death information; even when excellent source documentation is available, accurate classification of specific cases is sometimes problematic. We are therefore interested in methods for comparing the hazard of disease-specific mortality between the treatment regimens that do not require explicit categorization of deaths according to cause. One such procedure is the method of log rank subtraction, proposed and used by the Early Breast Cancer Trialists Collaborative Group (EBCTCG) in their ongoing series of quinquennial meta-analyses. We show that the validity of this procedure is predicated on strong assumptions that may not be reasonable in many applications. As an alternative, we propose a modification referred to as 'weighted log rank subtraction', that does not require these assumptions. We develop the operating characteristics of the new method, and compare its performance to that of log rank subtraction by analytical means and by simulation. We illustrate the procedure using data from clinical trials previously conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP).

email: bryant@nsabp.pitt.edu

64. MISSING DATA METHODS

MULTIPLE IMPUTATION IN SURVIVAL ANALYSIS WITH INCOMPLETE COVARIATES

Jia Li*, University of Pittsburgh
Stewart Anderson, University of Pittsburgh

We investigate a method for fitting regression models in survival analysis data which has missing information for both categorical and continuous covariates. A multiple imputation method based on a general location model is applied to the missing covariates. The missing data mechanism is assumed to be ignorable. Simulations via Schafer's MIX library were conducted to examine the properties of parameter estimates of multiple imputation. Both ignorable and nonignorable missing data were considered to examine the robustness of the model. Comparisons of the results from the multiple imputation method with those from the fully observed data and a traditional 'complete-case' analysis verify that the multiple imputation method generates reasonable estimators of coefficients and is more efficient than the complete-case analysis for both large and small sample sizes. We apply the multiple imputation method to survival data in breast cancer patients enrolled in Protocol B-06 of the National Surgical Adjuvant Breast and Bowel Project (NSABP). KEY WORDS: survival data, missing covariates, multiple imputation methods

email: li@nsabp.pitt.edu

SHARED PARAMETER MODELS WITH A FLEXIBLE RANDOM EFFECTS DISTRIBUTION

Roula Tsonaka*, Catholic University of Leuven
Geert Verbeke, Catholic University of Leuven
Emmanuel Lesaffre, Catholic University of Leuven

Longitudinal studies often generate data according to a MNAR mechanism, where missingness is related to the unobserved responses. In such cases, joint modelling of the measurement and missingness process is required to account for informative missingness. Shared parameter models provide a flexible framework for the joint modelling of these two processes, where it is assumed that a set of random effects induces the dependence between them. A key feature of these models is the conditional independence assumption, which implies that the measurement and the missingness processes are independent given the random effects. Although parametric assumptions are usually made for this latent process, these may be very restrictive or even unrealistic, provided that knowledge about the distribution of the unobservable random effects is not directly available. In addition, the choice of the random effects distribution can severely affect the validity of the results, since both the conditional independence assumption and the identification of the missing response components are strongly related to this distribution. In this work we propose to relax the commonly used distributional assumptions for the random effects to a more flexible semiparametric distribution given as a mixture of normal variables with an unspecified number of components and fixed support points. This choice for the random effects distribution allows for more general shapes of distribution than the commonly used normal distribution. In this way we also allow for types of association between the two processes other than elliptical which is important in order to validate the conditional independence assumption. The proposed model is estimated using a two step procedure that iterates until convergence between a Vertex Exchange Method step for the estimation of the mixing distribution over the class of arbitrary mixtures of normals, and an EM step for the estimation of the model parameters.

email: spyridoula.tsonaka@med.kuleuven.be

AN IMPUTATION STRATEGY FOR BINARY DATA

Hakan Demirtas*, University of Illinois at Chicago
Don Hedeker, University of Illinois at Chicago

New quasi-imputation and expansion strategies for correlated binary responses are proposed by borrowing ideas from random number generation. The core idea is to convert correlated binary outcomes to multivariate normal outcomes in a sensible way so that re-conversion to the binary scale, after performing multiple imputation, yields the original specified marginal expectations and correlations. This conversion process ensures that the correlations are transformed reasonably which in turn allows us to take advantage of well-developed imputation techniques for Gaussian outcomes. We use the phrase "quasi" because the original observations are not guaranteed to be preserved. We argue that if the inferential goals are well-defined, it is not necessary to strictly adhere to the established definition of multiple imputation. Our expansion scheme employs a similar strategy where imputation is used as an intermediate step. It leads to proportionally inflated observed patterns, forcing the dataset to a complete rectangular format. The plausibility of the proposed methodology is examined by applying it to a wide range of simulated datasets that reflect alternative assumptions on complete data populations and missing-data mechanisms.

email: demirtas@uic.edu

SIEVE MAXIMUM LIKELIHOOD ESTIMATION FOR MISSING COVARIATES IN REGRESSION MODELS

Qingxia Chen*, Vanderbilt University
Donglin Zeng, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill

Misspecification of the covariate distribution in missing data problems is studied for regression models with several missing covariates. We propose a new semiparametric method which specifies a fully nonparametric model for the conditional distribution of the missing covariates given the completely observed covariates, assuming the missing covariates are missing at random (MAR). For ease of exposition, we first deal with the problem of one missing continuous covariate. To obtain the estimates, the log of the fully unspecified covariate joint density is approximated by B-spline functions and the estimates of the regression coefficients are obtained by maximizing a pseudo-likelihood function over a sieve space. Such estimators are shown to be consistent and asymptotically normal with their asymptotic covariance matrix achieving the semiparametric efficiency bound. Profile likelihood methods are then used to numerically compute the asymptotic covariance matrix of the regression coefficients. Extension of the proposed methodology to several missing continuous covariates is also investigated and two separate approaches are proposed. Simulations are given to examine the finite sample properties of the estimates and a real dataset from a liver cancer clinical trial is analyzed using the proposed methods.

email: cindy.chen@vanderbilt.edu

A CAR-BART MODEL TO MERGE TWO DATASETS

Song Zhang*, University of Texas M.D. Anderson Cancer Center
Peter Muller, University of Texas M.D. Anderson Cancer Center
Tina Shih, University of Texas M.D. Anderson Cancer Center

One problem frequently encountered by public health researchers and health planners in the United States is the absence of socioeconomic data in many widely used and routinely collected sources of health and disease information. A common practice to solve this problem is to supplement individual-level record with the socioeconomic profile of the immediate neighborhood of the individual's residence. In this study, we are interested in the relationship between self-perceived health status and income, but they are only available in two different datasets. We use Bayesian additive regression trees (BART) to impute the missing income. The imputed income is further adjusted for spatial correlation through the implementation of a conditionally autoregressive (CAR) model. We demonstrate how to use the available BART libaray in R to make pure Bayesian inference based on samples from the posterior distribution.

email: yszhang@odin.mdacc.tmc.edu

A PSEUDOLIKELIHOOD METHOD FOR SEMIPARAMETRIC REGRESSION DATA WITH NONIGNORABLE NON-RESPONSE

Gong Tang*, University of Pittsburgh

Assume that the predictors are fully observed and the response variable is subject to missing values, standard maximum likelihood method requires modeling the missing-data mechanism when the missingness depends on the underlying value of the response. Tang, Little and Raghunathan (2003) proposed a pseudolikelihood method to analyze such data, with a parametric regression model for the complete data, and avoid specifying the missing-data mechanism. The method is extended to semiparametric regression models via adding a penalty term to address the nonparametric component of the regression. The properties of the estimators are to be discussed.

email: got1@pitt.edu

65. QUANTITATIVE-TRAIT LINKAGE ANALYSIS

POSITION-DEPENDENT CORRELATIONS IN EQTL MAPPING

Kwang-Youn A. Kim*, University of Iowa
Todd E. Scheetz, University of Iowa
Ruth Swiderski, University of Iowa
Alisdair R. Philp, University of Iowa
Thomas L. Casavant, University of Iowa
Edwin M. Stone, University of Iowa
Val C. Sheffield, University of Iowa
Jian Huang, University of Iowa

Recent advances in microarray technology have allowed us to examine expression patterns of thousands of genes. Analysis of gene expression data in related individuals have allowed us to identify locations of genetic elements that are responsible for variation in gene expression. This category of analysis is known as "expression as quantitative trait loci mapping" (often abbreviated as "eQTL mapping"). Previous studies have primarily examined the main effects of putative QTLs by performing a QTL analysis for each gene expression. There is a growing awareness that investigating the main effects can only explain a fraction of the total gene expression variation. Our study further extends the eQTL mapping into secondary effects by examining the correlations of gene expressions with the aid of genotypic information of the related individuals. We refer to this new methodology as position-dependent correlation. We have applied this method to a dataset consisting of gene expression values from 120 F2 rat eyes using Affymetrix® Rat Genome 230 2.0, and genotypes of 400 markers. We analyzed the correlations of eight Bardet-Biedl Syndrome (BBS) genes, which are members of a heterogeneous disease.

email: kwang-youn-kim@uiowa.edu

EFFICIENT MARKOV CHAIN MONTE CARLO ALGORITHMS FOR MAPPING GENOME-WIDE INTERACTING QTL

Nengjun Yi*, University of Alabama at Birmingham

Many complex traits are determined by multiple genetic and environmental influences. Gene-gene and gene-environment interactions play an important role in the genetic control and evolution of complex traits. Identification of genome-wide interacting quantitative trait loci (QTL) has been a daunting challenge, mainly due to huge model space. In this study, we propose a Bayesian model selection approach to identifying interacting QTL across the entire genome for complex traits in experimental crosses. The proposed method is able to model fixed or random covariates, and simultaneously detect gene-gene and gene-environment interactions. Computationally efficient Markov chain Monte Carlo (MCMC) algorithms are developed to sample from the posterior distribution. Statistical properties of the proposed algorithms are explored. We detail how to use prior knowledge to specify prior distributions for all unknowns. Several strategies to reduce the model space are incorporated into the proposed approach, allowing more rapid identification and exploration of important interactions. We illustrate the proposed method using two real data sets.

email: nyi@ms.soph.uab.edu

VARIABLE SELECTION FOR LARGE P SMALL N REGRESSION MODELS WITH INCOMPLETE DATA: MAPPING QTL WITH EPISTASIS

Min Zhang*, Purdue University
Dabao Zhang, Purdue University
Martin T. Wells, Cornell University

A Bayesian approach to select variables for multiple linear regression with large p, small n and incomplete data is presented. The most important characteristic of the parameter space under consideration is that, although high-dimensional, it is very sparse. Such information is incorporated into the prior via assuming a large probability mass concentrated at zero, by which the dimension of parameter space is reduced sufficiently. Thus our approach allows the simultaneous estimation of main effects as well as interactions within the same model. To account for possible asymmetry between positive and negative effects, the prior for non-zero effects is further decomposed into a mixture of two truncated normal distributions. In addition, our approach can naturally handle the imputation of missing data. The inference was carried out by using a Markov chain Monte Carlo (MCMC) sampling scheme. We evaluate the performance of our approach by simulation studies and illustrate the proposed method with a real data example.

email: minzhang@stat.purdue.edu

POOR PERFORMANCE OF BOOTSTRAP CONFIDENCE INTERVALS FOR THE LOCATION OF QUANTITATIVE TRAIT LOCI

Ani W. Manichaikul*, Johns Hopkins University
Karl W. Broman, Johns Hopkins University

The aim of many genetic studies is to locate the genomic regions (called quantitative trait loci, QTLs) that contribute to variation in a quantitative trait (such as time-to-death following bacterial infection). Confidence intervals (CIs) for the locations of QTLs are particularly important for the design of further experiments to identify the gene or genes responsible for the effect. Likelihood support intervals are the most widely used method to obtain CIs for QTL location, but the non-parametric bootstrap has also been recommended. Through extensive computer simulation, we show that bootstrap confidence intervals are poorly behaved and so should not be used in this context. Likelihood support intervals, on the other hand, behave appropriately. The profile likelihood (or LOD curve) for QTL location has a tendency to peak at genetic markers, and so the distribution of the maximum likelihood estimate (MLE) of QTL location has the unusual feature of point masses at genetic markers; this is the primary cause of the poor behavior of the bootstrap.

email: amanicha@jhsph.edu

NONPARAMETRIC FUNCTIONAL INTERVAL MAPPING OF QUANTITATIVE TRAIT LOCI

Jie Yang*, University of Florida
George Casella, University of Florida

Functional mapping is a powerful tool for detecting major genes responsible for different phenotypic curves by using a parametric functional form, usually derived from a biological law, to drive a maximum-likelihood-based test for a significant QTL (quantitative trait loci). However, in many situations there is no obvious functional form or not enough observations to fit the parametric form. So in these cases, this strategy will not be optimal. Here we propose to use nonparametric function estimation, typically implemented with B-splines, to estimate the underlying functional form of phenotypic trajectories, and then construct a nonparametric test to find evidence of existing quantitative trait loci. Using the representation of a nonparametric regression as a mixed model, we can easily derive a likelihood ratio test statistic throughout the linkage map. A simulation procedure can be adopted to decide the critical value. Other related statistical issues like how to estimate variance-covariance matrix are discussed. Simulation studies and application to a real dataset are provided to illustrate our method with comparison to other existing methods.

email: jyang81@ufl.edu

GENOMEWIDE FUNCTIONAL MAPPING FOR GENETIC CONTROL OF PROGRAMMED CELL DEATH: A SEMIPARAMETRIC MODEL

Yuehua Cui*, Michigan State University
Rongling Wu, University of Florida

Naturally-occurring' or 'programmed' cell death (PCD) in which the cell uses specialized cellular machinery to kill itself is a ubiquitous phenomenon that occurs early in organ development. Such a cell suicide mechanism that enables metazoans to control cell number and eliminate cells threatening the organism's survival has been thought to be under genetic control. In this article, we developed a novel statistical model for genomewide mapping specific genes or quantitative trait loci (QTL) that are responsible for the PCD process based on polymorphic molecular markers. This model incorporates the biological mechanisms of PCD that undergoes two different developmental stages, exponential growth and polynomial death. We derived a parametric approach to model the exponential growth and a nonparametric approach based on the Legendre function to model the polynomial death. A nonstationary model has been used to approximate the structure of the covariance matrix among cell numbers at a multitude of different times. The statistical behavior of our model is investigated through simulation studies and tested by a real example in rice.

email: cui@stt.msu.edu

NONLINEAR MIXED-EFFECT MIXTURE MODELS FOR FUNCTIONAL MAPPING OF LONGITUDINAL TRAITS

Wei Hou*, University of Florida
Rongling Wu, University of Florida

Nonlinear mixed-effects (NLME) models, aimed to model intra-and inter-individual variation in repeated measurements, have become a popular tool for longitudinal studies. In this talk, we will present an NLME mixture model for functional mapping of quantitative trait loci (QTL) that are responsible for differentiation in longitudinal traits. This model is constructed within a mixture model framework, with each mixture component assigned by biological rationale. The EM algorithm, implemented with the simplex algorithm, has been derived to estimate the parameters that construct longitudinal curves and covariance matrices. We used a real QTL mapping example for stem wood growth trajectories in poplar trees to demonstrate and validate this model. Our model provides a general plat for testing and studying the genetic architecture of longitudinal traits from a dynamic point of view.

email: whou@biostat.ufl.edu

66. SPATIAL-TEMPORAL MODELING

MULTIVARIATE SPATIOTEMPORAL MODELS FOR ENVIRONMENTAL EPIDEMIOLOGICAL DATA

Brian Reich*, North Carolina State University
Montserrat Fuentes, North Carolina State University
David Holland, U.S. Environmental Protection Agency

Motivated by the analysis of ambient particulate matter data, we develop a nonparametric Bayesian model for multivariate spatiotemporal data. Our model decomposes each site's multivariate time-series into a spatially-varying population curve and a site-specific deviate from the population curve. Each component is modeled using smoothing splines with a flat prior on the effective degrees of freedom. Applying these methods to data from the EPA's Air Quality System, we investigate the relationship between concentrations of fine and coarse particles and the association between these concentrations and mortality.

email: reich@stat.ncsu.edu

SPATIAL AND TEMPORAL ANALYSIS ON MISSOURI BLADDERPOD (LESQUERELLA FILIFORMIS)

William B. Leeds*, Truman State University
Elizabeth R. Bobzien, Truman State University
Hyun-Joo Kim, Truman State University
Michael I. Kelrick, Truman State University

Missouri bladderpod (Lesquerella filiformis) is a federally threatened plant species found in southwest Missouri and northern Arkansas. Understanding how its abundance changes in space and time, and in relation to its heterogeneous glade habitat, is crucial for successful conservation of the species. Abundance of Missouri bladderpod and percent cover of nine habitat attributes were recorded in four distinct years. Due to the nature of the data, both parametric and nonparametric approaches were considered. Logistic regression modeling, repeated measures ANOVA, and log-linear modeling were used in temporal analysis. For spatial analysis the data set was probed with a Geographical Information System software environment. The techniques and approaches used can be applied to many different problems in habitat modeling and conservation biology. Key Words: logistic regression, repeated measures ANOVA, log-linear modeling, conservation biology, Missouri bladderpod

email: wbl326@truman.edu

NONPARAMETRIC ESTIMATION OF CORRELATION FUNCTIONS IN LONGITUDINAL AND SPATIAL DATA, WITH APPLICATION TO COLON CARCINOGENESIS EXPERIMENTS

Yehua Li*, Texas A&M University
Naisyin Wang, Texas A&M University
Raymond J. Carroll, Texas A&M University

In longitudinal and spatial studies, observations often demonstrate strong correlations that are stationary in time or distance lags, and the times or locations of these data being sampled may not be homogeneous. We propose a nonparametric estimator of the correlation function in such data, using kernel methods. We develop a pointwise asymptotic normal distribution for the proposed estimator, when the number of subjects is fixed and the number of vectors or functions within each subject goes to infinity. Based on the asymptotic theory, we propose a weighted block bootstrapping method for making inference about the correlation function, where the weights account for the inhomogeneity of the distribution of the times or locations. The method is applied to a data set from a colon carcinogenesis study, in which colonic crypts were sampled from a piece of colon segment from each of the 12 rats in the experiment and the expression level of p27, an important cell cycle protein, was then measured for each cell within the sampled crypts. A simulation study is also provided to illustrate the numerical performance of the proposed method.

email: yehuali@stat.tamu.edu

PARTITIONING STATISTICAL EVIDENCE OF CAUSAL ASSOCIATION IN OBSERVATIONAL STUDIES: AN ILLUSTRATION USING SPATIOTEMPORAL DATA

Holly Janes*, Johns Hopkins Bloomberg School of Public Health
Scott L. Zeger, Johns Hopkins Bloomberg School of Public Health
Francesca Dominici, Johns Hopkins Bloomberg School of Public Health

In observational studies designed to evaluate a causal link between an exposure and outcome, unmeasured confounding is a widespread and principal concern. We propose partitioning the exposure into orthogonal components, and allowing each component to have a unique effect on the outcome. This allows the causal hypothesis to be evaluated; in simple situations, if the exposure causes the outcome, the effects of all exposure components will be the same. When exposure component effects vary, attention can be restricted to the effects that are thought to be less confounded. We show that the total exposure effect is a weighted average of the effects of its different components. This representation reveals the contributions of the various sources of statistical information. Our methods are illustrated using spatiotemporal data on particulate matter (PM) and mortality. PM is partitioned into its space, time, and space-by-time components. We focus on the time and space-by-time components which provide roughly equal amounts of information but have very different signs. Most of the evidence of a positive PM effect comes from the trend component, which is likely to be confounded. Our methods are useful more broadly for viewing and evaluating the sources of information for an exposure-outcome causal link.

