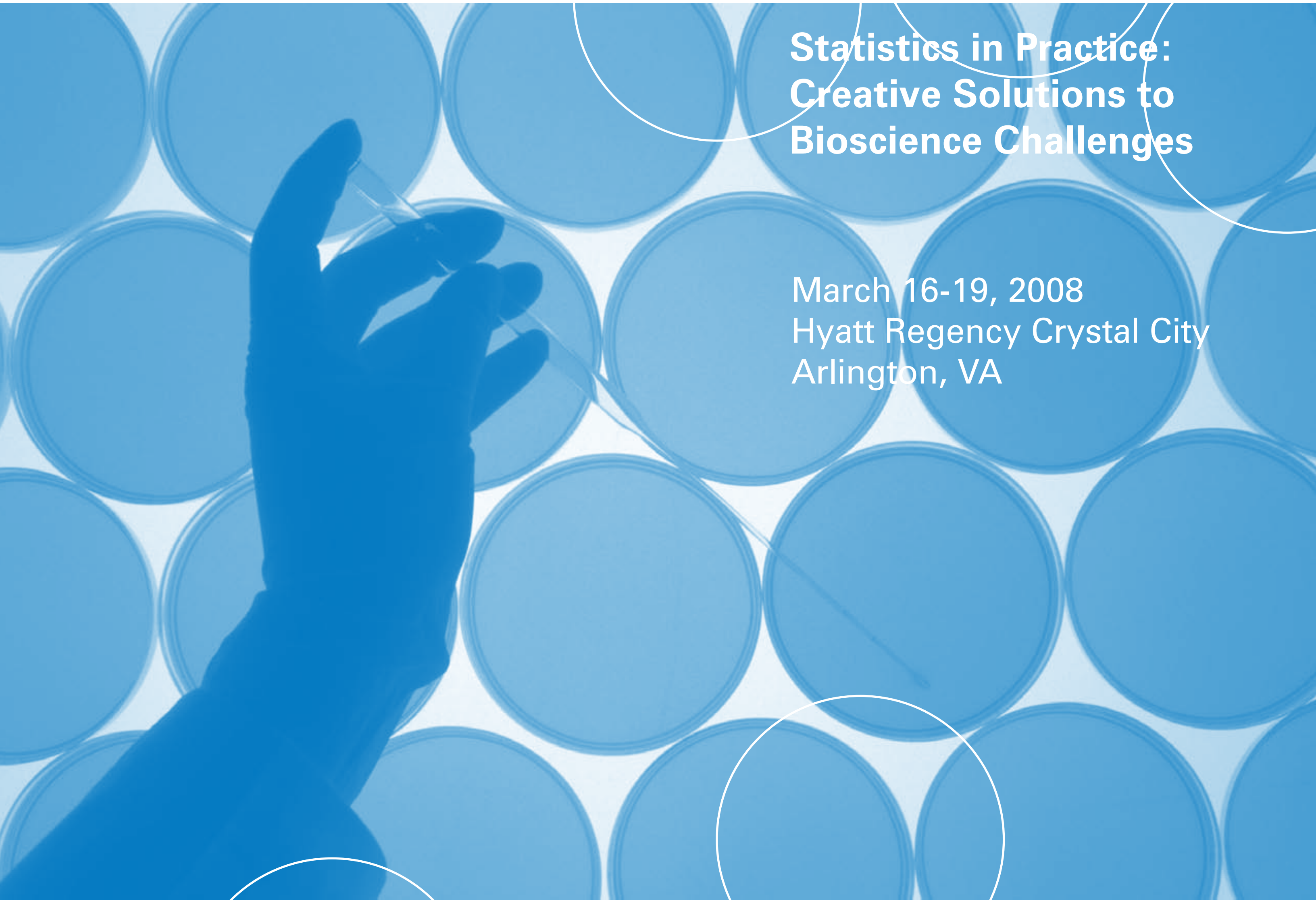
ENAR

International Biometric Society
Eastern North American Region

Spring Meeting with IMS and Sections of ASA



Statistics in Practice: Creative Solutions to Bioscience Challenges

March 16-19, 2008
Hyatt Regency Crystal City
Arlington, VA

PROGRAM AND ABSTRACTS

Contents

Acknowledgements

ENAR would like to acknowledge the generous support of the 2008 Local Arrangements Committee, chaired by Kimberly Drews and the faculty and staff of George Washington University and Guoqing Diao and the faculty and staff of George Mason University.