email: hjanes@jhsph.edu

67. RECENT DEVELOPMENTS IN BAYESIAN BIOINFORMATICS

BAYESIAN CALIBRATION OF MASS SPECTRA

Cavan Reilly*, University of Minnesota

The need for calibration of mass spectra with regard to the mass/charge axis is a widely recognized problem. Several approaches have been proposed, but these suffer from a number of weaknesses. Here we discuss a Bayesian approach to the problem that not only allows for calibration, but can be used to estimate average spectra within a patient population. The approach allows for nonlinear calibration curves for each subject. The posterior modes of the curves are found via an EM algorithm, and these modes are then used for calibration.

email: cavanr@biostat.umn.edu

REGULATORY BINDING SITE DETECTION FROM HIGH-DENSITY SEQUENCE AND CHIP-CHIP ARRAY DATA USING HIDDEN MARKOV MODELS

Mayetri Gupta*, University of North Carolina at Chapel Hill
Jonathan Gelfond, University of North Carolina at Chapel Hill
Joseph Ibrahim, University of North Carolina at Chapel Hill

Chromatin immunoprecipitation followed by microarray hybridization (ChIP-chip) over high-density tiling arrays has been recently developed as a tool to measure protein-DNA interactions across the genome. Sites of interactions between the genome and a specific protein often tend to show a conserved sequence pattern (called a motif). However, genome-wide sequence data or tiling array data, when considered separately, frequently yield contradictory predictions of regulatory sites. We propose a model in which sequence and array data can be considered simultaneously to elicit meaningful binding site predictions, and show how the Bayesian formulation can successfully deal with hidden sources of bias and experimental artefacts.

email: gupta@bios.unc.edu

BAYESIAN ROBUST INFERENCE FOR DIFFERENTIAL GENE EXPRESSION

Raphael Gottardo*, University of British Columbia
Adrian Raftery, University of Washington
Ka Yee Yeung, University of Washington
Roger Bumgarner, University of Washington

We consider the problem of identifying differentially expressed genes under different conditions using gene expression microarrays. Because of the many steps involved in the experimental process, from hybridization to image analysis, cDNA microarray data often contain outliers. We develop a robust Bayesian hierarchical model for testing for differential expression. Errors are modeled explicitly using a t-distribution, which accounts for outliers. The model includes an exchangeable prior for the variances which allow different variances for the genes but still shrink extreme empirical variances. We compare our method to six other baseline and commonly used techniques, namely the t-test, the Bonferroni-adjusted t-test, Significance Analysis of Microarrays (SAM), Efron's empirical Bayes, and EBarrays in both its Lognormal-Normal and Gamma-Gamma forms. The method is illustrated with two publicly available gene expression data sets: one HIV experiment (cDNA microarray) and one spiked-in experiment (Affymetrix platform) with probe summary computed with RMA, gcRMA, MAS 5 and Dchip. Our method performs better than its competitors on each dataset, on the basis of false positives, false negatives and false discovery rate.

email: raph@stat.ubc.ca

68. TO POOL OR NOT TO POOL: SYSTEMATIC REVIEWS AND POOLED ANALYSES

CASE STUDIES OF THE USE OF META-ANALYSIS IN PHARMACOEPIDEMIOLOGY

Jesse A. Berlin*, Johnson and Johnson Pharmaceutical Research and Development

Meta-analysis is used to provide a quantitative summary of the results of multiple studies of the same clinical question. How broadly or narrowly to define the question is a matter of some debate. Meta-analysis is often used to address questions about effectiveness of a therapeutic intervention, but it is in the safety arena that combining data across studies may be particularly useful, especially when suspected adverse events are uncommon. A randomized trial of the use of erythropoietin (EPO) in metastatic breast cancer patients found an increased risk of mortality in the EPO arm. This ultimately led to a meeting of FDA's Oncologic Drug Advisory Committee (ODAC), in May of 2004, to review potential safety concerns about this drug class. Three pharmaceutical manufacturers presented to the committee. In this presentation, I will review data presented by the companies to ODAC and the FDA. I will present key results from the breast cancer trial to help set the context in which the meeting was held, and to help generate hypotheses about potential explanations for the increase in risk. Related data from previous trials will also be relevant to these hypotheses. Time permitting, I'll present other examples illustrating other uses of meta-analysis in pharmacoepidemiology, including uses early in the drug development process.

email: jberlin@prdus.jnj.com

RECENT ADVANCES IN SURROGATE ENDPOINT EVALUATION

Tomasz Burzykowski*, Hasselt University-Belgium

It often appears, that the most sensitive and relevant ("true") clinical endpoint, which might be used to assess the efficacy of the treatment, might be difficult to apply in a clinical trial. In such cases, a seemingly attractive solution is to replace the true endpoint by another one, which is measured earlier, more conveniently, or more frequently. Such "replacement" endpoints are termed "surrogate" endpoints (Ellenberg and Hamilton 1989). Before a clinical endpoint can be used as a surrogate, it should be "validated". In this paper, we discuss formal statistical methods that are useful to address the validation issue. In particular, we focus on the methodology proposed by Daniels and Hughes (1997), Buyse et al. (2000), and Gail et al. (2000), that use data from multiple clinical trials. Recently, the methodology have been extended in several ways. For instance, Baker (2005) and Korn, Albert, and McShane (2005) have developed new approaches. Burzykowski and Buyse (2005) have proposed a new measure, a so-called surrogate threshold effect, which is easier to interpret and provides more information than the measures suggested by Buyse et al. (2000). For the latter, Alonso (2005) have provided an interesting interpretation using concepts from the information theory.

email: tomasz.burzykowski@uhasselt.be

A SIMPLE META-ANALYTIC APPROACH FOR BINARY SURROGATE ENDPOINTS

Stuart G. Baker*, National Cancer Institute

In a standard meta-analysis, the goal is to estimate the effect of intervention on endpoint with a smaller variance than with an individual study, pooling is over the estimated effect of intervention on outcome, and increased sample size typically trumps heterogeneity to increase precision. In the simple surrogate-endpoint meta-analysis, the goal is to predict the effect of intervention on true endpoint in a target trial with only a surrogate endpoint, pooling is over the estimated predicted effect of intervention on true endpoint (which is based on the association of surrogate and true endpoints in each previous trial and the surrogate endpoint in the target trial), and heterogeneity typically makes a large contribution to variability. Validation of the surrogate endpoint approach is via the Average Prediction Error for the Predicted effect of intervention (APEP), which is a weighted sum of the absolute values of the differences between the predicted and observed intervention effects in a set of trials with both surrogate and true endpoints. APEP is compared with the average prediction error of a standard meta-analysis using only true endpoints and the average clinically meaningful difference in true endpoints implicit in the trials.

email: sb16i@nih.gov

69. NEW STATISTICAL METHODS FOR ESTIMATING MEDICAL EXPENDITURES AND COST EFFECTIVENESS FROM OBSERVATIONAL DATA

BAYESIAN COST EFFECTIVENESS ANALYSIS

Daniel F. Heitjan*, University of Pennsylvania

In recent years there has been considerable debate about how to conduct a proper cost-effectiveness analysis. Topics of interest have included which cost-effectiveness parameter to estimate and whether to approach the estimation from the Bayesian or frequentist perspective. Analyses are complicated by the need to model association between cost and effectiveness outcomes, which can be particularly difficult when some data are censored, as is common in randomized trials in cancer and cardiovascular disease. In this talk I describe a Bayesian methodology for estimating the cost-effectiveness of a new treatment compared to a standard in a clinical trial, when censoring of survival, the effectiveness variable, induces censoring of total cost. The statistical model assumes that survival follows a Weibull distribution and that total health care cost follows a gamma distribution whose mean has a linear regression on survival time. I summarize the posterior distributions of key parameters by importance sampling. I illustrate the method with an analysis of data from a randomized clinical trial of a treatment for cardiovascular disease.

email: dheitjan@cceb.upenn.edu

ESTIMATING MEDICAL EXPENDITURES FOR SMOKING-RELATED DISEASES VIA SMOOTH QUANTILE RATIO ESTIMATION

Francesca Dominici*, Johns Hopkins Bloomberg School of Public Health

The methodological development of this paper is motivated by a common problem in econometrics where we are interested in estimating the difference in the average expenditures between two populations, with and without a disease, as a function of the covariates. Smooth Quantile Ratio Estimation (SQUARE) is a novel approach for estimating the difference in average expenditures by smoothing across percentiles the log-transformed ratio of the two quantile functions. In this paper we extend SQUARE to a regression model. More specifically, we apply the basic definition of SQUARE to compare expenditures for the cases and controls having 'similar' covariate profiles. We determine strata of cases and control with 'similar' covariate profiles by use of propensity score matching. We then apply two-part regression SQUARE to the 1987 National Medicare Expenditure Survey to estimate the difference between persons suffering from smoking attributable diseases and persons without these diseases as a function of the propensity of getting the disease. Using a simulation study, we compare frequentist properties of two-part regression SQUARE with maximum likelihood estimators for the log-transformed expenditures.

email: fdominic@jhsph.edu

ESTIMATING THE COST-EFFECTIVENESS OF MEDICAL THERAPIES FROM OBSERVATIONAL DATA VIA PROPENSITY SCORES

Nandita Mitra, University of Pennsylvania
Alka Indurkhya*, University of Massachusetts

Statistical estimates of the cost-effectiveness of treatments using the net monetary benefit are used by policy makers to help make decisions on allocation of societal resources for competing medical treatments. Propensity score adjustment of the net monetary benefit (in a generalized linear model framework) has been recently shown to provide less biased estimates in the presence of significant differences in baseline measures and demographic characteristics between treatment groups in observational studies. We evaluate the sensitivity of propensity score adjusted estimates of net monetary benefits when: (a) important covariates are unobserved and (b) covariates, cost, and treatment components have different patterns of missing data . The methods are illustrated using Surveillance, Epidemiology and End Results -Medicare data of 49,375 elderly men diagnosed with localized prostate cancer.

email: nmitra@cceb.upenn.edu

COVARIATE ADJUSTMENT IN CENSORED COST DATA

Andrew R, Willan*, SickKids Research Institute and University of Toronto
Danyu Lin, University of North Carolina
Andrea Manca, University of York

We propose a system of seemingly unrelated regression equations to provide a general method for prognostic factor adjustment and subgroup analysis in cost-effectiveness studies with censored data. The method can be used in either an incremental cost-effectiveness ratio approach or an incremental net benefit approach, and does not require that the set of covariates for costs and effectiveness be the same. The use of covariate adjustment is essential for observational studies, since patient groups are likely to differ with respect to factors affecting both costs and clinical outcomes. In clinical trials, however, regression models are used most effectively for identifying treatment by prognostic factor interactions in the examination of subgroup effects. We propose using inverse probability weighting for parameter estimation. The measure of effectiveness can be either survival time or quality-adjusted survival time. Inverse probability weighting is required because, even when censoring is completely at random, censoring on the cost scale and the quality-adjusted survival scale will be informative. The methods will be illustrated using data from recent clinical trials.

email: andy@andywillan.com

70. STATISTICAL CHALLENGES IN PRE-CLINICAL PHARMACEUTICAL RESEARCH

ANALYSIS IN PRECLINICAL PHARMACEUTICAL RESEARCH: CHALLENGES AND OPPORTUNITIES

J. Alan Menius*, GlaxoSmithKline

The role of the preclinical statistician in the pharmaceutical industry has changed dramatically during the last decade. High throughput screening technologies are now used to screen a company's entire compound collection. Gene chip technologies create thousands of measurements per experimental observation. The statistical challenges include: optimizing the experimental procedures, data normalization, variable selection, multiple testing, combining data from multiple 'omics platforms, and finally testing hypotheses and creating new ones using these data. Preclinical statistician must be familiar with numerous techniques such as partial least squares, variable clustering, recursive partitioning, support vector machines and bayesian networks. Though the challenges are great, the opportunities for statisticians have never been better. Through the development and application of analysis procedures, statisticians have become a critical member of any scientific team hoping to harness the information contained in the gigabytes of data being generated in pharmaceutical preclinical research.

email: alan.j.menius@gsk.com

HOW TO FIND DRUGS WITH TREES: APPLICATIONS OF ENSEMBLE METHODS IN QSAR MODELING

Andy Liaw*, Merck Research Laboratories
Christopher Tong, Merck Research Laboratories
Ting-chuan Wang, Merck Research Laboratories
Vladimir Svetnik, Merck Research Laboratories

Ensemble learning methods such as bagging, boosting, and random forests have been shown to be very effective for building predictive models. We present how these methods have been applied in some areas of drug discovery research. In particular, we will discuss their applications in high throughput screening, structure-activity relationship (SAR) elucidation, and molecular profiling. Some comparison with other popular methods such as k-nn, PLS, SVM, and Naive Bayes will be presented.

email: andy_liaw@merck.com

SEARCHING FOR OPTIMUMS IN HIGH DIMENSIONAL SPACE: AN APPLICATION OF THE SELC ALGORITHM IN DRUG DISCOVERY

Kjell Johnson*, Pfizer, Inc.
Abhyuday Mandal, University of Georgia
C.F. Jeff Wu, Georgia Institute of Technology

Mandal, Wu, and Johnson (in press, Technometrics) present a novel method for identifying optimal factor combinations in high dimensional space. This technique is a modification of Wu, Mao, and Ma's (1990) sequential elimination of levels (SEL) method and incorporates concepts of forbidden arrays and genetic algorithms to efficiently find optimal combinations. In this presentation, we will explain the sequential elimination of level combinations (SELC) method and will illustrate its usefulness on an example from combinatorial chemistry.

email: kjell.johnson@pfizer.com

OPEN ISSUES IN THE CLASSIFICATION OF OLIGONUCLEOTIDE MICROARRAY DATA

Max Kuhn*, Pfizer Global Research and Development

A drug discovery case study will be presented where a model is developed to classify candidate antibiotic compounds into one of eight methods of action (e.g. protein synthesis inhibitors) based on the results of Affymetrix gene chip assays. This experiment will be used to illustrate open issues and potential solutions.

e-mail: max.kuhn@pfizer.com

71. IMS: MEDALLION LECTURE

SHRINKAGE ESTIMATION: AN EXPANDING STATISTICAL THEME

Lawrence D. Brown*, University of Pennsylvania

Stein (1955) suprised the statistical world with his discovery that the ordinary least squares estimator of a multivariate normal mean is not admissible in the usual setting. James and Stein (1961) then produced their classic estimator which often provides significant improvement over the ordinary estimator. 'Shrinkage' is a core feature of the estimator. Although not immediately apparent this core idea is present in a wide and growing variety of contemporary statistical settings. It can be found in fixed effects linear models, random effects models, longitudinal and panel data, spatial statistical models, time series models, nonparametric regression and density estimation and even in versions of ordinary linear regression. These connections will be surveyed in this talk. The empirical Bayes interpretation was first proposed by Stein (1962) and effectively exploited in this setting by Lindley and Smith (1972), Efron and Morris (1972) and others. This interpretation and its hierarchical fully Bayes first cousin, as first developed for this problem by Strawderman (1972), provide an important link to the manifestations of shrinkage in the various contemporary methodologies. Some of the basic general points in this talk will be illustrated with two interesting data analyses.

email: lbrown@wharton.upenn.edu

72. SEMIPARAMETRIC AND NONPARAMETRIC METHODS IN LONGITUDINAL AND SURVIVAL ANALYSIS

IDENTIFYING LATENT CLUSTERS OF VARIABILITY IN LONGITUDINAL DATA

Michael R. Elliott*, University of Michigan School of Public Health

Means or other central tendency measures are by far the most common focus of statistical analyses. However, as Carroll (2003) noted, "systematic dependence of variability on known factors" may be of more fundamental concern in certain settings. We develop a latent class model that relates underlying "clusters" of variability to baseline or outcome measures of interest. Because estimation of variability is inextricably linked to estimation of trend, assumptions about underlying trends are minimized by using nonparametric regression estimates. The resulting residual errors are then clustered into unobserved classes of variability that are in turn related to subject-level predictors of interest. An application is made to psychological affect data.

email: mrelliot@umich.edu

BAYESIAN SEMIPARAMETRIC REGRESSION FOR LONGITUDINAL BINARY PROCESS DATA

Li Su*, Brown University
Joseph W. Hogan, Brown University

Longitudinal studies with binary repeated measures are widespread in biomedical research. Marginal regression approaches for balanced binary data have been developed while for binary process data, where measurement times are irregular and may differ by individuals, likelihood-based methods for marginal regression analysis are less well developed. In this article, we develop a Bayesian regression model for analyzing longitudinal binary process data. Our model specifies both the marginal mean and serial correlation structures semiparametrically. In particular, serial correlation is allowed to depend on the time lag between adjacent outcomes nonparametrically. When data are missing at random (MAR) and a correctly specified joint distribution for binary processes is required, the proposed semiparametric model gives considerable flexibility. Estimation and inference proceed by a fully Bayesian approach. The methods are illustrated using longitudinal viral load data from the HIV Epidemiology Research Study (HERS) and the performance is evaluated using simulations.

email: lisu616@gmail.com

SEMIPARAMETRIC ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE OBSERVATION TIMES

Jianguo Sun, University of Missouri-Columbia
Do-Hwan Park*, University of Nevada-Reno
Liuquan Sun, Chinese Academy of Sciences
Xingqiu Zhao, McMaster University

Statistical analysis of longitudinal data is an important topic faced in a number of applied fields including epidemiology, public health and medicine. In general, the information contained in longitudinal data can be divided into two parts. One is the set of observation times that can be regarded as realizations of an observation process and the other is the set of actually observed values of the response variable of interest that can be seen as realizations of a longitudinal or response process. For their analysis, a number of methods have been proposed and most of them assume that the two processes are independent. This greatly simplifies the analysis since one can rely on conditional inference procedures given the observation times. However, the assumption may not be true in some applications. We will consider situations where the assumption does not hold and propose a semiparametric regression model that allows the dependence between the observation and response processes. Inference procedures are proposed based on the estimating equation approach and the asymptotic properties of the method are established. The results of simulation studies will be reported and the method is applied to a bladder cancer study.

email: dopark@unr.edu

ON SEMIPARAMETRIC REGRESSION MODELS FOR NONHOMOGENEOUS BIRTH-BIRTH PROCESS

Hao Liu*, University of California-Davis

We consider situations where patients can experience two distinct types of recurrent events, and covariate effects are of primary interest. Bivariate counting processes with proportional intensity models are flexible for this situation, but usually based on a working independence assumption for nonhomogeneous Poisson processes. In this paper, we study the semiparametric regression models via nonparametric MLE based on a non-homogeneous birth-death process assumption, which introduces dependence between the two types of recurrent events and yields relatively simple estimation equations. The consistency and asymptotic normality of the estimators are derived and confirmed by a simulation study. The idea is illustrated by analyzing a real data example.

email: liuhao888@gmail.com

A BAYESIAN MODEL FOR SPARSE FUNCTIONAL DATA

Wesley K. Thompson*, University of Pittsburgh
Ori Rosen, University of Texas-El Paso

We propose a Bayesian methodology for analyzing data which consist of curves on multiple individuals,i.e., functional data. The curves are fit using a mixed-effects model for a B-spline basis. A Bayesian MCMC algorithm fits the parameters of this model while automatically selecting the local-level of smoothing through a latent indicator variable knot selection scheme. We also derive Bayesian posterior credible intervals for both overall mean and individual curves. This methodology is applicable to situations where the curves are sampled sparsely and/or at irregular timepoints. Covariate information can also be incorporated into the model. Other extensions of this model are considered. The proposed methodology is demonstrated via Monte Carlo simulation studies and by application to functional data on body fat percentage of adolescent girls pre- and post-menarche.

email: wesleyt@pitt.edu

73. SPATIAL MODELING OF DISEASE

A SPATIAL SCAN STATISTIC FOR ORDINAL DATA

Inkyung Jung*, Harvard Medical School-Harvard Pilgrim Health Care
Martin Kulldorff, Harvard Medical School-Harvard Pilgrim Health Care
Ann C. Klassen, Johns Hopkins Bloomberg School of Public Health

Spatial scan statistics are widely used for count data to detect geographical disease clusters of high or low incidence, mortality or prevalence and to evaluate their statistical significance. Some data are ordinal or continuous in nature, however, so that it is necessary to dichotomize the data to use a traditional scan statistic for count data. There is then a loss of information and the choice of cut-off point is often arbitrary. In this paper, we propose a spatial scan statistic for ordinal data, which allows us to analyze such data incorporating the ordinal structure without making any further assumptions. The test statistic is based on a likelihood ratio test and evaluated using Monte Carlo hypothesis testing. The proposed method is illustrated using prostate cancer grade and stage data from the Maryland Cancer Registry.

email: inkyung_jung@harvardpilgrim.org

CANCER CLUSTER DETECTION IN SEMI-PARAMETRIC MODELS WITH RANDOM EFFECTS: A SCORE-BASED TESTING APPROACH

Matteo Bottai, Arnold School of Public Health-University of Southern California
Marco Geraci*, Arnold School of Public Health-University of Southern California

We develop methods for spatial multilevel semi-parametric models for relative risk surface estimation, in which the spatial correlation is modeled through splines with random coefficients associated to a set of knots. A score-based test statistic is derived, partially by applying some of the results available for singular information problems. The proposed procedure is applied to testing that the variance of the random effects vanishes, that is, no clustering. Confidence intervals for the variance parameter are also derived. Its properties are assessed across different scenarios through an extensive simulation. Across all the scenarios, this overall, general test shows correct levels, and is highly sensitive to clustering. Once a departure is detected, a second, finer grid of knots can be superimposed on the existing grid, and the proposed procedure can be applied to test the homogeneity within two or more sub-areas. Along with the simulated dataset, the models developed were applied to the lip cancer dataset from former East Germany in 1980-1989, which is comprised of observations over 219 small-areas. The dataset had been previously analyzed by others. The inference based on our models appears consistent with theirs, providing greater insight about local clustering and comparison of geographical sub-areas.

email: Mbottai@gwm.sc.edu

SPATIAL ASSOCIATION ON CANCER INCIDENCE IN THE COMMUNITY SURROUNDING THE ROCKETDYNE FACILITY IN SOUTHERN CALIFORNIA

Sunkyung Yu*, University of Michigan
Jennifer B. Dimmer, University of Michigan
Hal Morgenstern, University of Michigan

The Santa Susana Field Laboratory (SSFL) and associated facilities of Atomics International (AI) and Rocketdyne, formerly a division of Rockwell International and now a division of the Boeing Company, are located in and around the Simi Hills of Ventura and Los Angeles counties in Southern California. The previous reports of increased cancer incidence among the workers at Rocketcyne / AI and in the surrounding community certainly warrant further examination of the potential cancer risk associated with proximity of SSFL facilities. In the response to strong community concerns about the continued presence of radioactive and toxic substances and the possibility of excess cancer occurrence in the area surrounding the SSFL, we conduct a study to examine cancer incidence in Ventura and Los Angeles Counties in relation to distance from SSFL using census block information provided by the California Tumor Registry. We consider a multivariate spatial regression model based on a Poisson log-linear model controlling with some confounding factors, such as age, gender, and race/ethnicity, and model the spatial structured components which stands for a spatial process. We apply a Bayesian hierarchical model with a hierarchical prior in drawing inferences for spatial and temporal trends in the incidence rate for each census block in the surrounding the SSFL.

email: skyu@umich.edu

GEOSTATISTICAL HIERARCHICAL MODEL FOR TEMPORALLY INTEGRATED DATA MEASURED WITH ERROR

Brian J. Smith*, The University of Iowa
Jacob J. Oleson, The University of Iowa

In the search for important determinants of disease, epidemiologists often face the task of retrospectively estimating past exposures of interest. Such is the case in modern studies of the effect on lung cancer risk of residential radon - a naturally occurring environmental gas that is correlated over space and time. Exposure assessment is limited in these epidemiologic studies because radon measurements are not available for the locations at which individuals spent time prior to enrollment. In such settings, there is a need for prediction at unmeasured geographic sites and time periods. We develop a hierarchical Bayesian geostatistical model for predicting unmeasured radon concentrations over space and time. Our work arises from a study of residential radon in Iowa, where measurements were taken as yearly averages and subject to detector measurement error. Much focus has been given lately to geostatistical methods for data that are obtained as integrated averages over geographic regions. We show how these techniques work in the time domain as well. Unlike the numerical approximations that are needed to integrate over geographic regions, we also provide closed-form solutions for the integration that must be performed over temporal periods.

email: brian-j-smith@uiowa.edu

SIGNAL QUALITY MEASUREMENTS FOR CDNA MICROARRAY DATA

Tracy L. Bergemann*, University of Minnesota
Lue Ping Zhao, Fred Hutchinson Cancer Research Center

Concerns about the reliability of transcription rates estimates derived from microarray data inspires ongoing research into measurement error in these experiments. Error happens at both the technical level within the lab and the experimental level. Here, we are concerned with measuring spot-specific error. We outline two different approaches to quantify the reliability of spot-specific intensity estimates. In both cases, the spatial correlation between pixels and its impact on spot quality is accounted for. The first method is a straightforward parametric estimate of within-spot variance that assumes a Gaussian distribution and accounts for spatial correlation via an overdispersion factor. The second method employs a non-parametric quality estimate referred to throughout as the mean square prediction error (MSPE). The MSPE first smooths a pixel region and then measures the difference between actual pixel values and the smoother. Both methods herein are compared for real and simulated data to assess numerical characteristics and the ability to describe poor spot quality. Further, we examine the use of spot quality estimates to increase efficiency in downstream analysis.

email: berge319@umn.edu

APPROXIMATE METHODS IN BAYESIAN POINT PROCESS SPATIAL MODELS

M. M. Hossain*, University of South Carolina
Andrew B. Lawson, University of South Carolina

A range of point process models have been developed for spatial epidemiology applications, where case event data arises, based largely on a heterogeneous Poisson Process formulation. In addition, Diggle and Rowlingson (1994) proposed a conditional logistic regression formulation, in relation to analyse Larynx cancer data in order to assess a putative hazard source (incinerator) effect. These unconditional and conditional models can be flexibly extended with random effects which include prior spatial correlation. However, the implementation of such models has been limited due in part to the difficulty of dealing with normalising integrals in the unconditional model and the general lack of suitable software. In this paper a range of approaches to random effect modeling for case event data are considered, based on different approximations. The ability of the methods to recover known parameter values from simulations is also examined. One advantage of the methods described is that all the statistics for inference are readily interpretable and implementation in WinBUGS is straightforward.

email: hossain@gwm.sc.edu

EFFECTS OF AIR POLLUTION, SOCIAL ECONOMICAL STATUS, AND SPATIAL CLUSTERING EFFECTS ON LUNG CANCER MORTALITY RATES OF NORTH CAROLINA COUNTIES OF 2000

Kuo-Ping Li*, University of North Carolina at Chapel Hill
Chirayath Suchindran, University of North Carolina at Chapel Hill

A certain air pollution species can affect lung function and thus may increase the risk of lung cancer. Other risk factors of lung cancers include smoking as well as life style-related factors. These factors may or may not be related to a community's social economical status (SES). In this study, we propose a Hierarchical Bayesian model that links county mortality rates due to lung cancer, measurements of county air pollution emissions, and a certain SES (per capita income, education levels such as number of college graduates, etc). The possible link between spatial clustering and these mortality rates is also investigated by including the effects of spatial association through a conditionally autoregressive (CAR) term in the model. We studied the lung cancer mortality rates in 100 North Carolina counties of year 2000. It is found that, in general, the pollutions emission does not have significant effects on the county lung cancer mortality rates, while some SES has a certain degree of effects. We also found that spatial clustering plays an important role in the variability of these mortality rates.

email: kpli@email.unc.edu

74. MULTIPLE TESTING AND FALSE DISCOVERY RATES

A MODIFIED SPATIAL SCAN STATISTIC FOR CLUSTER DETECTION

Ronald E. Gangnon*, University of Wisconsin-Madison

The spatial scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) is a widely applied tool for cluster detection. The 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 the standard Bonferroni adjustment. In many settings, the extent of the multiple comparisons problem varies across the study region. For example, urban areas have many overlapping clusters, while rural areas have few. The spatial scan statistic provides a single uniform multiplicity adjustment. We propose new two spatially-varying multiplicity adjustments for spatial cluster detection, one based on a nested Bonferroni-type adjustment and one based on local averaging. The geography of power for the three statistics is explored through simulation studies, and the tests 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

SCREENING AND REPLICATION USING THE SAME DATA SET: TESTING

Amy J. Murphy*, Harvard School of Public Health
Matthew B. McQueen, Harvard School of Public Health
Benjamin A. Raby, Brigham and Women's Hospital-Harvard Medical School
Kady Schneider, Harvard School of Public Health
Jessica Su, Harvard School of Public Health
Juan Celedon, Harvard School of Public Health
Edwin K. Silverman, Brigham and Women's Hospital-Harvard Medical School
Nan M. Laird, Harvard School of Public Health
Scott T. Weiss, Brigham and Women's Hospital-Harvard Medical School
Christoph Lange, Harvard School of Public Health

In a recent methodological development for genome-wide association studies, two-stage testing strategies have been proposed for quantitative traits in family-based designs (Van Steen et al., 2005a). Such testing strategies use the same data set for genomic screening and replication, outperforming standard methodology, such as false-discovery rate (FDR, Benjamini and Hochberg, 1995), by several magnitudes, making genome-wide association feasible, even for relatively moderate sample sizes. However, the proposed testing strategies suffer from the limitation that they require phenotypic variation (Lange et al., 2003a,b) in the trait of interest and therefore cannot be applied to the primary phenotype of most family-based studies, affection status. In many family-based studies, only affected probands are recruited and affection status is the primary end point of the study. In this paper, we present a novel two-stage testing strategy that can be applied in such situations. It also uses the same data set from genomic screening and replication, but requires neither variation in the phenotype nor specification of a phenotypic mean model. We estimate the power of this approach empirically via simulation studies. The practical relevance is illustrated by an application to a partial genome-scan of an asthma study.

email: amurphy@hsph.harvard.edu

AN ADAPTIVE ALPHA-SPENDING ALGORITHM IMPROVES THE POWER OF STATISTICAL INFERENCE IN MICROARRAY DATA ANALYSIS

Jacob P. Brand*, Pennington Biomedical Research Center
Lan Chen, University of Birmingham at Alabama
Xiangqin Cui, University of Birmingham at Alabama
Alfred A. Bartolucci, University of Birmingham at Alabama
Grier P. Page, University of Birmingham at Alabama
Kyoungmi Kim, University of Birmingham at Alabama
Stephen Barnes, University of Birmingham at Alabama
Vinodh Srinivasasainagendra, University of Birmingham at Alabama
Mark T. Beasley, University of Birmingham at Alabama
David B. Allison, University of Birmingham at Alabama

We propose an adaptive alpha-spending algorithm that incorporates additional contextual evidence (including correlations among genes) about differential expression by means of assigning gene-specific significance levels to each gene resulting in alpha-spending adjusted p-values. We have shown that the Bonferroni correction applied to the alpha-spending adjusted p-values approximately controls the Family Wise Error Rate under the complete null. Under certain conditions it is plausible that the method in (Benjamini and Hochberg, 1995) applied to the alpha spending adjusted p-values also controls the False Discovery Rate (FDR). Although here we only discuss the alpha-spending algorithm for inferential testing in microarray data analysis, this algorithm may be helpful in any situation in High Dimensional Biology involving many hypothesis tests. We compared the power between the original p-values and post-processed p-values for 700 simulated genes simulated under a variety of conditions including: different correlations among differentially expressed genes, different sample sizes and different levels of differential expression. The simulation study confirms that application of the method in (Benjamini and Hochberg, 1995) controls the FDR. We found that on average the adaptive alpha-spending algorithm yielded higher power and that it yielded its greatest benefits with increasing sample sizes and correlation among genes.

email: BrandJP@pbrc.edu

THE EFFECT OF CORRELATED GENE EXPRESSION ON TESTS OF FUNCTIONAL CATEGORY ENRICHMENT

William T. Barry*, University of North Carolina
Andrew B. Nobel, University of North Carolina
Fred A. Wright, University of North Carolina

A primary application of microarrays has been the identification of differential expression across a set of experimental conditions. With the availability of annotation databases such as GO and Pfam, several methodologies have been proposed to test the involvement of entire functional categories. By comparing the genes within a category against their complement, traditional parametric and non-parametric tests have been used to detect either a difference in the proportion of genes called significant (Chi-squared or Fisher's Exact test), or a shift in their average differential expression (Z-test or Wilcoxon rank sum). These methods falsely assume independence among the genes, such that positive gene-gene correlation inflates the variances of statistics, leading to anti-conservative tests. Permutation analysis of a real microarray dataset is used in a manner that preserves gene-gene correlations, in order to quantify the true error rates of these methods. For 1823 GO and Pfam categories, histograms of realized p-values demonstrate increases in Type I error above nominal levels, while the minimum p-value for the four tests passed a Bonferroni correction to alpha=0.05 in 39.5%, 36.5%, 97.2% and 96.9% of the realizations, respectively. These results indicate a fully permutation-based approach to testing is more appropriate for the data structure in microarrays.

email: wbarry@bios.unc.edu

A NOTE ON USING PERMUTATION BASED FALSE DISCOVERY RATE ESTIMATE TO COMPARE DIFFERENT ANALYSIS METHODS FOR MICROARRAY DATA

Yang Xie*, University of Minnesota
Wei Pan, University of Minnesota
Arkady B. Khodursky, University of Minnesota

In practice, with the true False discovery rate (FDR) unknown, an estimated FDR can serve as a criterion to evaluate the performance of various statistical methods under the condition that the estimated FDR approximates the true FDR well, or at least, it does not improperly favor or disfavor any particular method. Permutation methods have become popular to estimate FDR in genomic studies. The purpose of this paper is two-fold. First, we investigate theoretically and empirically whether the standard permutation based FDR estimator is biased, and if so, whether the bias inappropriately favors or disfavors any method. Second, we propose a simple modification of the standard permutation to yield a better FDR estimator, which can in turn serve as a more fair criterion to evaluate various statistical methods. The results show that the standard permutation method over-estimates FDR. The over-estimation is the most severe for the sample mean statistic while the least for the t-statistic with the SAM-statistic lying between the two extremes, suggesting that one has to be cautious when using the standard permutation-based FDR estimates to evaluate various statistical methods. In addition, our proposed FDR estimation method is simple and outperforms the standard method.

email: yangxie@biostat.umn.edu

FALSE DISCOVERY RATE ADJUSTMENT FOR TREE MODELS

Carol J. Etzel*, University of Texas-M.D. Anderson Cancer Center
Sumesh Kachroo, University of Texas-M.D. Anderson Cancer Center

Alternative pruning rules are useful in pruning back large disease outcome trees. However, the need to control for tree-wise type 1 error while still maintaining power to identify disease-related nodes is crucial. In this investigation, we compare power to detect disease-related nodes and tree-wise type 1 error rates (for null nodes) for a false discovery rate adjustment method and an alpha-allocation method via simulation of a disease-outcome tree. We then apply these methods to a lung cancer case/control study to identify high-risk subgroups while controlling for false discovery.

email: cetzel@mdanderson.org

FINITE SAMPLE PROPERTIES OF ESTIMATORS OF THE FALSE DISCOVERY RATE

Naim U. Rashid*, Duke University
Anindya Roy, University of Maryland

Multiple testing and the False Discovery Rate (FDR) are currently some of the most active research areas in statistics and bioinformatics. The increasingly complex structure of modern biomedical datasets demands the use of such new statistical methodologies to yield greater power and accuracy in the multiple hypothesis testing setting. Microarray, gel imaging, and other types of high-throughput systems often create such complex datasets, containing data from thousands of different samples from two or more treatment groups. To control the increase in global statistical error often observed in multiple hypothesis testing situations, researchers have introduced methods such as the Bonferroni correction, FDR, positive FDR (pFDR) and others. These concepts help in reducing, for example, the error rate of discovering statistically significant but false biological differences between corresponding samples within such datasets. We introduce an empirical analysis of the sensitivity of the FDR with respect to parameter specifications in several commonly used models through statistical simulation. We also apply the methodology to simulated two-dimensional gel electrophoresis data to demonstrate the efficacy of such methods in enhancing the accuracy of discovering biologically relevant differences in protein expression. Furthermore, we introduce methodology using maximum likelihood estimation to enhance the accuracy of estimators for the FDR.

email: nur2@duke.edu

75. COMPETING RISKS AND CURE RATES

IMPACT OF CHANGE IN LEVEL OF RISK FACTOR(S) AND PROPORTION OF CURED/IMMUNE INDIVIDUALS ON THE POPULATION ATTRIBUTABLE RISK: A SIMULATION BASED STUDY

Jayawant N. Mandrekar*, Mayo Clinic
Melvin L. Moeschberger, The Ohio State University

An important question in the assessment of a clinical trial is to determine whether treatment cures some patients. In the public health field, it is also important to find out the amount of disease burden in a population that could be eliminated or reduced, if risk factors were eliminated or reduced. Cure/immune models can estimate cure rate when long-term survivors are present and consist of two parts; an accelerated failure time model for survival time of susceptible patients and a logistic model for cure/immune proportion estimation. The goals of our research were to assess the cure rate and the impact of the change in the level of risk factor (or factors) and the proportion of cured/immune individuals in the population on the population attributable risk (PAR). Our simulations demonstrated that the PAR estimates increase (or decrease) with the increasing (or decreasing) levels of the risk factor, but remain approximately unchanged (as expected) with increases or decreases in the cured/immune proportion. The PAR is underestimated by a Cox proportional hazards model in the presence of cured/immune individuals unlike a cure model. In conclusion, the results from our approach are more precise when cured/immune individuals are present.

email: mandrekar.jay@mayo.edu

BAYESIAN ADDITIVE-MULTIPLICATIVE CURE RATE MODEL

Guosheng Yin*, M. D. Anderson Cancer Center–The University of Texas
Luis E Nieto-Barajas, Departmento de Estadistica ITAM

We propose a new class of additive-multiplicative cure rate models for right-censored failure time data. The new class of models offers a wide variety of innovative modeling structures and it provides a natural and formal way to examine the existence of a survival fraction. An inherent parameter constraint needs to be incorporated into the model formulation due to the additive effects of covariates. Within the Bayesian paradigm, we take a Markov gamma process prior with a moving partition to model the baseline hazard rate nonparametrically. A Markov chain Monte Carlo computational scheme is implemented for sampling from the full conditional distributions of the parameters. This family of models is illustrated with a real dataset involving a bone marrow transplantation study.