Sponsors

We gratefully acknowledge the support of:

Abbott Laboratories
Amgen Inc.
AstraZeneca/MedImmune
Biogen Idec
Bristol-Myers Squibb Co.
Cepahlon, Inc.
Cytel Inc.
Daiichi Sankyo
Eli Lilly and Company
GlaxoSmithKline
ICON Clinical Research
Johnson & Johnson
Merck & Co., Inc.
Novartis Pharmaceuticals Inc.
PPD, Inc.
Procter & Gamble
Quintiles, Inc.
Roche Biostatistics
Rho, Inc.
SAS
Schering-Plough
Statistics Collaborative, Inc.
Takeda Global Research & Development
The Emmes Corporation
Wiley- Blackwell
Wyeth Research

Exhibitors

We gratefully acknowledge the support of:

Allergan
American Statistical Association
The Cambridge Group Ltd.
Cambridge University Press
CRC Press – Taylor & Francis Group
Cytel, Inc.
Elsevier
Insightful Corporation
JMP
Kforce Clinical Research
Oxford University Press
PPD, Inc.
Salford Systems
SAS
SAS Publishing
SIAM – Society for Industrial and Applied Mathematics
Smith Hanley Associates LLC
Springer
statistics.com
Wiley-Blackwell

Officers and Committees

January – December 2008

EXECUTIVE COMMITTEE - OFFICERS

President	Eric (Rocky) Feuer
Past President	Lisa LaVange
President-Elect	Lance Waller
Secretary (2007-2008)	José Pinheiro
Treasurer (2008-2009)	Scarlett Bellamy

REGIONAL COMMITTEE (RECOM)

President (Chair) Eric Feuer
Eight ordinary members (elected to 3-year terms): and RAB Chair (Amy Herring)

2006-2008	2007-2009	2008-2010
John Bailer	Karen Bandeen-Roche	Jianwen Cai
Stacy Linborg	F. Dubois Bowman	Bradley Carlin
Tom TenHave	Paul Rathouz	Peter Macdonald

REGIONAL MEMBERS OF THE COUNCIL OF THE INTERNATIONAL BIOMETRIC SOCIETY

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

APPOINTED MEMBERS OF REGIONAL ADVISORY BOARD (3-year terms)

Chair: Amy Herring

2006-2008	2007-2009	2008-2010
Lloyd Edwards	Christopher S. Coffey	Karla V. Ballman
Michael Hardin	Hormuzd A. Katki	Craig Borkowf
Eileen King	Lan Kong	Avital Cnaan
Carol Lin	Yi Li	Kimberly Drews
Keith Muller	Lillian Lin	Matthew Gurka
Soomin Park	Laura Meyerson	Monica Jackson
Shyamal Peddada	Gene Pennello	Robert Johnson
Jeremy Taylor	Tamara Pinkett	Robert Lyles
Melanie Wall	John Preisser	Peter Song
Position to be Filled	Douglas E. Schaubel	Ram Tiwari

PROGRAMS

2008 Spring Meeting – Crystal City, VA

Program Chair: Avital Cnann
Program Co-Chair: KyungMann Kim
Local Arrangements Co-Chairs: Guoqing Diao and Kimberly Drews

2009 Spring Meeting – San Antonio, TX

Program Chair: Brent Coull
Program Co-Chair: Mahlet Tadesse
Local Arrangements Chair: Chen-Pin Wang

2008 Joint Statistical Meeting	2009 Joint Statistical Meeting
Robert Johnson	Lloyd Edwards

Biometrics Editor	Marie Davidian
Biometrics Co-Editors	Geert Molenberghs, Naisyin Wang, and David Zucker
Biometric Bulletin Editor	Urania Dafni
JABES Editor	Carl Schwarz
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

Eric Feuer *(President)*

Lisa LaVange *(Past-President)*

Lance Waller *(President-Elect)*

ENAR Standing/Continuing Committee Chairs

Nominating (2007)	Jane Pendergast
Sponsorship (2008)	Christine Clark

ENAR Representative on the ASA Committee on Meetings

Maura Stokes (January 2006-December 2008)

American Association for the Advancement of Science

(Joint with WNAR) Terms were 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: Eric Feuer

Workshop for Junior Researchers

Brent Coull (Chair)

Debashis Ghosh

Amy Herring

Yi Li

Fostering Diversity Workshop

Scarlett Bellamy *(co-organizer)*

DuBois Bowman *(co-organizer)*

Stacy Lindborg

Amita Manatunga

Renee Moore

Dionne Price

Dejuran Richardson

Louise Ryan

Kimberly Sellers

Keith Soper

Mahlet Tadesse

Tom Ten Have

Lance Waller

Student Awards

ENAR Student Award Committee
Jane F. Pendergast (Chair)
Sudipto Banerjee
Christopher Bilder
Jianwen Cai
David B. Dunson
Elizabeth S. Garrett-Mayer
Debashis Ghosh
Liang Li
Yi Li
Jeffrey S. Morris
Sunil Rao
Joshua Tebbs
Heping Zhang