email: gsyin@mdanderson.org

EFFECTS OF COMPETING CAUSES OF DEATH IN MA.17, A PLACEBO-CONTROLLED TRIAL OF LETROZOLE AS EXTENDED ADJUVANT THERAPY FOR BREAST CANCER PATIENTS

Daniel Q. Meng*, NCIC-CTG, Queen's University-Canada
Judith-Anne W. Chapman, NCIC-CTG, Queen's University-Canada
Lois E. Shepherd, NCIC-CTG, Queen's University-Canada
Wendy Parulekar, NCIC-CTG, Queen's University-Canada
James N. Ingle, Mayo Clinic
Paul E. Goss, Massachusetts General Hospital Cancer Center

A patients likelihood of dying from breast cancer (BC) can be confounded by the existence of competing risks: dying from other malignancy (OM) or dying from other causes (OC). A supplementary study of the clinical trial data MA.17 consisting of 5170 breast cancer patients who, after preliminary surgery, had received five years previous adjuvant tamoxifen before registered in the clinical trial was conducted. The effects of 11 exploratory factors on the overall survival and cause-specific survival were examined by employing Cox, log-normal, and Weibull models. Logistic regression model was used to assess the probability of cause-specific death. The study showed that, after more or less five years tamoxifen, non-breast cancer deaths accounts for more than 75% of total death in patient group older than 70 years of age. Though older patient had shorter survival time to death from BC, the death rate of BC in two age groups were not different much. In addition to significant age effect, nodal status, osteoporosis, and cardiovascular disease had significant positive effect on the probability of dying from BC, OM, and OC respectively.

email: mengdan@gmail.com

FLEXIBLE CURE RATE MODELING UNDER LATENT ACTIVATION SCHEMES

Freda W. Cooner*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Bradley P. Carlin, University of Minnesota
Debajyoti Sinha, Medical University of South Carolina

Survival models have been and continue to be extremely popular in analyzing time-to-failure data, where failure is disease relapse or death. With rapid improvements in medical treatment and health care, many survival data sets now reveal a substantial portion of patients who are cured. Extended survival models called cure rate models account for the probability of a subject being cured. Popular cure models can be broadly classified into the classical mixture models of Berkson and Gage or the hierarchical classes of Chen, Ibrahim and Sinha. Recent developments in formulating Bayesian hierarchical cure models have evoked significant interest regarding relationships and preferences between these two classes of models. Our present work proposes a unifying class of cure rate models that facilitates flexible hierarchical model-building while including both existing cure model classes as special cases. This unifying class also elucidates the relationship between classical and hierarchical cure models. Issues such as regressing on the cure fraction and propriety of the associated posterior distributions under different modelling assumptions are also discussed. Finally, we offer a simulation study and also illustrate with two data sets (one on melanoma and the other on breast cancer) that reveal our model's ability to distinguish among underlying mechanisms that lead to relapse and cure.

email: xiyunwu@yahoo.com

PARAMETRIC REGRESSION ON CUMULATIVE INCIDENCE FUNCTION

Jong-Hyeon Jeong*, University of Pittsburgh
Jason Fine, University of Wisconsin-Madison

It is noted that an improper distribution needs to be considered for the baseline distribution for parametric regression in competing risks analysis. We propose a simple form of Gompertz distribution to be the improper baseline distribution of the events of interest. Maximum likelihood inference on regression parameters and associated cumulative incidence function is proposed under a flexible link function of the generalized odds-rate model. We also note that estimation and prediction of the cured proportion of patients with cause-specific events is straightforward under the parametric setting. Simple parametric test statistics are also discussed for testing model assumptions such as proportional hazards and proportional odds for the events of interest between groups. The proposed parametric regression modelling is applied to analyze local or regional recurrences in breast cancer data.

email: jeong@nsabp.pitt.edu

MODELING BIVARIATE COMPETING RISK EVENTS VIA MARKOV CHAINS

Mireya Diaz*, Case Western Reserve University

Complex time-to-event analyses commonly arise in medical data. One such scenario is the one posed by bivariate competing risk events such as competing outcomes in similar organs of an individual. Under this setting, a progressive Markov chain model provides an elegant and flexible formulation. Via simulations, the effects of parameters such as correlation between events, rate of competing risks, censoring and sample size on the estimates of transition probabilities are assessed. The simulation scheme is based on a series of independent latent waiting times for each cause of failure and cluster component. When correlation between cluster components is weak the stationary distribution in terms of the failure cause tends to the distribution of an offspring's locus from heterozygous parents, and the events rarely occur simultaneously. Sample size and censoring affect the bias of the estimates in opposite directions under this condition. When the correlation increases, the stationary distribution loses this equilibrium leaning towards simultaneous and same-cause events. Future work will extend the model to deal with cluster asymmetry, new clocks, and covariates.

email: mcd8@cwru.edu

76. GENE EXPRESSION ANALYSIS

FLEXIBLE TEMPORAL EXPRESSION PROFILE MODELLING USING THE GAUSSIAN PROCESS

Ming Yuan*, Georgia Institute of Technology

Time course gene expression experiments have proved valuable in a variety of biological studies (e.g., Chuang et al., 2002; Edwards et al., 2003). A general goal common to many of these time course experiments is to identify genes that exhibit different temporal expression profiles across multiple biological conditions. Such experiments are, however, often hampered by the lack of data analytical tools. Taking advantage of the great flexibility of Gaussian processes, we propose a statistical framework for modelling time course gene expression data. It can be applied to both long and short time series and also allows for multiple differential expression patterns. The method can identify a gene's temporal differential expression pattern as well as estimate the expression trajectory. The utility of the method is illustrated on both simulations and an experiment concerning the relationship between longevity and the ability to resist oxidative stress.

email: myuan@isye.gatech.edu

NORMALIZATION OF MICROARRAYS IN TRANSCRIPTION INHIBITION

Yan Zheng*, Cavan Reilly, University of Minnesota

Almost all of the existing methods for normalization assume that not too many of the genes differ in expression levels across arrays. Hence when the level of expression for many genes is not roughly constant across arrays, these standard methods are inappropriate. Here we develop a model for normalization in the context of an experiment that attempts to measure mRNA halflifes by stopping transcription and then measuring gene expression at certain later times. This model does not assume that most genes are constant across arrays, but rather assumes some genes have long halflifes. By supposing there are genes with long halflifes relative to the duration of the experiment, the model allows estimation of normalizing terms. Certain weaknesses of the basic model are noted, and a more sophisticated model is developed that addresses these shortcomings.

email: yanzheng@biostat.umn.edu

COMPARISON OF VARIOUS STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIAL GENE EXPRESSION IN REPLICATED MICROARRAY DATA

Seo Young Kim, Chonnam National University
Jae Won Lee*, Korea University
In Suk Sohn, Korea University

DNA microarray is a new tool in biotechnology, which allows the simultaneous monitoring of thousands of gene expression in cells. The goal of differential gene expression analysis is to identify those genes whose expression level changes significantly by the experimental conditions. Although various statistical methods have been suggested to confirm the differential gene expression, only a few studies which compared the performance of the statistical tests are performed. In our study, we extensively compared three types of parametric methods such as T-test, B-statistic and Bayes T-test, and three types of non-parametric methods such as samroc, SAM (Significance Analysis of Microarray) and a modified mixture model using both the simulated datasets and the three real microarray experiments.

email: jael@korea.ac.kr

APPLICATIONS OF RELIABILITY COEFFICIENTS IN CDNA MICROARRAY DATA ANALYSIS

Wenqing He*, University of Western Ontario
Shelley B. Bull, Samuel Lunenfeld Research Institute and University of Toronto

Gene expression microarray technology has been widely used in areas such as human cancer research to identify molecular characteristics of sample specimens. The microarray study, however, is a very complicated procedure that involves numerous sources of variability that may be either systematic or random. Systematic variation is often eliminated through normalization procedures. At present there are no standard criteria available to evaluate the performance of a particular normalization approach. We propose a reliability-type coefficient as a criterion to assess the effectiveness of normalization procedures in eliminating systematic variation. Informative gene screening is an important issue for analysis such as clustering procedures. We propose to apply a gene-specific reliability coefficient for informative gene screening. Simulation studies are conducted to evaluate the performance of the proposed methods and examples from ongoing microarray studies are employed for illustration.

email: whe@stats.uwo.ca

IMPROVED PARAMETER ESTIMATION FOR RMA IN BACKGROUND CORRECTION OF GENE EXPRESSION MICROARRAYS

Monnie McGee*, Southern Methodist University
Zhongxue Chen, Southern Methodist University

Affymetrix gene expression microarrays (or 'gene chips') allow the response of thousands of genes to various stimulants to be assessed all at once. Before one can analyze microarray data, the data must be corrected for non-biological sources of variation. Then, the data are normalized and summarized in order to obtain gene expression values. A popular way perform background correction is by the use of a convolution model, as in the Robust Multichip Average (RMA) algorithm. The convolution model states that the true signal, S, is a convolution of a normal noise distribution and an exponential signal distribution. In order to use the convolution model, the parameters of the normal and exponential distributions must be estimated. I show simulation results indicating that parameter estimates used in the current implementation of RMA are quite poor. I propose new parameter estimates, and show that the algorithm with the new parameter estimates gives better results than the current implementation of RMA.

email: mmcgee@smu.edu

IDENTIFYING FUNCTIONAL GENE CATEGORIES IN MICROARRAY EXPERIMENTS WITH NONPARAMETRIC METHODS

Hua Liu*, University of Kentucky
Christopher P. Saunders, University of Kentucky
Constance L. Wood, University of Kentucky
Arnold J. Stromberg, University of Kentucky

High-throughput technologies such as microarray chips allow investigators to study tens of thousands of genes at the same time. After assessing the statistical significance of differential gene expression, the identification of over-represented functional gene categories is commonly of interest. Existing statistical methods for identifying over-represented functional gene categories include Fisher exact tests, two-sample Kolmogorov-Smirnov tests, and Wilcoxon Mann-Whitney tests. These tests are not sensitive to differences in distributions over specific subsets of the real line. To improve the power, we propose a truncated Wilcoxon Mann-Whitney statistic that evaluates the partial area under the receiver operating characteristic curve and a one-sided version of this statistics that evaluates the partial area under the curve when the receiver operating characteristic curve is above the uniform cumulative distribution function. The asymptotic properties of these two statistics are derived. Both methods were applied to a microarray experiment, and the results compared with the Fisher exact test, the two-sample Kolmogorov-Smirnov test and the Wilcoxon Mann-Whitney test.

email: hualiu@ms.uky.edu

77. RECENT ADVANCES IN STATISTICAL METHODS FOR GENETIC EPIDEMIOLOGY

ROBUST ESTIMATION OF HAPLOTYPE/ENVIRONMENT INTERACTIONS

Andrew S. Allen*, Duke University
Glen A. Satten, Centers for Disease Control and Prevention

Case-control genetic association studies are popular designs for examining the genetic influences of complex disease. Investigating the effect of haplotypes is often a goal of such studies as haplotype-based analyses may provide information regarding the function of a gene and may lead to greater power relative to single loci methods. Haplotypes, however, are rarely observed directly but instead need to be reconstructed from multilocus genotype data. Likelihood methods addressing this problem are forced to model the nuisance distribution of haplotypes and can, if this distribution is misspecified, lead to substantial bias in parameter estimates even when complete genotype data is available. To address this problem, we use a geometric approach to estimation in the presence of nuisance parameters and develop locally-efficient estimators of the effect of haplotypes on disease that are robust to incorrect estimates of haplotype frequencies. Relations between existing estimators are discussed and the methods are demonstrated with a simulation study of a case-control design.

email: andrew.s.allen@duke.edu

ASSOCIATION TESTING WITH RELATED INDIVIDUALS

Mary Sara McPeek*, University of Chicago

A fundamental problem of interest to human geneticists is to understand the genetic risk factors that predispose some people to get a particular complex disease. We consider a study design for association testing in which cases are sampled from families with multiple affected members. Controls may be related to each other and to affecteds or unrelated. This type of design arises naturally when association testing follows linkage analysis, for which multiplex families may have been collected. The relatedness of the the individuals in the study has an impact in two ways: 1) the dependence among individuals due to their relationship can be used to weight the individuals to improve power (and in any case, must be taken into account to control type I error) and 2) a trait-predisposing allele or genotype may be enriched in individuals with multiple affected relatives, and one can use this information to improve power. We propose a new, computationally efficient method for case-control association studies of binary traits suitable for any set of related individuals, provided that their genealogy is known. This is joint work with Timothy Thornton, Xiaodong Wu, and Catherine Bourgain.

email: mcpeek@galton.uchicago.edu

TWO-PHASE DESIGNS IN STUDIES OF GENE-ENVIRONMENT INTERACTION

Nilanjan Chatterjee*, National Cancer Institute

In the “indirect” approach for fine mapping of disease genes , the association between the disease and a genomic region is studied using a set of marker SNPs that are likely to be in linkage disequilibrium with the underlying causal loci/haplotypes, if any exist. The SNPs themselves may not be causal. In this study, we propose novel strategies for testing associations in marker-based genetic association studies incorporating gene-gene and gene-environment interactions, two sources of heterogeneity expected to be present for complex diseases like cancers. We propose a parsimonious approach to modeling interactions by exploiting the fact that each individual marker within a gene is unlikely to have its own distinct biologic functions, but, instead, the markers are likely to be “surrogates” for a common underlying “biologic phenotype” for the gene, which, by itself, or by interacting with other genetic or/and environmental products, causes the disease. We use this approach to develop powerful tests of association in terms of observable genetic markers, assuming that the biologic phenotypes themselves are latent (not directly observable). We illustrate applications of the proposed methodology using both real and simulated data.

email: chattern@mail.nih.gov

78. ADAPTIVE BAYESIAN MODELING OF FUNCTIONAL BIOMEDICAL DATA

ADAPTIVE BAYESIAN SMOOTHING OF FUNCTIONAL PREDICTORS IN LINEAR MIXED MODELS USING WAVELET SHRINKAGE

Elizabeth J. Malloy*, Harvard School of Public Health
Brent A. Coull, Harvard School of Public Health
Jeffrey S. Morris, M.D. Anderson Cancer Center
Sara D. Dubowsky, Harvard School of Public Health
Helen H. Suh, Harvard School of Public Health

Frequently data is measured over time on a grid of discrete values which collectively define a functional observation. We develop a method that incorporates repeated measures of fixed effects functional data into a linear mixed model. Of particular interest is the situation where the functional fixed effects exhibit local features, such as spikes. While standard functional data analytic methods tend to smooth these features away, our method uses Bayesian wavelet shrinkage to preserve any dominant local features. The model is fit in the wavelet space by applying a discrete wavelet transform to the functional fixed effects. Estimation and inference are performed using a Bayesian paradigm. Properties of the method are examined using simulation studies. We apply the method to data from a study examining the association between hour air pollution measurements and acute systemic infammation.

email: emalloy@hsph.harvard.edu

BAYESIAN ADAPTIVE REGRESSION SPLINES FOR HIERARCHICAL DATA

Jamie L. Bigelow*, University of North Carolina at Chapel Hill & National Institute of Environmental Health Sciences
David B. Dunson, National Institute of Environmental Health Sciences

Motivated by the problem of describing and classifying hormone trajectories, we propose a flexible semiparametric Bayesian methodology for hierarchical functional data. The approach is based on a hierarchical spline model, with the number and location of knots and the distribution of the random spline coefficients treated as unknown. Assuming a parametric distribution for the spline coefficients, we obtain a very flexible regression model. Relaxing distributional assumptions and assuming a nonparametric discrete distribution for the spline coefficients, we obtain a procedure that clusters trajectories into classes, without pre-specification of the class-specific trajectories, the number of classes, or the allocation of subjects to classes. This is accomplished through a generalization of the Dirichlet process to a collection of unknown distributions having varying dimension. An efficient reversible jump Markov chain Monte Carlo algorithm is developed by constructing dependency within this collection of distributions. The methods are illustrated using progesterone data from the North Carolina Early Pregnancy Study.

email: bigelow@niehs.nih.gov

BAYESIAN HIERARCHICAL SPATIALLY CORRELATED FUNCTIONAL DATA ANALYSIS WITH APPLICATION TO COLON CARCINOGENESIS

Veera Baladandayuthapani*, The University of Texas M.D. Anderson Cancer Center
Bani K. Mallick, Texas A&M University
Mee Young Hong, Texas A&M University
Raymond J. Carroll, Texas A&M University

In this talk we present new methods to analyze data from an experiment using rodent models to investigate the biological mechanisms surrounding p27, an important biomarker predictive of early colon carcinogenesis. The responses modeled here are essentially functions nested within a two-stage hierarchy. Standard functional data analysis literature focusses on a single stage of hierarchy and conditionally independent functions with near white noise. In contrast, our functions are not necessarily conditionally independent and thus require new methodology. In fact, there is substantial biological motivation for existence of spatial correlation between the functions, which arise from biological structures called colonic crypts, a phenomenon we term crypt signalling. This talk focusses on modeling the spatial correlation between functions in the presence of a natural hierarchy. Our approach is fully Bayesian and uses Markov Chain Monte Carlo methods for inference and estimation. Analysis of this dataset gives new insights into the structure of p27 expression in early colon carcinogenesis and suggests the existence of significant crypt signalling.

email: veera@mdanderson.org

WAVELET-BASED FUNCTIONAL MIXED MODELS

Jeffrey S. Morris*, University of Texas M.D. Anderson Cancer Center
Raymond J. Carroll, Texas A&M University

An ever-increasing number of scientific studies yield functional data, in which the ideal units of observation are curves and the observed data consist of sets of curves sampled on a fine grid. In this talk, we discuss new methodology that generalizes the linear mixed model to the functional mixed model framework, and fits the model using a Bayesian wavelet-based approach. This method is very flexible, with the full range of fixed effects structures and between-curve covariance structures available in the mixed models framework. It yields nonparametric estimates of the fixed and random effects functions that are adaptively regularized using a nonlinear shrinkage prior on the fixed effects' wavelet coefficients, as well as estimates of any between-curve and within-curve correlation surfaces. Because we have posterior samples for all model quantities, we can also perform many types of Bayesian inference and prediction. This method is appropriate for functional data characterized by numerous local features like peaks, since our adaptive regularization procedure regularizes the functions with minimal attenuation of dominant local features. We show the results of applying this method to functional data from a study of children's activity levels using accelerometers.

email: jeffmo@mdanderson.org

79. CURRENT TRENDS IN SMALL AREA ESTIMATION

BAYESIAN METHODOLOGY WHICH ACCOUNTS FOR UNCERTAINTY ABOUT THE COMMONALITY OF A SET OF SMALL AREA ESTIMATES

Guofen Yan*, University of Virginia
J. Sedransk, Case Western Reserve University

We describe and evaluate Bayesian methodology to improve inference for 'small areas.' Inference for each small area will be improved by pooling data from other, like, entities. However, the pooled data must be concordant with that from the small area of direct interest. Our methodology ensures this concordance while the methods in current use may not. We show this using a set of samples.

email: gy4g@virginia.edu

USING SMALL AREA ESTIMATION METHOD TO COMBINE TWO HEALTH SURVEYS

Shijie Chen*, RTI International

To collect information on health status and activities of daily living (ADL), the Center of Medicare and Medicaid Services (CMS) conducted a Medicare Health Survey (MHS). In order to get sufficient sample with limited budget the MHS questionnaire was attached to the on-going Consumer Assessment of Health Plans Survey (CAHPS), which is much larger than the MHS. However, these two surveys gave different results in some interested domains in that the CAHPS estimates have smaller sampling errors with larger sampling biases and the MHS estimates have larger sampling errors with much smaller sampling biases. This situation makes it difficult to combine these two surveys directly. To improve the quality of analysis, we propose using small area estimation (SAE) techniques when combining these surveys. We will develop a proper SAE model reflecting this situation.

email: schen@rti.org

A SMALL AREA ESTIMATION APPROACH IN REDUCING SURVEY COSTS

Partha Lahiri*, University of Maryland
Paul D. Williams, National Center for Health Statistics

The demand for producing various statistics at sub-national levels is steadily increasing at a time when survey agencies are looking for ways to reduce total survey costs to meet the fixed budgetary requirements. In addition to the usual sampling errors, survey researchers are also paying more attention to various kinds of non-sampling errors resulting in higher survey costs. In many instances, estimates that use external data sources (with or without survey data) have proven to be more reliable than the customary design-based survey statistics and these small area statistics, when aggregated, do not produce the usual survey weighted statistics at the national level. In view of all these changes in the complex survey environment, agencies such as the U.S. National Center for Health Statistics and the Bureau of Labor Statistics are exploring various innovative approaches to cut survey costs. The main idea of this paper is to demonstrate how a change of the sampling design and the use of estimation techniques found in small area estimation literature can potentially help in reducing total errors of a survey statistic for a large area.

email: plahiri@survey.umd.edu

80. INFERENCE IN THE PRESENCE OF NON-IDENTIFIABILITY: APPLICATIONS TO THE ANALYSIS OF COARSE DATA

BAYESIAN INFERENCE IN THE PRESENCE OF NON-IDENTIFIABILITY

Paul Gustafson*, University of British Columbia

Realistic modeling of observational data in the health sciences often leads to a nonidentified model. For instance, this might arise from acknowledging imperfections such as exposure misclassification or selection bias, along with uncertainty about the extent of the imperfections. One inferential strategy is to simply carry out Bayesian analysis with the best prior information that can be ascertained, as a lack of identifiability does not preclude the application of Bayes theorem. This talk compares the Bayesian approach to other strategies, such as making overly-strong assumptions in order to obtain an identified model, or carrying out a sensitivity analysis. These comparisons are facilitated by a simple characterization of how much Bayesian learning can occur for a given parameter in a given nonidentified model.

email: gustaf@stat.ubc.ca

DRAWING INFERENCE FROM REGIONS OF ESTIMATES: IGNORING BOUNDS OR BOUNDING IGNORANCE?

Stijn Vansteelandt*, Ghent University-Belgium
Els Goetghebeur, Ghent University-Belgium
Mike Kenward, London School of Hygiene and Tropical Medicine-U.K.
Geert Molenberghs, Hasselt University-Belgium

Analyses of incomplete data are problematic as parameters describing the target population are typically not identified without untestable assumptions. To make sense of the incomplete data while avoiding misguided conclusions due to incorrect assumptions, sensitivity analyses have replaced the classical point estimate by a region of parameter estimates. Molenberghs, Kenward and Goetghebeur (2001) call such region an estimated ignorance region because it expresses ignorance due to the missing data. The idea of worst case-best case regions is not new, but some major open challenges remain. First, special algorithms are needed to allow for computationally feasible estimation of ignorance regions. Second, the region itself is subject to sampling variation and this needs to be acknowledged. Third, incorporating expert opinions is not straightforward. In this talk, we address these practically important questions and illustrate them with some solutions in a number of practical applications. Our development will lead to a formalism for sensitivity analysis which naturally allows to combine model uncertainty or 'ignorance', and sampling uncertainty or 'imprecision' into overall Estimated Uncertainty RegiOns (EUROs).

email: stijn.vansteelandt@ugent.be

A DISTRIBUTIONAL APPROACH FOR SENSITIVITY ANALYSIS IN OBSERVATIONAL STUDIES

Zhiqiang Tan*, Johns Hopkins University

The approach is developed from a nonparametric likelihood perspective. Regardless of the no-confounding assumption, the marginal distribution of covariates times the conditional distribution of observed outcome given each treatment assignment and covariates is estimated. For a fixed bound on unmeasured confounding, the marginal distribution of covariates times the conditional distribution of counterfactual outcome given each treatment assignment and covariates is explored to the extreme and then compared with the composite distribution corresponding to observed outcome given the same treatment assignment and covariates. We illustrate the methods by analyzing the data from an observational study on right heart catheterization.

email: ztan@jhsph.edu

81. IMS: BAYESIAN MODEL SELECTION

MCMC WITH MIXTURES OF SINGULAR DISTRIBUTIONS: APPLICATION TO BAYESIAN MODEL SELECTION

Adrian E. Raftery*, University of Washington
Raphael Gottardo, University of British Columbia

Markov chain Monte Carlo (MCMC) methods for Bayesian computation are mostly used when the dominating measure is the Lebesgue measure, the counting measure or a product of these. Many Bayesian problems give rise to distributions that are not dominated by the Lebesgue measure or the counting measure alone. In this paper, we introduce a simple framework for using MCMC algorithms in Bayesian computation with mixtures of mutually singular distributions. The idea is to find a common dominating measure that allows the use of traditional Metropolis-Hastings algorithms. We show how our formulation can be used in Bayesian model selection. When the full conditionals are available, the Gibbs sampler can be used. We compare our formulation with the reversible jump approach, and show that the two are closely related. We show how the method can be used for testing a normal mean, for variable selection in regression, and for hypothesis testing for differential gene expression under multiple conditions. This allows us to compare the three methods considered: Metropolis-Hastings with singular distributions, Gibbs sampler with singular distributions, and reversible jump. In our examples, we found the Gibbs sampler with singular distributions to be more precise and to need considerably less computer time than the other methods.

email: raftery@stat.washington.edu

USING BAYESIAN MODEL AVERAGING TO ASSESS THE EFFECT OF SOCIAL INTERACTIONS ON RECIDIVISM

Sibel Sirakaya*, Departments of Economics and Statistics, and Center for Statistics and the Social Sciences, University of Washington

Using a national sample, this paper identifies the risk factors for recidivism among Female, Male, Black, White and Hispanic felony probationers. The individual hazard function is assumed to depend on individual and neighborhood characteristics as well as social interactions among probationers. In selecting the covariates from a set of potential candidates, Bayesian model averaging is used both to account for model uncertainty and the subsequent inference. The results point to social interactions as one of the most significant factors affecting recidivism among all gender, ethnicity and race groups. When a frailty parameter is added to account for the possibility of unobserved risk factors shared by probationers within neighborhoods, the empirical results remain robust indicating negligible unobserved neighborhood-level heterogeneity.

email: sirakaya@stat.washington.edu

BAYESIAN MODEL SELECTION FOR DISCRIMINANT ANALYSIS

Russell J. Steele*, McGill University
Michelle E. Ross, McGill University

Discriminant analysis is a classical tool for used for prediction of classification according to a set of observed covariates. It has become popular again with the wide availability of model-based clustering tools, but model selection with the BIC and AIC can be difficult to formulate because of the underlying assumptions of the model. We reformulate the classical model in such a way that model selection via penalized likelihood methods (such as the BIC) is possible. We show that the new BIC criterion roughly approximates a set of integrated likelihoods for standard prior distributions for the mean, variance, and proportion parameters. The results are presented for a small-sample biomarker assay problem where immunohistochemistry is used to predict colo-rectal cancer treatment efficacy.

email: steele@math.mcgill.ca

82. RE-SAMPLING AND ROBUST METHODS AND APPLICATIONS

GENERAL OUTLIER DETECTION FOR A HOMOGENEOUS POISSON PROCESS WITH SUM-QUOTA ACCRUAL SCHEME

Jonathan T. Quiton*, University of South Carolina
Edsel A. Peña, University of South Carolina
James D. Lynch, University of South Carolina

In this talk we consider the problem of detecting whether a specific observed inter-event time in a recurrent event data, which is subject to a sum-quota accrual scheme, is an outlier. In addition, we also address the problem of determining if a particular subject or unit is an outlier in relation to the other observed units in light of the recurrent event data. We limit this talk to a situation where the stochastic process governing event occurrences for each subject or unit is a homogeneous Poisson process (HPP), and the mathematical framework utilized to develop the outlier detection procedures is similar to that of Neyman's smooth embedding. Through this framework, we are able to derive jackknife-based procedures. We discuss several conditioning schemes for the sampling distributions of the test statistics, in particular, we discuss the relevance of the Conditionality Principle in this problem. Results of simulation studies regarding power comparisons of the different conditioning schemes will be presented, and the procedures will be illustrated by applying to real data sets in biomedical and engineering settings. Key words: Conditionality Principle; outlier detection; homogeneous Poisson process; recurrent events with sum-quota accrual; jackknife residuals.

email: quiton@stat.sc.edu

DETERMINING OPTIMAL EXPERIMENTAL DESIGNS FOR NONLINEAR MODELS USING LIKELIHOOD RATIO BASED INFERENCE

Sharon D. Yeatts*, Virginia Commonwealth University
Chris Gennings, Virginia Commonwealth University

For nonlinear models, Wald-type inference is based on the linear Taylor series approximation to the nonlinear function. If the approximation is inadequate, the resulting inference can be misleading. For this reason, it is often preferable to apply likelihood ratio-based inference techniques to nonlinear models. Designing an experiment for likelihood ratio-based inference is problematic. Some common alphabetic optimality criteria are based on the asymptotic variance-covariance matrix. A criterion based on the likelihood ratio-based confidence interval, however, requires advance knowledge of the response at each of the candidate design points. We have developed a method for constructing the optimal second-stage design for likelihood ratio-based inference. We want to augment the original experiment in such a way as to increase the precision associated with the parameters of interest. Our method involves the simulation of data points through the random resampling of standardized residuals from the first-stage design. The estimate of the confidence region for each candidate design, however, depends on the residuals sampled. Therefore, the optimal design search is based on a bootstrap estimate of the confidence region for each candidate design. The method and resulting second-stage designs are illustrated for a nonlinear dose-threshold model.
Supported by T32 ES07334-01A1 (NIEHS, NIH).

email: the.yeatts@netzero.com

COMPONENTS OF THE BOOTSTRAP-VARIANCE OF THE AREA UNDER THE ROC CURVE

Andriy I. Bandos*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh
David Gur, University of Pittsburgh

The area under the Receiver Operating Characteristic curve (AUC) is a summary index that is widely used in the evaluation of diagnostic tests. This index and subsequent inferences can be affected by a variety of factors including between-subjects and between-readers variability. Thus, knowledge about these sources of variability in the AUC is important for various stages of study design and data analysis. In this paper we extend to the multi-reader setting a general approach we previously employed for constructing a closed-form solution for the ideal bootstrap variance of the nonparametric estimator of AUC difference. We then derive formulae for the variance-components within a bootstrap space obtained by bootstrapping hierarchical data based on independent sets of subjects and a sample of readers. These solutions enable computing of the ideal bootstrap-variance-components directly instead of estimating them by sampling the bootstrap sample space (Monte Carlo estimators), and hence, eliminate the sampling error and burdensome computations associated with repeated sampling. Also, the closed-form expressions permit the analytical assessment of the bias of some of the bootstrap variance components. We derive the expectation of the reader-related component of the ideal bootstrap-variance and illustrate an approach for constructing a potentially less biased estimator.

email: anb61@pitt.edu

MARGINAL ANALYSIS OF CORRELATED FAILURE TIME DATA WITH INFORMATIVE CLUSTER SIZES

Xiuyu Cong*, Rice University
Guosheng Yin, University of Texas-M. D. Anderson Cancer Center
Yu Shen, University of Texas-M. D. Anderson Cancer Center

We consider modeling correlated survival data when cluster sizes may be informative to the outcome of interest based on a within-cluster resampling approach and a weighted marginal model. We derive the large sample properties for the within-cluster resampling estimators under the Cox proportional hazards model. We establish consistency and asymptotic normality of the regression coefficient estimators, and the weak convergence property of the estimated baseline cumulative hazard function. The weighted marginal model is constructed by incorporating the inverse of cluster sizes as weights into the estimating equations. We conduct extensive simulation studies to assess and compare the finite-sample behaviors of the estimators and apply the proposed methods to a real data example as an illustration.

email: xcong@rice.edu

PERMUTATION-BASED TEST FOR IDENTIFYING LONGITUDINAL GENE EXPRESSIONS ASSOCIATED WITH THE TIME TO AN EVENT

Natasa Rajicic*, Harvard School of Public Health
Dianne M. Finkelstein, MGH Biostatistics and Harvard School of Public Health
David A. Schoenfeld, MGH Biostatistics and Harvard School of Public Health

Development of methods for linking gene expressions to various clinical and phenotypic characteristics is an active area of genomic research. Scientists hope that such analysis may, for example, describe relationships between gene function and clinical events such as death or recovery. Methods are available for relating gene expression to measurements that are categorized or continuous, and there is less work in relating expressions to survival-type endpoints such as time to death, response, or relapse. When gene expressions are measured over time, methods have been proposed for differentiating temporal patterns. However, no methods so far have been proposed for the survival analysis of longitudinally collected genearrays. We describe an approach to the survival analysis of longitudinal gene expression data. We construct a measure of association between time to a survival endpoint and gene expressions collected over time using the approach of a conditional score. The issue of high dimensionality and dependence when assessing statistical significance is addressed using permutations and the control of false discovery. Our proposed method is illustrated on a dataset from a multi-center research study of inflammation and response to injury that aims to uncover the biological reasons why patients can have dramatically different outcomes after suffering a traumatic injury.

email: nrajicic@partners.org

ESTIMATING THE LOCATION FROM A SKEWED SAMPLE: TO TRANSFORM OR NOT TO TRANSFORM

Abutaher M. Minhajuddin*, University of Texas-Southwestern Medical Center
Xian-Jin Xie, University of Texas-Southwestern Medical Center

Often the real data provide evidence of skewness of the population. In estimating the location from such data, the usual practice is to transform the data using a transformation such as the Box-Cox transformation. However, transforming the data gives rise to a different set of problems with no clear solution to some of them. In this article, we want to discuss some of these problems associated with transforming the data. We also want to present some robust procedures based on bootstrap methods. We will show via simulation results that these alternative procedures do provide viable alternatives to transformation of the data.

email: minhajuddin@gmail.com

A NOVEL STATISTICAL APPROACH IDENTIFYING AND LIMITING THE EFFECT OF INFLUENTIAL OBSERVATIONS

Tamekia L. Jones*, University of Alabama at Birmingham
David T. Redden, University of Alabama at Birmingham

Outliers are observations with extreme standardized deviations between the observed dependent variable and the predicted dependent variable. Within linear regression, outliers are detected using studentized residuals. Leverage is a measure of the standardized deviation of an observation's independent variables from the center of the independent variables. Within linear regression, leverage is assessed using the diagonal of the projection matrix H=X(X`X)-1X`. An observation that is both an outlier and a leverage point is an influential observation. Influential observations affect estimation of regression parameters leading to poor estimation of the regression line, incorrect inference, and inaccurate predictions. Detection of these observations is challenging due to the masking effect. This effect occurs when some or all influential points are difficult to identify using regression diagnostics because the extremeness of one observation obscures the extremeness of another. We present a robust regression method that extends Rousseeuw's least trimmed squares method using the minimum covariance determinant. Results utilizing ordinary least squares, the least trimmed squares method, and our proposed method are illustrated and compared using simulations. We illustrate that our proposed approach overcomes the masking effect, properly identifies influential observations (outliers and leverage points), and is robust to their influence.

email: tljf81@uab.edu

AN EVALUATION OF PSEUDO OBSERVATIONS IN MULTI STATE MODELS

Pinaki Biswas, Department of Biostatistics, University of Michigan, Ann Arbor
Jack D. Kalbfleisch, Department of Biostatistics, University of Michigan, Ann Arbor

Multi-state models have long been in use for modeling life history data. It is straightforward to model the transition intensities using a Cox or relative risk type model, but this approach leads to complicated relationships between the covariates and the occupancy probabilities for the various states. Andersen et al. (2003, Biometrika) suggest approximate generalized linear models for pseudo observations generated from estimated occupancy probabilities and proposed GEE methods for model fitting. We examine this suggestion and compare it with bootstrap methods for estimating parameters in the approximate GLMs. We find that the bootstrap tends to perform better, particularly in the presence of censoring where the pseudo observation technique encounters difficulty.

83. PHARMACOKINETICS, PHARMACODYNAMICS, AND TOXICOLOGY

THRESHOLD DOSE-RESPONSE MODEL WITH RANDOM EFFECTS FOR TERATOLOGICAL DATA

Daniel L. Hunt*, St. Jude Children's Research Hospital
Shesh N. Rai, St. Jude Children's Research Hospital

In developmental toxicity experiments, animals are randomly allocated to exposure levels of some environmentally ambient toxic substance and, subsequently, responses are measured for the litters. There are many issues in the planning and analysis of such studies: selection of doses, allocation of animals at different doses, correlated data, threshold effect, hormesis effect, estimation and testing problems. The litter effect due to natural clustering of the data is well-known. Additionally, for many dose-response studies of toxins acting on animal development, a threshold effect has been observed whereby responses for certain dose groups (below threshold) are equivalent. We incorporate a random component (for the litter effect) into a general threshold dose-response model to characterize the dose-response relationship. Of primary interest is determining threshold significance. Hence, appropriate characterization of all other effects is vital to successful threshold estimation. We also consider a more flexible spline-based approach for estimation and compare the two models. We fit these models to a well-known data set in the field of teratology. Key Words: Beta-binomial; Dose-response; Random effects; Regression Spline; Threshold; Teratological experiment.

email: daniel.hunt@stjude.org

INTERVAL ESTIMATION OF EFFECTIVE DOSES IN TOBIT REGRESSION MODEL

Nan Lin*, Washington University in St. Louis
Douglas Simpson, University if Illinois at Urbana-Champaign
Ringo M. Ho, Nanyang Technology University

In this article, we propose various interval estimation techniques of the effective doses Ed100q in tobit regression models. These methods are developed based on bootstrap methods and their adjustment, the delta method, the Fieller interval and the likelihood ratio method. For these asymptotically equivalent methods, we conducted extensive simulations to study their finite sample performance, in particular, their robustness to misspefication of error distributions. The performance of these nine methods are affected by the sample size, the value of q, and the type of misspecification. The delta method has displayed relatively stable performance and therefore recommended when one does not have much knowledge about the error distribution.

email: nlin@math.wustl.edu

STOCHASTIC MODELS FOR COMPLIANCE ANALYSIS USING INTER-DOSING TIMES

Junfeng Sun*, University of Nebraska Medical Center
Haikady N. Nagaraja, The Ohio State University

Compliance is the extent to which a patient follows the prescribed regimen. Good compliance is crucial for maintaining drug concentration in patients' body, thus very important in both clinical trials and medical practice. Even though many different compliance indices have been proposed in the literature, there is no published systematic study of the statistical properties of these compliance indices. We utilize the information-rich electronic event monitoring (EEM) data, build realistic stochastic models to describe the inter-dosing times, and study the statistical properties of the following clinically meaningful compliance indices: the therapeutic coverage, the delayed medication index and the premature medication index. We apply Markov-dependent mixture models to describe the inter-dosing times of each subject, use empirical Bayes approach to pool data subjects, and combine the pharmacokinetic (PK) model of the drug with the inter-dosing times. We establish asymptotic normality of these indices and analyze the data from an AIDS clinical trial as an illustration.

email: junfengsun@unmc.edu

NONPARAMETRIC BAYES TESTING OF CHANGES IN A RESPONSE DISTRIBUTION WITH AN ORDINAL PREDICTOR

Michael L. Pennell*, University of North Carolina at Chapel Hill, NIEHS
David B. Dunson, NIEHS

In certain biomedical studies, one may anticipate changes in the shape of a response distribution across the levels of an ordinal predictor. For instance, in toxicology studies, a likely trend is that the skewness of non-Gaussian outcomes, such as hematology or clinical chemistry data, increases as the dose of the exposure increases. To address this issue, we propose a Bayesian nonparametric method for testing for distribution changes across an ordinal predictor. Using a dynamic mixture of Dirichlet processes, we allow the response distribution to change flexibly at each level of the predictor. In addition, by assigning mixture priors to the hyperparameters, we can obtain posterior probabilities of no effect of the predictor and identify the lowest level for which there is an appreciable change in distribution. The method also provides a natural framework for performing tests across multiple outcomes. We will demonstrate our method using data from a toxicology experiment and, in doing so, will discuss how historical control data may be used in prior elicitation.

email: pennell@niehs.nih.gov

A PHYSIOLOGICALLY BASED PHARMACOKINETIC MODEL FOR GAVAGE AND IV ADMINISTRATION OF METHYLEUGENOL IN F344/N RATS AND B6C3F1 MICE

Petra K. LeBeau*, Rho, Inc.
Shree Y. Whitaker
Christopher J. Portier, NIEHS

The current presentation will discuss the development and implementation of a diffusion-limited PBPK model that was developed to quantitatively describe the process involved in methyleugenol toxicokinetics. The PBPK model mathematically represents the absorption, distribution, metabolism, and elimination of methyleugenol in rats and mice. The differential equations used in this model describe mass transfer of methyleugenol between tissue compartments and have parameters representing physiological quantities and chemical-specific parameters. Likelihood-based statistical methods were used to test hypotheses regarding the structure of the model and similarities across species and/or sexes. The model demonstrates that absorption of methyleugenol in rats and mice is rapid and complete; a small but significant fraction of the methyleugenol is absorbed and carried through the lymph and blood system by chylomicrons; the distribution of methyleugenol to tissues is not hindered by capillary permeability; metabolism of methyleugenol follows linear kinetics; and an extra-hepatic site for metabolism/elimination is needed for both species. The model also serves as a tool to evaluate the potential for an increase in human cancers using observed serum concentrations in the U.S. general population.

email: plebeau@rhoworld.com

A REGRESSION BASED APPROACH FOR DEVELOPING A LIMITED SAMPLE MODEL FOR PHARMACOKINETIC DATA

Alfred F. Furth*, Mayo Clinic
Sumithra J. Mandrekar, Mayo Clinic
Andrea Rau, Mayo Clinic
Joel M. Reid, Mayo Clinic
Angelina Tan, Mayo Clinic
Sara J. Felten, Mayo Clinic
Charles Erlichman, Mayo Clinic
Matthew M. Ames, Mayo Clinic
Alex A. Adjei, Mayo Clinic

Area under the drug concentration curve (AUC) is a measure of therapeutic activity of an agent. Our goal was to develop a simple predictive model for AUC using drug concentrations collected from a few time points. The primary challenge is in utilizing the multitude of data collected from a small cohort of patients for developing a clinically useful limited sample model (LSM). Blood samples were collected at 11 different time points, spanning 25 hours from 34 patients on a Mayo phase I clinical trial. Graphical methods and Pearson correlations were used for assessing functional forms and univariate relationships. Several Pharmacokinetic data modeling techniques such as dose-normalization were investigated. LSMs were developed using multivariate regression techniques. Sensitivity analysis, model diagnostics and prediction accuracy were used to evaluate the performance characteristics of the different LSMs. Internal validation was performed using bootstrap resampling techniques. Dose was not a significant predictor of AUC. Using log-transformed data, the two and three time point LSMs predicted 94.1% and 97.1% of values within 5% of the actual log-AUC. While the bootstrap validation approach confirmed multivariate results, it was less applicable due to the accelerated titration design used for this trial. Prospective validation of this approach is underway.

email: furth.alfred@mayo.edu

RELATIONSHIP ASSESSMENT BETWEEN PK AND PD

Tao Liu, University of Pennsylvania
Longlong Gao*, GlaxoSmithKline Company

Pharmacokinetic and Pharmacodynamic (PK/PD) study is aimed to investigate the process how a specific agent enters human body and how human body responds to it. Generally, several indicators of drug concentrations and response concentrations are measured during a certain period of time for a patient after the treatment administration. A quantity of clinical importance is the correlation between these indicators, which are usually measured as curves over time. This paper considers the methods that are suitable for estimating the PK/PD correlation. The methods include the standard method based on the landmarks (scalar) of PK/PD curves, the canonical correlation method (PK/PD as vectors), the functional canonical correlation method (PK/PD as curves, (Leurgans, Moyeed, and Silverman, 1993)), and the recently published 'function dynamical correlation' method (PK/PD as curves $\in L^2$, (Dubin and M\'{u}ller, 2005)). We apply the methods to assess the relationship between PK and PD in a clinical trial. The calculated correlations are compared, and their interpretations are discussed.

email: liutao_94@tsinghua.org.cn

84. SURVIVAL ANALYSIS II

NONPARAMETRIC ESTIMATION OF THE BIVARIATE SURVIVOR FUNCTION UNDER RIGHT TRUNCATION WITH APPLICATION TO PANIC DISORDER

Xiaodong Luo*, Columbia University
Wei-Yann Tsai, Columbia University

In studies, investigators are interested in estimating the distribution of the onset ages of a generic disorder in successive generations. The empirical distribution is inappropriate for this purpose due to right-truncation, i.e. only parent-child pairs with onset ages prior to the ages of interview were included in the sample. In this article, we propose a simple nonparametric method to the underlying bivariate distribution of the onset ages of parent-child pair. This estimator has a closed form and performs satisfactory in the simulation study. We also apply our approach to estimate Age-at-Onset anticipation of the panic disorder.

email: xl2013@columbia.edu

A PENALIZED LIKELIHOOD APPROACH TO JOINT MODELING OF LONGITUDINAL AND TIME-TO-EVENT DATA

Wen Ye*, University of Michigan
Xihong Lin, Harvard University
Jeremy M.G. Taylor, University of Michigan

Joint models for longitudinal and time-to-event data have attracted recent attention. Full joint likelihood approach using EM algorithm (Wolfsohn and Tsiatis, 1997) or Bayesian method of estimation (Faucett and Thomas, 1996) not only eliminate the bias in naïve and two-stage methods, but also improves the inference and efficiency. However, both an EM algorithm and a Bayesian method are computationally intensive, which limits the utilization of joint models. We propose to use an estimation procedure of maximizing a penalized joint likelihood generated by a Laplace approximation of a joint likelihood, which combines the likelihood of the longitudinal data and the partial likelihood of the time-to-event data. The results of a simulation study show that this penalized likelihood approach performs as well as the corresponding EM algorithm under a variety of scenarios, but only requires a fraction of the computation time. An additional advantage of this approach is that it does not require estimation of the baseline hazard function. The procedure is applied to a prostate cancer study.

email: wye@umich.edu

NEW METHODOLOGY: CHALLENGING THE LOGRANK AND WILCOXON TESTS FOR NONPARAMETRIC SURVIVAL COMPARISON

Gabriel P. Suciu*, Nova Southeastern University

Hypothesis tests of the equality of Kaplan-Meyer survival curves is typically accomplished using one of the available methods designed for this purpose but, without doubt, the logrank test is the one most commonly used. Perhaps the reason for the popularity of the logrank test rests in its ready availability in almost all statistical software packages. However, what many users do not appreciate is that the logrank test has very low power for some alternative hypotheses. Furthermore, the alternative hypothesis for which the log rank has good power may not be at all what the investigator has in mind. There are also assumptions that underlie the appropriate use of this test and these assumptions are often ignored (if not violated) and may lead to a test with very low power. A new methodology that unifies the impressive arsenal of survival comparison is presented. The new taxonomy and the guidelines conserve the power and the efficiency of the tests, contributing to a better treatment efficacy assessment in clinical trials.

email: suciu@nova.edu

A SAMPLE SIZE FORMULA FOR RECURRENT EVENTS DATA USING ROBUST LOG-RANK STATISTICS

Rui Song*, University of Wisconsin-Madison
Michael R. Kosorok, University of Wisconsin-Madison

Recurrent events data are frequently encountered in clinical trials. This article develops a sample size formula for robust log-rank statistics applied to recurrent events data with arbitrary numbers of events. The formula is derived based on the asymptotic normality of the robust log-rank statistics under certain local alternatives in the recurrent data context. It reduces to the same form as Schoenfeld's (1983) formula for cases of a single event or independence within subjects when the test size is small enough. We carry out simulations to study the control of type I error and the comparison of powers between several methods. The proposed sample size formula is illustrated using data from an rhDNase study and a bladder cancer study.

email: rsong@stat.wisc.edu

REGRESSION MODELS FOR THE MEAN OF THE QUALITY-OF-LIFE-ADJUSTED RESTRICTED SURVIVAL TIME USING PSEUDO-OBSERVATIONS

Adin-Cristian Andrei*, University of Michigan
Susan Murray, University of Michigan

In this research we develop generalized linear regression models for the mean of a quality-of-life-adjusted restricted survival time. Parameter and standard error estimates are obtained from generalized estimating equations (GEE) applied to pseudo-observations. Simulation studies with moderate sample sizes are conducted and an example from the International Breast Cancer Study Group Ludwig Trial V is used to illustrate the newly developed methodology.

email: andreia@umich.edu

85. TOPICS IN STATISTICS: SEQUENTIAL METHODS, GOODNESS-OF-FIT TESTS, AND MULTIVARIATE ANALYSIS

COMPARISON OF SEQUENTIAL EXPERIMENTS FOR ESTIMATING THE NUMBER OF CLASSES IN A POPULATION

Tapan K. Nayak*, George Washington University
Subrata Kundu, George Washington University

This talk will deal with stopping rules in sequential sampling from a population with an unknown number of classes or species. Adopting Blackwell's (1951) ideas we define "more informative stopping rule," which induces a partial ordering of all stopping rules. Some consequences of more informativeness, and certain complete class results will be discussed. Necessary and sufficient conditions for a stopping rule to yield complete sufficient statistics will be presented and explained geometrically. Unbiased estimation of the class size and some functions of it will also be discussed.

email: tapan@gwu.edu

AN EFFECTIVE MAXIMUM LIKELIHOOD ESTIMATION METHOD FOR A FINITE MIXTURE MODEL IN HIGH DIMENSIONAL BIOLOGY

Qinfang Xiang*, University of Missouri-Rolla
Gary L. Gadbury, University of Missouri-Rolla

Finite mixture models have found use in the analysis of high dimensional data such as result from microarray experiments. Maximum likelihood estimation has often been used to estimate the model parameters using numerical optimization techniques. However, finding maximum likelihood estimates (MLEs) representing a unique maximum can be challenging when the likelihood surface is flat, a situation observed to occur when mixture components are "close." Many methods can then be sensitive to starting values of the optimization routine, sometimes converging to saddlepoints and giving negative asymptotic variance estimates. This further complicates the ability to estimate the precision of MLEs and other derived quantities like estimates of false discovery. An effective approach based on number theoretic method is applied to a finite mixture of one uniform and one beta. This type of mixture has been used to model a distribution of P-values. Compared to Newton-type methods, this method does not require initial values and can efficiently find MLE's. Also, interval estimation for parameters of this mixture is considered by computing the Hessian matrix evaluated at the MLE's.

email: xiang@umr.edu

NEW TESTS OF UNIFORMITY AND NORMALITY

David B. Kim*, Manhattan College

This presentation introduces a new test of goodness of fit based on a quantum mechanical representation of a probability density function (pdf). We represent a pdf as a squared magnitude of the probability amplitude, or the wave function, which can be represented as a normalized linear combination of orthogonal functions. In contrast to Good and Gaskins (1971), who developed the penalized likelihood to estimate the coefficients of such a linear combination, we estimate the coefficients by minimizing an empirical distance measure in the space of square integrable functions, namely, the L2E criterion of Scott (2001). Test statistics for uniformity and normality are based on the L2E estimators, and their asymptotic distributions under the null hypothesis are easily obtained. Simulated power comparisons with existing tests show that the new tests are competitive.

email: david.kim@manhattan.edu

THE ASYMPTOTIC DISTRIBUTION OF MODIFIED SHAPIRO-WILK STATISTICS FOR TESTING MULTIVARIATE NORMALITY

Christopher P. Saunders*, University of Kentucky
Constance L. Wood, University of Kentucky

A class of tests for multivariate normality is proposed which are based on the Shapiro-Wilk statistic and the related correlation statistics applied to the dependent univariate data that arises with a data-suggested linear transformation, which projects the data vectors onto the real line. In particular, we consider a sequence of random linear transformations converging in probability to a fixed linear transformation. The asymptotic properties of the statistics in the proposed class are established. The statistics based on the random linear transformations are shown to be asymptotically equivalent to the statistics using the fixed linear transformation. The statistics based on the fixed linear transformation have the same critical points as the corresponding tests of univariate normality; this allows for an easy implementation of these tests for multivariate normality. In addition, an extensive simulation study is performed to characterize the small sample behavior of these statistics.

email: saunders@ms.uky.edu

ESTIMATING EQUATIONS FOR CANONICAL CORRELATIONS

Hye-Seung Lee*, Columbia University
Myunghee Cho Paik, Columbia University
Joseph H. Lee, Columbia University

Maximum canonical correlation is a useful tool to best summarize the relationship between two groups of variables. In this paper, we develop a regression approach to examine the association between canonical correlation and pair specific covariates (regression CCA). Through this approach, the regression parameters are estimated for the association between canonical correlation and pair specific covariates, allowing the portion of variables to be discrete as well as continuous. This regression CCA is applied to the familial correlation analysis for the memory scores from the study of familial Alzheimer's disease (AD) in Caribbean Hispanics.

email: hl660@columbia.edu

ARRAYBLAST: DATA-MINING TOOL FOR GENE EXPRESSION SIGNATURES

Yajun Yi, Chun Li*, Vanderbilt University
Alfred L. George, Vanderbilt University

Comparisons of microarray studies across platforms and between laboratories are difficult for a variety of reasons. However, these types of comparisons offer enormous opportunity to learn about shared patterns of gene expression among different experiments, tissues, species and experimental systems. We have developed and implemented ArrayBlast, an approach for systematically organizing and comparing microarray datasets from a variety of sources. Rather than attempting to compare raw or even normalized datasets directly, ArrayBlast was developed to compare expression signatures that are represented by genes exhibiting significant differences in expression along with their corresponding statistical measurements. This informatics tool, which is implemented in Perl, operates in a manner analogous to the widely used nucleotide alignment tool, BLAST. To enable this strategy, we first converted a large reference dataset of microarray experiments into a searchable and comparable resource. This includes defining gene expression signatures for each dataset using common statistical measures. Then, a query dataset is compared to every subject dataset in the collection and scored based upon their similarities of gene expression signatures. The final output of the program reports a summary of statistically significant matches at three levels: dataset, gene expression signatures and genes exhibiting significant differences in expression.

email: chun.li@vanderbilt.edu

RANDOMIZED DISCONTINUATION TRIALS: DESIGN AND EFFICIENCY

Tao Liu, University of Pennsylvania
Valeri V. Fedorov*, GlaxoSmithKline Co.

Randomized Discontinuation Trials (RDT) are two-phase designs and become more and more popular across a number of therapeutic areas (oncology is one of the most known (Rosner, Stadler and Ratain, 2002)). In this design, a single arm trial, called the open phase, is followed by a randomized blinded two-arm trial at the second phase to compare two treatments (generally one is placebo). Intuition and simulation exercises show that potentially RDT can increase the sensitivity of trials relatively to the more traditional patient allocations. This increase can be substantial if the open phase provides a reliable separation of responders from nonresponders. We compare RDT with the traditional two-arm randomized clinical trial (RCT) when the outcomes are binary and the population of interest consist of three groups, placebo responders, treatment-only responders and nonresponders. Our results are derived in the 'likelihood parameter estimation' setting and are based on the comparisons of estimator variances. We identify conditions under which RDT is superior to RCT, including the response rates, the misclassification rates and the randomization strategy at the second stage. Transition to hypothesis testing is rather straightforward.

email: daniel.d.liu@gsk.com

86. STATISTICAL ISSUES IN THE DESIGN, EVALUATION, AND MONITORING OF CLINICAL TRIALS WITH LONGITUDINAL AND SURVIVAL ENDPOINTS

BAYESIAN EVALUATION OF LONGITUDINAL/SURVIVAL TRIALS

Donald A. Berry*, University of Texas M. D. Anderson Cancer Center

I will describe Bayesian approaches in the design and analysis of clinical trials with long-term endpoints. The goals are (i) more efficient clinical trials and clinical development programs, and (ii) treating patients more effectively, both those in and those outside of trials. Many types of innovative designs in these kinds of trials have been used at my home institution, the University of Texas M. D. Anderson Cancer Center, in national oncology studies and in pharmaceutical and medical device industry-sponsored trials. I will provide some background on Bayesian designs for clinical trials and give case studies of the Bayesian approach used in actual designs and analyses presented to the FDA. These examples include the possibility of early stopping and variations on themes such as: seamless phases II and III trials with sequential sampling and exploiting correlations of early endpoints (such as biomarkers) with the primary long-term endpoint. I will specifically address the issue of monitoring trial results using Bayesian predictive probabilities, both prospectively and by data and safety monitoring committees.

email: dberry@mdanderson.org

SAMPLE SIZE RE-ESTIMATION IN SURVIVAL STUDIES

Thomas D. Cook*, University of Wisconsin-Madison

Long term randomized clinical trials with failure time endpoints are typically designed based on a set of simple assumptions regarding accrual rates, dropout rates, hazard rates and treatment effects. Some assumptions may be that accrual or hazard rates are constant, or that hazards are proportional. As the trial unfolds, it is often desirable that the validity of these assumptions be evaluated and that the effect that any observed deviations may have on the integrity or feasibility of the trial be assessed. There are additional complications that arise when dealing with interim data such as delays in event reporting and adjudication. This presentation will include a discussion of useful methods for performing the required assessments using interim data. Some practical examples will be discussed.

email: cook@biostat.wisc.edu

STOCHASTIC CURTAILMENT ESTIMATION IN SURVIVAL STUDIES

Dan L. Gillen*, University of California-Irvine

Researchers frequently elect to evaluate new therapies on the basis of patient survival over a well-defined period of time. For example, clinicians might consider five-year survival when investigating drugs developed for use in childhood cancer, or 28-day survival when investigating the treatment of sepsis in patients suffering traumatic injury. However, for ethical reasons data must be periodically analyzed for early indications of efficacy, futility, or harm. In this case, group sequential methodology is typically used to generate multiple criteria for guiding the decision of whether a trial should be stopped early. As currently implemented, these criteria generally assume proportional hazards and are defined for settings where the average effect of treatment up to the interim analysis is the same as that which would be observed if the trial continued on to maximum duration. In this talk, we address the uncertainty of future observations under potentially nonproportional hazards alternatives. We propose a method of imputation of future treatment effects based on random walks, which assumes minimally informative Bayesian prior distributions on the smoothness of survival. Imputation of future survival differences is carried out using standard Bayesian predictive distributions, thereby allowing for estimation of measures of stochastic curtailment.

email: dgillen@uci.edu

EVALUATION OF STOPPING RULES AND SECONDARY ENDPOINTS FOR LONGITUDINAL STUDIES

John M. Kittelson*, University of Colorado Health Sciences Center

We consider clinical trials with longitudinal endpoints in which the primary outcome is a summary measure over a long follow-up time period (e.g., the 5-year rate of change). In such trials interim analyses may be conducted before any subject has been observed for the entire follow-up period so that the primary outcome cannot be estimated. One possible solution is to calculate the same summary measure over a shorter interval (e.g., the 2-year rate of change), however we show that this can give biased inference relative to the primary (longer-term) outcome, and can result in later interim analyses occurring with less statistical information than earlier analyses. We propose an alternative solution that uses interim decision rules based on a posterior distribution for the primary summary measure. We explore parameterizations of a prior distribution to capture a range of clinically feasible outcome trajectories. Early data are then used to construct a posterior distribution for outcome trajectories which forms the basis for inference over the entire follow-up period. We explore the robustness of this approach and illustrate its application using data from a study of a new treatment for peripheral arterial disease.

email: john.kittelson@uchsc.edu

VARIABLE SELECTION IN NONPARAMETRIC RANDOM EFFECTS MODELS

David B. Dunson*, NIEHS
Bo Cai, NIEHS

In analyzing longitudinal, clustered or multivariate data, latent variable models are very commonly used. Examples include random effects models, factor analytic models and structural equation models, which are broadly used in social science applications. A common criticism of latent variable models is possibly sensitivity to modeling assumptions, including the latent variable and residual distributions, number of latent variables, structural relationships among the latent variables, and measurement structure. In performing inferences and making predictions based on such models, it is important to accommodate model uncertainty. Unfortunately, standard methods of Bayes model averaging may not be appropriate due to parameter constraints and other issues. In this talk, I describe novel Bayesian methods for a broad class of latent variable model uncertainty problems and illustrate these methods using biomedical data examples.

email: dunson1@niehs.nih.gov

LONGITUDINAL PROFILING OF HEALTH CARE UNITS BASED ON CONTINUOUS AND DISCRETE PATIENT OUTCOMES

Michael J. Daniels*, University of Florida
Sharon-Lise Normand, Harvard Medical School

Monitoring health care quality involves combining continuous and discrete outcomes measured on subjects across health care units over time. This article describes a Bayesian approach to jointly modeling multilevel multidimensional continuous and discrete outcomes with serial dependence. The overall goal is to characterize trajectories of traits of each unit. Underlying normal regression models for each outcome are used and dependence among different outcomes is induced through latent variables. Serial dependence is accommodated through modeling the pairwise correlations of the latent variables. Methods are illustrated to assess trends in quality of health care units using continuous and discrete outcomes from a sample of adult veterans discharged from one of twenty-two Veterans Integrated Service Networks with a psychiatric diagnosis between 1993 and 1998.

email: mdaniels@stat.ufl.edu

LATENT VARIABLE MODELS FOR MULTIPLE NON-COMMENSURATE OUTCOMES

Armando Teixeira-Pinto*, Harvard Graduate School of Arts & Science and Faculty of Medicine University of Porto
Sharon-Lise T. Normand, Harvard Medical School and 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 present a multivariate model to analyze binary and continuous correlated outcomes using a latent variable. We contrast this method with other approaches in the literature and show that this model is equivalent to the factorization method proposed by Catalano and Ryan (1992). A motivating example evaluating the quality of treatment of schizophrenia in patients who were and were not enrolled in managed care illustrates the different approaches.

email: apinto@fas.harvard.edu

BAYESIAN APPROACHES TO HIERARCHICAL FACTOR ANALYSIS

A. James O'Malley*, Harvard Medical School
Alan M. Zaslavsky, Harvard Medical School

Health care quality surveys in the USA are administered to individual respondents to evaluate performance of health care units (e.g. health plans). We analyze relationships between quality measures at the unit level, by applying techniques such as factor analysis to covariance structure estimated at the unit level in a hierarchical model. At the lower (patient) level we first fit generalized variance-covariance functions that take into account the nonresponse patterns in the survey responses. A between unit covariance matrix is then estimated using a hierarchical model, which evaluates the fitted generalized variance-covariance functions to account for sampling variation. Maximum quasilikelihood and Bayesian inferential procedures are used for model fitting. At the second (e.g. plan) level, we propose to estimate an unstructured covariance matrix and then apply an exploratory factor analysis to summarize relationships between measures. Results will include a description of the computational difficulties encountered and solutions used to overcome these.

email: omalley@hcp.med.harvard.edu

88. ROC ANALYSIS IN BIOMEDICAL INFORMATICS

OPTIMAL ESTIMATION OF ROC CURVES OF CONTINUOUS-SCALE TESTS

Xiao-Hua A. Zhou*, University of Washington and VA Puget Sound Health Care System
Huazhen Lin, University of Washington and Sichuan University

In this talk, we introduce a new semi-parametric maximum likelihood estimate of an ROC curve that satisfies the property of invariance of the ROC curve. In our simulation studies, we demonstrate that the proposed estimator has the best performance among all existing semi-parametric estimators. Finally, we illustrate the application of the proposed estimator using a real data set.

email: azhou@u.washington.edu

NON-PARAMETRIC SEQUENTIAL TESTING OF THE AREA UNDER THE ROC CURVES

Aiyi Liu*, National Institute of Child Health and Human Development
Chengqing Wu, National Institute of Child Health and Human Development
Enrique F. Schisterman, National Institute of Child Health and Human Development

We consider evaluation and comparison of the diagnostic accuracy of biomarkers with continuous test outcomes. We develop nonparametric group sequential testing procedures to 1)evaluate the area of a single biomarker under its receiver operating characteristic (ROC) curve, and 2)compare the area of two biomarkers under their ROC curves, with either independent or paired test outcomes. These procedures rely on the construction of a two-dimensional statistics of Whitehead (1999) so that design methods based on Brownian motion can be applied.

email: liua@mail.nih.gov

ASSESSING RATER PERFORMANCE IN IMAGE SEGMENTATION

Simon K. Warfield*, Harvard Medical School
Kelly H. Zou, Harvard Medical School
William M. Wells, Harvard Medical School

Our objective was to develop a mechanism for assessing human and machine performance in image segmentation despite the absence of a known reference standard for clinical images. Simultaneous estimation of a reference standard and rater performance parameters has been investigated for labelled images. Image segmentations represented by distance from the boundary contour may not be directly addressed by previous methods. We developed a new statistical estimator that operates on a collection of segmentations, and is capable of estimating a reference boundary contour and rater performance in terms of mean offset from the true boundary and variance of boundary location. The estimation problem is formulated as an Expectation-Maximization optimization. The estimation scheme was applied to synthetic phantoms with known ground truth, and was found to accurately estimate the true parameters. The method was also applied to magnetic resonance images of patients with brain tumors. The method was able to accurately identify poor segmentations and to characterize rater performance. Our new algorithm is readily applicable to clinical imaging data and enables assessment of human and machine rater performance. It generalizes prior algorithms that are limited to multi-category segmentations.

email: warfield@crl.med.harvard.edu

89. STATISTICAL CONTRIBUTIONS TO THE FRONTIERS OF HIV/AIDS RESEARCH

A GENERAL GAMMA-BASED HISTORY OF SURVIVAL AFTER AIDS: 1984-2004

Alvaro Muñoz, Johns Hopkins Bloomberg School of Public Health
Christopher Cox*, Johns Hopkins Bloomberg School of Public Health
Haitao Chu, Johns Hopkins Bloomberg School of Public Health
Michael Schneider, Johns Hopkins Bloomberg School of Public Health

The first objective is to present a taxonomy of the generalized gamma distribution in terms of the shape of the hazard functions. Specifically, the plane defined by the scale and shape parameters is divided in four regions according to the identity and inverse functions so that hazards in the 'west' are increasing, in the 'north' have the shape of an arc, in the 'east' the hazards are decreasing and in the 'south' the hazards have a bathtub shape. In addition, the general gamma includes the Weibull, lognormal, standard gamma and their corresponding inverses as particular cases. The second objective is to use a period analysis of data from the Multicenter AIDS Cohort and Women's Interagency HIV studies to characterize the survival times after AIDS diagnosis in four eras defined by the use of different types of antiretroviral therapies. The results not only indicate how progressively effective therapies have been, but they demonstrate how a fatal infectious disease has become a condition which resembles a chronic disease. In addition, we assessed agreement of delta method and bootstrap-based procedures for confidence intervals for non-proportional hazards and non-proportional times. These methods provide ways to link cohort studies with public health.

e-mail: ccox@jhsph.edu

CHALLENGES IN DESIGNING A STUDY TO EVALUATE WHETHER USE OF ANTRIRETROVIRAL THERAPY TO PREVENT MOTHER TO CHILD HIV TRANSMISSION IMPACTS FUTURE MATERNAL TREATMENT OPTIONS

Michael D. Hughes*, Harvard School of Public Health

A single dose of the antiretroviral drug, nevirapine (SD-NVP), given during labor is effective in reducing mother to child transmission (MTCT) of HIV. It is inexpensive and widely recommended for use in resource-limited settings. However, there are now considerable data showing that most women who receive SD-NVP have NVP-resistant virus circulating in their blood during the immediate postpartum period. NVP and other drugs in the same class (non-nucleoside reverse transcriptase inhibitors, NNRTIs) are also widely recommended as part of the initial treatment for HIV-infected individuals and have been widely adopted in treatment programs in resource-limited settings. It is unknown whether the presence of NVP-resistant virus during the postpartum period affects the efficacy of NNRTI-based treatment when the mother subsequently needs treatment for her own health. The talk will describe the challenges in designing a study to address this question. The study comprises a linked pair of randomized trials undertaken in parallel: one among women with and one among women without prior SD-NVP exposure.

e-mail: mhughes@sdac.harvard.edu

MODELING THE POPULATION LEVEL EFFECTS OF AN HIV-1 VACCINE IN AN ERA OF HAART

Wasima Rida*, Statistics Collaborative, Inc.
Sonja Sandberg, Framingham State College

A model for HIV transmission is formulated for a homosexual population in which the use of highly active antiretroviral therapy (HAART) to treat HIV infection is incorporated. The basic reproduction number R0 for the spread of HIV is derived under this model. The model is then expanded to include the potential effects of a prophylactic HIV vaccine. The basic reproduction number Rf under the combination of HAART and vaccination is derived and is shown to equal the dominant eigenvalue of the next generation matrix of the epidemic. The critical vaccination fraction f* necessary to eliminate disease is defined as the minimum vaccination fraction f for which Rf is less than or equal to 1.0. When the basic reproduction number R0 is large or an HIV vaccine is only partially effective, the critical vaccination fraction may exceed 1.0. HIV vaccination, however, may still reduce the prevalence of disease. For a vaccine with limited ability to reduce susceptibility to infection, prevalence of disease can still be reduced if the reduction in infectiousness is at least as great as the reduction in the rate of disease progression. In particular, a vaccine that reduces infectiousness during acute infection may have an important public health impact especially if coupled with counseling to reduce risk behavior.

e-mail: wasima@statcollab.com

INVESTIGATING ASSOCIATIONS BETWEEN FUNCTIONAL PATTERNS OF IMMUNE RESPONSE TO HIV AND DISEASE PROGRESSION

Martha Nason*, Biostatistics Research Branch, NIAID, NIH

Establishing an immune correlate of protection from infection is crucial to the development of vaccines for HIV. In the absence of a protective vaccine, understanding correlates of disease progression in infected individuals is a first step. Historically, neither the quantity nor breadth of the HIV-specific T cell response correlate conclusively with protection from disease progression. Therefore, interest is turning towards the quality of the HIV-specific T cell response, and towards exploring relationships between progression and patterns of T cell functionality. The data available to assess these patterns are complex, with simultaneous measurement of several parameters on individual immune cells for each person, perhaps measured over time and when stimulated with each of a number of peptides. This talk will focus on some of the unique challenges in analyzing this type of data and describe methods and results from an analysis comparing functional patterns in CD8+ T-cells for HIV+ progressors and non-progressors. We describe a two-stage process for choosing measures of functionality based on a comparison of the cohorts, followed by an assessment of the relationship between the measures of functionality and disease progression within the progressors.

e-mail: mnason@Niaid.nih.gov

90. MISSING DATA IN LONGITUDINAL DATA ANALYSIS

A LATENT-CLASS MIXTURE MODEL FOR INCOMPLETE LONGITUDINAL DATA

Caroline Beunckens*, Hasselt University
Geert Molenberghs, Hasselt University
Geert Verbeke, Biostatistical Center-K. U. Leuven

In the analyses of incomplete longitudinal clinical trial data, there has been a shift, away from simple ad hoc methods that are valid only if the data are missing completely at random (MCAR), to more principled (likelihood-based or Bayesian) ignorable analyses, which are valid under the less restrictive missing at random (MAR) assumption. The availability of the flexible standard statistical software allows for such analyses in practice. While the possibility of data missing not at random (MNAR) cannot be ruled out, it is argued that analyses valid under MNAR are not well suited for the primary analysis in clinical trials. Therefore, rather than either omitting or blindly shifting to an MNAR framework, the optimal place for MNAR analyses arguably is within a sensitivity analysis context. There are several sensitivity analysis routes available. We propose a flexible model, based on a latent-class mixture formulation, sharing features with selection, pattern-mixture, and shared-parameter models.

e-mail: caroline.beunckens@uhasselt.be

A CENSORED MULTINOMIAL MODEL FOR BINARY, LONGITUDINAL SURVEY DATA WITH MISSING VALUES

Steven J. Mongin, University of Minnesota School of Public Health
Timothy R. Church*, University of Minnesota School of Public Health

Longitudinal surveys often have missing data due to occasional non-response. A multinomial model accommodates the missing values as censoring. For application to a longitudinal study of cancer screening behavior, estimation and inference is obtained through a Bayesian analysis with a uniform prior on the joint distribution of the multinomial probabilities. A simple computational approach for Gibbs sampling is developed for use with the model. Consideration is given to the consequences of the uniform prior in terms of the induced priors on other usual parameters of interest in longitudinal studies.

e-mail: trc@cccs.umn.edu

MISSING PHENOTYPE DATA IMPUTATION FOR LONGITUDINAL PEDIGREE DATA ANALYSIS

Mariza de Andrade, Mayo Clinic
Brooke Fridley*, Mayo Clinic

Mapping complex traits to relatively small genetic effects whose phenotypes may be modulated by temporal trends is challenging. Missing data complicates matters in the longitudinal data analysis. Since most analytical methods developed for the analysis longitudinal pedigree data require no missing data, the researcher is left with the option of dropping those individuals with missing data from the analysis or imputing values for the missing data. Though methods to handle missing data have been an area of statistical research for many years, little has been done within the context of a longitudinal pedigree analysis. We will present the use of data augmentation within a Bayesian longitudinal polygenic model to produce k complete longitudinal datasets. The data augmentation will take into account the observed familial information and observed subject information available at other time points. These k complete longitudinal datasets can then be used to fit a longitudinal pedigree model using the S-Plus library MULTIC. By producing a set of k complete datasets and thus k set of parameter estimates, the total variance associated with an estimate can be partitioned into a within-imputation and a between-imputation component. The method will be illustrated using the Genetic Analysis Workshop (GAW13) simulated data.

e-mail: mandrade@mayo.edu

SENSITIVITY ANALYSIS AND INFORMATIVE PRIORS FOR LONGITUDINAL BINARY DATA WITH OUTCOME-RELATED DROPOUT

Joo Yeon Lee*, Brown University
Joseph W. Hogan, Brown University

Dropouts are common in longitudinal data, and sensitivity analysis is indispensable to evaluate the effect of untestable assumptions about dropout mechanisms on study conclusion. This paper develops pattern mixture models, composed of marginalized transition models within pattern, for repeated binary data. Within this framework we propose several approaches to sensitivity analyses and incorporation of prior information that allow the analyst to explore the effects of possibly outcome-related dropout. We show how to represent and convey uncertainty about common assumptions such as MAR using prior distributions on the sensitivity parameters, and how to use prior distributions that reflect beliefs about the distribution of missing responses. Methods are illustrated using data from the OASIS study.

e-mail: jooyeon@stat.brown.edu

AN EXTENSION OF LATENT VARIABLE MODEL FOR INFORMATIVE INTERMITTENT MISSING DATA

Li Qin*, University of Pittsburgh
Lisa A. Weissfeld, University of Pittsburgh
Melissa Kalarchian, University of Pittsburgh
Marsha Marcus, University of Pittsburgh

In longitudinal studies, subjects are followed over time, so missing data are a frequent problem. We propose a latent variable model for informative intermittent missingness which is an extension of Royi¯s (2003) latent dropout class model. In our model, the value of the latent variable is affected by the missing pattern and it is also as a covariate in modeling the longitudinal response. Using this approach, the latent variable links the longitudinal response and the missing process. In our model the latent variable is continuous instead of categorical and we assume that it is from a normal distribution with unity variance. To simplify the analysis for intermittent missing patterns, we define two variables: one for the dropout time, and the other for the number of missing time points before dropout. The EM algorithm is used to obtain the estimates of the parameter we are interested in and Gauss-Hermite quadrature is used to approximate the integration of the latent variable (Sammel, et al., 1997). The standard errors of the parameter estimates are obtained from the inverse of the Fisher information matrix of the final marginal likelihood. This method is illustrated using data from a pediatric obesity study on evaluating the effectiveness of family-based intervention. We use the generalized Pearson residuals to assess the fit of the model.

e-mail: liq1@pitt.edu

IMPUTATION APPROACH FOR RESPONDERS ANALYSIS IN LONGITUDINAL STUDIES WITH RANDOM MISSING DATA

Liqiu Jiang*, North Carolina State University
Kaifeng Lu, Merck & Co., Inc.
Anastasios A. Tsiatis, North Carolina State University

Often a binary variable is generated by dichotomizing a underlying continuous measurement at the primary time point of interest according to a prespecified threshold value. Normally, people would use a logistic model to estimate covariates effects on the binary responses. In the event that the underlying continuous measurements are from a longitudinal study, the repeated measurements are often analyzed using a repeated measures model because of mathematical and computational convenience of available off-the-shelf software. This practical advantage motivates us, in this article, to use repeated measures model as an imputation approach in the presence of missing data on the responder status as a result of patient drop-out before normal completion of the study. We, then, apply the logistic regression model on the observed or otherwise imputed responder status. Large sample properties of the resulting estimators are derived and simulation studies carried out to assess the performance of the estimators in the situation in which either continuous repeated measurements are misspecified as multinormal distribution or responser status as logistic distribution as the incompatibility of these two models. KEY WORDS: missing data; multiple imputation; repeated measures; logistic regression.

e-mail: liqiu00@yahoo.com

A SELECTION MODEL FOR FUNCTIONAL MAPPING OF LONGITUDINAL TRAITS WITH NON-IGNORABLE MISSING DATA

Hongying Li*, University of Florida
Rongling Wu, University of Florida

Functional mapping has emerged as a powerful statistical tool for mapping and identifying quantitative trait nucleotides (QTN) that regulate the process and pattern of longitudinal responses. In this talk, we will present a statistical framework for embedding the idea of functional mapping within a general clinical trial, aimed to unveil the genetic architecture of inter-patient variation in longitudinal responses with incomplete data. In clinical trials, there are always some patients who drop out early due to side effects or limited duration, presenting a significant challenge in statistical inference. We develop a selection model for functional mapping of longitudinal traits incorporating this so-called non-ignorable drop-out information, whose missing responses depend on unobserved and maybe observed information. We derive the EM algorithm to estimate the parameters that are related to the haplotype frequencies of QTN, the action and interaction effects of risk haplotypes and the autoregressive or antedependence structure of covariance matrix of the longitudinal process. The model is validated by a real example for the pharmacogenomic study of drug response. Our model, in a couple with routine clinical trials, will have great implications for the detection and mapping of specific DNA variants that encode differentiation in longitudinal responses.

e-mail: hli@stat.ufl.edu

SAMPLING WEIGHTED RELATIVE RISK REGRESSION

Rickey E. Carter*, Medical University of South Carolina
Stuart R. Lipsitz, Medical University of South Carolina

Recently, the estimation of relative risk (in lieu of the odds ratio) in prospective studies has received increased attention. However, iterative routines maximizing the log-linear model based on the Bernoulli likelihood may encounter convergence difficulties, especially when the underlying probability of success approaches 1.0. Others have proposed the use the Poisson likelihood and robust variance estimator to consistently estimate relative risk and its associated confidence interval. This approach, while successfully addresses the convergence limitations of the Bernoulli likelihood, has one important limitation. Namely, the estimates of the success probabilities for certain configurations of the linear component may exceed 1.0. In this presentation, we present a modified Bernoulli likelihood that consistently estimates relative risk, improves convergence over the unmodified Bernoulli likelihood, and provides estimates of the success probability in the range of 0 to 1. The method is illustrated using a prospective clinical trial.

e-mail: carterre@musc.edu

MISSPECIFICATION TESTS FOR DISCRETE DATA MODELS

Marinela Capanu*, Memorial Sloan-Kettering Cancer Center
Brett Presnell, University of Florida

The IOS test of Presnell & Boos (2004) is a general purpose goodness-of-fit test based on a ratio of in-sample and out-of-sample likelihoods. For large samples, the IOS statistic can be viewed as a contrast between two estimates of the information matrix that are equal under correct model specification. Both the IOS test and its large sample approximation are simple to compute and broadly applicable for testing the adequacy of parametric models. We compare the performance of IOS with existing goodness-of-fit tests for a variety of discrete data models. Our findings suggest that IOS is strongly competitive, not only with other general purpose tests, but even with tests designed especially for a specific model.

e-mail: capanum@mskcc.org

A TEST FOR NON-INFERIORITY WITH A MIXED MULTIPLICATIVE/ADDITIVE NULL HYPOTHESIS

Xiaodan Wei*, University of Wisconsin-Madison
Richard J. Chappell, University of Wisconsin-Madison

The equivalence of two treatments or drugs with binomial outcomes is often measured in terms of the difference, ratio or odds ratios between two sample proportions. In this paper, we propose a new setting of the non-inferiority hypothesis, called mixed null hypothesis, which combines the nonzero difference and non-unity ratio tests together. It tests ratios when proportions are small and differences when they are big. The mixed null hypothesis approach does not require us to choose between nonzero difference and non-unity ratio equivalence tests either in planning the trial or at the time of analysis. A test statistic is derived for the mixed null hypothesis, and asymptotic properties of this test statistic are provided. Simulation results show the actual size of test is close to the actual significance level .05. The mixed null hypothesis approach has almost the same power as the difference or ratio tests when the same equivalence limit is used. Two real clinical trials are studied here to conlcude that the mixed null hypothesis test gives almost same p-value as ratio test when event rates are low and similar result as difference test when they are high.

e-mail: xwei@stat.wisc.edu

A PENALIZED LATENT CLASS MODEL FOR ORDINAL RESPONSES

Stacia M. DeSantis*, Harvard School of Public Health
E. Andres Houseman, Harvard School of Public Health
Brent A. Coull, Harvard School of Public Health
Rebecca A. Betensky, Harvard School of Public Health

Applying latent class methodology to correlated, high-dimensional ordinal data poses many challenges. Unconstrained analyses may not result in an adequate model fit. Thus, information contained in ordinal variables may not be exploited by researchers. Using the ridge penalty, we develop a penalized latent class model to analyze high-dimensional ordinal data. By regularizing maximum likelihood estimation, this technique allows us to fit an ordinal latent class model that would otherwise not be feasible without applying strict constraints. We illustrate our methodology in the context of schwannoma, a peripheral nerve sheath tumor. Many ordinal histological features are measured in order to characterize clinical subsets of schwannoma, which may have different prognoses. The methodology is applied to the data in order to arrive at latent classes that may elucidate subsets.

e-mail: sdesanti@hsph.harvard.edu

ASYMPTOTICALLY OPTIMAL INFERENCE WITH PARTIALLY OBSERVABLE BINARY DATA: APPLICATIONS TO PLANT DISEASE ASSESSMENT

Joshua M. Tebbs*, University of South Carolina
Melinda H. McCann, Oklahoma State University

Insect-vectored plant diseases impact the agricultural community each year by affecting the economic value, the quantity, and the quality of crops. Controlling the spread of disease is an important area in risk assessment and understanding the dynamics of vector populations helps researchers to develop effective treatments. In this talk, we investigate an experimental design commonly-used by researchers who study plant disease, and we derive large sample hypothesis tests that may be used to characterize disease-transmission behavior in a stratified population. Small-sample results via simulation are also provided. As we illustrate, analyzing data from such experiments can present a challenge because it involves binary random variables which are not be directly observable. We illustrate our proposed methods using an Argentinean study that examines the 'Mal de Rio Cuarto' virus and its transmission to susceptible maize.

e-mail: tebbs@stat.sc.edu

MULTIVARIATE LOGISTIC MODELS

Bahjat F. Qaqish*, University of North Carolina
Anastasia Ivanova, University of North Carolina
Eugenio Andraca, University of North Carolina

The multivariate logistic transform is a reparametrization of cell probabilities in terms of marginal logistic contrasts. However, given an arbitrary set of contrasts, the inverse transform may not exist. We present an efficient algorithm for detecting whether the inverse exists, and for computing it if it does. We compare the algorithm with iterative proportional fitting and Newton-Raphson algorithms. We also discuss the implications for fitting multivariate regression models to correlated binary outcomes.

e-mail: bahjat_qaqish@unc.edu

BAYESIAN ESTIMATE OF ODDS RATIOS FOR SMALL SAMPLE SIZE 2 X 2 TABLES WITH INCOMPLETELY CLASSIFIED DATA

Yan Lin*, Medical University of South Carolina
Stuart R. Lipsitz, Medical University of South Carolina
Debajyoti Sinha, Medical University of South Carolina
Barbara C. Tilley, Medical University of South Carolina
Rickey Carter, Medical University of South Carolina

In studies involving association between two binary variables within each subject, the estimation of odds ratios is often of primary interest. However, often in many studies either as a result of missing observation or due to the nature of observation process, some subjects cannot be fully cross-classified in the 2 x 2 contingency table. Our study is especially focus on the situation that each subject is classified at least by one binary variable and missing mechanism is assumed to be either missing completely at random (MCAR) or missing at random (MAR). In this paper, we propose a Bayesian estimate of odds ratio for 2 x 2 contingency tables with marginal supplements and small sample sizes. The Bayesian point and interval estimates of odds ratio using the Dirichlet priors for cell probabilities will be compared to the corresponding MLE's in terms of the bias of the estimates, the mean square errors, and the actual coverage of interval estimates.

e-mail: liny@musc.edu

92. ANALYZING MICROARRAY DATA

PREDICTIVE MODEL BUILDING FOR MICROARRAY DATA USING GENERALIZED PARTIAL LEAST SQUARES MODEL

Baolin Wu*, University of Minnesota

Microarray technology enables the simultaneous monitoring of tens of thousands of gene expression values in an entire genome. This results in the microarray data with the number of genes p far exceeding the number of samples n. Traditional classification methods do not work well when $n<<p$. Dimension reduction methods are often required before applying standard statistical methods, popular among them are principal component analysis (PCA) and the partial least squares (PLS) etc. We propose a novel maximum likelihood based framework specially designed for $n<<p$ situation. For continuous response we show that the maximum likelihood estimations correspond to the PLS. For discrete response, we derive the logistic regression model with generalized PLS, and reveal its close connection to the diagonal linear discriminant analysis method. The proposed likelihood based approach can easily incorporate gene selection and estimation shrinkage in a unified model, which is shown closely linked to the commonly used nearest shrunken centroid classifier (also known as PAM). Applications to public microarray data confirm the superiority of the proposed methods. Simulation studies are conducted to study the proposed methods.

e-mail: baolin@biostat.umn.edu

OVERVIEW ON STRUCTURAL ASSOCIATION TESTING AND REGIONAL ADMIXTURE MAPPING

David B. Allison*, University of Alabama at Birmingham
T.M. Beasley, University of Alabama at Birmingham
Jose R. Fernandez, University of Alabama at Birmingham
David T. Redden, University of Alabama at Birmingham
Hemant K. Tiwari, University of Alabama at Birmingham
Jasmin Divers, University of Alabama at Birmingham
Robert Kimberly, University of Alabama at Birmingham

Individual genetic admixture estimates, determined both across the genome and at specific genomic regions, have been proposed for use in identifying specific genomic regions harboring loci influencing dichotomous phenotypes in regional admixture mapping (RAM). Estimates of individual ancestry can be used in structured association tests (SAT) to reduce confounding induced by various forms of population substructure. Although presented as two distinct approaches, we provide a conceptual framework in which both RAM and SAT are special cases of a more general linear model which allows for greater modeling flexibility, adaptation to multiple designs, inclusion of covariates, interaction terms, and multi-locus models. We clarify which variables it is sufficient to control for in analyses and also provide a simple closed-form 'semi-parametric' method of estimating the reliability of individual admixture estimates used as individual ancestry estimates that makes an inherent errors-in-variables problem tractable. This approach offers SAT and RAM methods enormous flexibility, enabling application to a richer set of phenotypes, populations, covariates, and situations.

e-mail: dallison@uab.edu

INCORPORATING GENE FUNCTIONS 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 activity of a transcription factor (TF) on all the genes; for example, any TF is assumed to be either an activator or a repressor, but not both, while it is known that some TFs can be dual regulators. 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. 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. An application to a yeast dataset and simulations lend support for our proposed methods. 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 methods

e-mail: weixx035@umn.edu

HAPLOTYPE FREQUENCY ESTIMATION FROM POOLED GENOTYPES: A CONTINGENCY TABLE PERSPECTIVE

Yaning Yang, University of Science and Technology of China
Jinfeng Xu*, Columbia University
Zhiliang Ying, Columbia University
Jurg Ott, Laboratory of Statistical Genetics-Rockefeller University

Pooling DNA samples of multiple individuals has been advocated as a method to reduce genotyping costs. Under such a scheme, instead of haplotype counts, only allele counts at each locus are observed. We develop a systematic way for analyzing such data by formulating the problem under a contingency table perspective. We show that each pooled sample can be conveniently expressed as margins of a contingency table, with haplotype counts corresponding to (unobserved) cell counts. The cell frequencies can be uniquely determined from the marginal frequencies under the usual Hardy-Weinberg equilibrium assumption. The maximum likelihood estimates of haplotype frequencies are shown to be consistent and asymptotically normal as the number of tables increases. An explicit formula for the asymptotic variance-covariance matrix of the maximum likelihood estimates is derived and shown to be closely related to the generalized hypergeometric distribution. Simpler lower and upper bounds for the asymptotic variance-covariance matrix are given.

e-mail: xu@stat.columbia.edu

EIGENGENE BASED LINEAR DISCRIMINANT MODEL FOR GENE EXPRESSION DATA ANALYSIS

Ronglai Shen*, University of Michigan
Zhaoling Meng, Sanofi-Aventis
Debashis Ghosh, Comprehensive Cancer Center-University of Michigan
Arul M. Chinnaiyan, Comprehensive Cancer Center-University of Michigan

We propose an eigengene based linear discriminant analysis (ELDA) for tumor classification using high-dimensional gene profiling data. The algorithm selects essential eigengenes to be included in a classifier by orthogonalizing the highly-correlated data structure. It allows compact gene signatures to be identified and shows robust classification/prediction performances in two cancer microarray data sets. A second feature of our algorithm is the incorporation of a misclassification cost matrix. Differential penalization of one type of classification error over another is desirable in many important decision makings, and yet receives little investigation. In the breast cancer gene profiling study (van't Veer et al., 2002), false negative error is considered more costly to patients as it leads to treatment delays by predicting a poor prognosis patient to do well. We derived a cost-adjusted discriminant function that allows unequal penalization of misclassification errors. By assigning a larger cost to false negatives, a high-sensitivity classifier can be obtained while maintaining the overall errors at a reasonable level. Desirable sensitivities and specificities can be obtained by tuning the cost matrix parameters, and generalization to multi-class problems is immediate.

e-mail: rlshen@umich.edu

INCORPORATING BIOLOGICAL KNOWLEDGE INTO TUMOR CLASSIFICATIONS WITH MICROARRAY DATA

Feng Tai*, University of Minnesota
Wei Pan, University of Minnesota

In the context of sample (e.g. tumor) classifications with microarray gene expression data, many methods have been proposed. However, almost all the methods ignore existing biological knowledge, and treat a priori all the genes equally. On the other handm some genes are known to have biological functions or to be involved in pathways related to disease, and thus these genes are likely to be more relevant. Here we propose incorporating such biological knowledge into building a classifier to improve interpretability and prediction performance of the resulting model.
Key words: Gene expression; Gene ontology (GO); PAM; Penalization.

e-mail: fengtai@biostat.umn.edu

93. NONPARAMETRIC AND SEMIPARAMETRIC METHODS

MIXED-EFFECTS, POSTERIOR MEANS AND PENALIZED LEAST SQUARES

Yolanda Munoz Maldonado*, UT-HSC School of Public Health at Houston

In this paper we extend the known equivalence between the numerical solutions of 1) penalized splines type estimators, 2) best linear unbiased predictor of a particular mixed-effects model and 3) the posterior mean of Gaussian signal-plus-noise model with diffuse initial conditions. We prove that, when considering the more general cases of penalized least-squares error criterion, the mixed-effects model, and the posterior mean of a random-effects model with a diffuse prior in some of the random effects, the equivalence between the numerical answers of these different settings still holds. Three examples are chosen to illustrate the application of this general model in the frameworks of varying coefficient models, ridge regression and randomized block design. A small theorem shows that the methods of generalized cross-validation, generalized maximum likelihood estimation and unbiased risk prediction can be used to estimate the variance components or smoothing parameters in the three settings. The application of our main result allows to utilize in an ample number of cases an efficient order $O(n)$ Kalman filter algorithm to obtain the desired predictors and corresponding Bayesian confidence intervals.

e-mail: Yolanda.M.Munoz@uth.tmc.edu

PENALIZED FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS USING A KULLBACK-LEIBLER CRITERION

Robert T. Krafty*, University of Pennsylvania School of Medicine
Wensheng Guo, University of Pennsylvania School of Medicine

In this article we propose a new method for nonparametric functional principal component analysis. The method is conceptually derived from a two-stage analysis, in that we first apply a penalized smoothing operator to extract an underlying smooth signal from each individual curve, and then calculate the covariance function and the functional principal components from the smoothed curves. The penalty plays a different role than it does in the estimation of subject curves in that it controls the smoothness of the principal components as opposed to the smoothness of the subject curves. We propose to use a Kullback-Leibler criterion to measure the goodness-of-fit of our estimator to the empirical distribution and develop a leave-one-curve-out cross validation for estimation of the smoothing parameter and the number of principal components. The method can be applied to unequally spaced sparse data, and is shown to perform better than other existing methods in our simulation. Conistency rates for the covariance estimator and for the principal components are computed. We apply our method to an epileptic EEG data set to understand the seizure generation mechanism.

e-mail: rkrafty@cceb.upenn.edu

ESTIMATING LINEAR FUNCTIONALS OF INDIRECTLY OBSERVED INPUT FUNCTIONS

Eun-Joo Lee*, Illinois College

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@ic.edu

A NONPARAMETRIC LIKELIHOOD RATIO TEST TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES FROM MICROARRAY DATA

Sunil Mathur*, University of Mississippi
Sankar Bokka, University of Mississippi

Motivation: Microarray experiments contribute significantly to the progress in disease treatment by enabling a precise and early diagnosis. One of the major objectives of microarray experiments is to identify differentially expressed genes under various conditions. The statistical methods, currently available in literature to analyze microarray data are not up to the mark, mainly due to the lack of understanding of the distribution of microarray data. Results: We present a nonparametric likelihood ratio (NPLR) test to identify differentially expressed genes using microarray data. The NPLR test is highly robust against extreme values and does not assume the distribution of parent population. Simulation studies show that the NPLR test is more powerful than some of the commonly used methods, such as two-sample t-test, Mann-Whitney U-test and Significance Analysis of Microarray (SAM). When applied to microarray data, it is found that the NPLR test identifies more differentially expressed genes than its competitors. The asymptotic distribution of the NPLR test statistic, and the p-value function is presented. The application of NPLR method is shown is using both synthetic and real-life data. Biological significance of some of the genes detected only by NPLR method is discussed.

e-mail: skmathur@olemiss.edu

MODELING OF HORMONE SECRETION-GENERATING MECHANISMS: A SPLINE AND PSEUDO-LIKELIHOOD APPROACH

Anna Liu*, University of Massachusetts
Yuedong Wang, University of California-Santa Barbara

It is commonly believed that hormones are secreted in basal and pulsatile manners. The level of the basal secretion, the pulse frequency, the hormone decay rate from peak to nadir and the pulse amplitude are characteristics of hormone series. How they are altered by various diseases is of interest to medical researchers. Most current approaches require identification of pulse locations in order to estimate these parameters, which is a difficult task as the pulse signals are often masked due to the slow hormone decay rate. Assuming the pulses are generated from a counting process, this paper uses a stochastic model which convolves a pulse shape function and the process, bypassing the requirement of knowing pulse locations. Simultaneous inference on the parameters is based on a nonparametric pseudo-likelihood constructed with the first two moments of the stochastic model, where the intensity function of the counting process is estimated through smoothing splines. One appealing feature of the proposed approach is its robustness to various distributional assumptions of the stochastic components. The approach also deals with multiple hormone series as opposed to most current methods that are two-stage analyses, thus allowing unconditional inferences. Both simulations and applications suggest that the method provides reasonable estimates and valid inferences.

e-mail: anna@math.umass.edu

RANK ESTIMATION OF ACCELERATED LIFETIME MODELS WITH DEPENDENT CENSORING

Limin Peng*, Emory University
Jason P. Fine, University of Wisconsin-Madison

Under independent censoring, estimation of the covariate effects in the accelerated lifetime model may be based on censored data rank tests. Similar rank methodology has been developed with bivariate accelerated lifetime models for dependent censoring, but employs artificial censoring which may lead to substantial information loss. We present a new artificial censoring technique using pairwise ranking and establish the asymptotic properties of a pairwise rank estimator. Simulations show that the pairwise approach achieves large reductionsin artificial censoring and large efficiency gains over the existing rank estimator. The simulations also evidence moderate efficiency gains under independent censoring over a rank estimator which is semiparametric efficient under independent censoring. An AIDS data analysis illustrates the practical utility of the inferential procedures.

e-mail: lpeng@sph.emory.edu

ENAR

Index of Participants

Index of Participants

Index of Participants

Index of Participants

Index of Participants

Index of Participants

Index of Participants

Index of Participants

Index of Participants

Index of Participants

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

NOTES

ENAR WEBSITE
WWW.ENAR.ORG