Student Award Winner
Van Ryzin Award Winner
Lei Xu, University of Michigan

Award Winners
Kwun Chuen Gary Chan, Johns Hopkins University
Baojiang Chen, University of Waterloo
Chongzhi Di, Johns Hopkins University
Kevin Hao Eng, University of Wisconsin – Madison
Bo Huang, University of Wisconsin - Madison
Seowen Jin, Southern Methodist University and University of Texas at El Paso
Hanjoo Kim, University of Pennsylvania School of Medicine
Hong Li, Brown University
Zhiguo Li, University of Michigan
Megan Othus, Harvard University
Qing Pan, George Washington University
Ju-Hyun Park, University of North Carolina – Chapel Hill
Jing Qian, Emory University
Xiaoyan Shi, University of North Carolina - Chapel Hill
Chi Wang, Johns Hopkins University
Michael Wu, Harvard University
Benhuai Xie, University of Minnesota
Min Zhang, North Carolina State University
Xiaoxi Zhang, University of Michigan

Future Meetings of the International Biometric Society

XXIVth International Biometric Conference
Dublin, Ireland, 13-18 July 2008

2009 ENAR Spring Meeting
San Antonio, Texas

2010 ENAR Spring Meeting
New Orleans, LA

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

2008 Special Thanks

2008 ENAR Program Committee
Avital Cnaan (Chair), University of Pennsylvania
KyungMann Kim (Co-Chair), University of Wisconsin
Jesse A. Berlin, Johnson & Johnson
Nilanjan Chatterjee, National Cancer Institute
David B Dunson, National Institute of Environmental Health Sciences
Hongzhe Li, University of Pennsylvania
Karen M. Kuntz, University of Minnesota
Constantine Gatsonis, Brown University
Lance A. Waller, Emory University

ASA Section Representatives
David L. Banks (Section on Statistics in Defense and National Security)
Duke University
Scott R. Evans (Section on Teaching Statistics in the Health Sciences)
Harvard University
Kalyan Ghosh (Biopharmaceutical Section), Merck & Co.
Steven G. Heeringa (Survey Research Methods Section)
University of Michigan
Mevin Hooten (Section on Statistics and the Environment) Utah State University
Timothy D. Johnson (Biometrics Section), University of Michigan
A. James O'Malley (Health Policy Statistics Section), Harvard University
Susan Simmons (Section on Risk Analysis),
University of North Carolina, Wilmington
Dalene K. Stangl (Section on Statistical Education), Duke University
Elizabeth R. Zell (Statistics in Epidemiology Section)
Centers for Disease Control & Prevention

IMS Program Chair
Ram C. Tiwari, National Cancer Institute

ENAR Education Advisory Committee
Barry I. Graubard, National Cancer Institute
Lisa M. LaVange, University of North Carolina
Estelle Russek-Cohen, US Food and Drug Administration
Linda J. Young, University of Florida

Local Arrangements Committee
Kimberly L. Drews (Chair), George Washington University
Guoqing Diao (Co-Chair), George Mason University

ENAR Student Awards Chair
Jane F. Pendergast, University of Iowa

ENAR Diversity Workshop Committee
Scarlett L. Bellamy (Co-chair), University of Pennsylvania
F. DuBois Bowman (Co-chair), Emory University

ENAR Workshop for Junior Researchers
Brent Coull, Harvard University

ENAR Presidential Invited Speaker

Donald A. Berry, Ph.D.

Invited Address: Biostatistics, Science, and the Public Eye

Statisticians hold the scepter of science. Too often we wield it in a supporting role. The public sees us as number people: we can multiply and divide and we know things like the altitude and volume of the world's highest lake. Many of our medical colleagues see us as reservoirs of sample size calculations--the statistician as technician. I fight this attitude, in defense of myself and in defense of our profession. I frequently use my scepter as a weapon. It leaves a mark. And I don't always escape unscathed. But the controversies that arise make life interesting ... and fun! For example, I've been quoted in The New York Times scores of times about such diverse issues as the risks and benefits of cancer screening, mad cow disease, death counts in Iraq, and whether dogs can sniff out cancer. I'll relate some of these stories, addressing statistics, science and politics along the way. While I don't aim to convert all statisticians to be controversy seekers, my goal is not only to educate, but also to encourage each of you to find your own unique way to lift your scepter.

Biography

Donald Berry holds the Frank T. McGraw Memorial Chair for Cancer Research at The University of Texas M. D. Anderson Cancer Center, where he is Head of the Division of Quantitative Sciences and Chairman of the Department of Biostatistics. In addition, he serves as the faculty statistician on the Breast Cancer Committee of the Cancer and Leukemia Group B (CALGB), a national oncology group. Through Berry Consultants, LLC he has consulted with many pharmaceutical and medical device companies on clinical trial design and analysis issues. He is well known as a developer of adaptive designs that minimize sample size while having a greater chance of detecting true signals of drug activity, efficiently using information that accrues over the course of the trial. He is also co-developer (with Giovanni Parmigiani) of BRCAPRO, a widely used program that provides individuals' probabilities of carrying mutations of breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. Dr. Berry received his Ph.D. in statistics from Yale University, and previously served on the faculty at the University of Minnesota and at Duke University, where he held the Edger Thompson Professorship in the College of Arts and Sciences. Dr. Berry is the author of more than 250 published articles as well as several books on biostatistics. In the last two years he has had first-authored publications in the New England Journal of Medicine, the Journal of the American Medical Association, and Nature Reviews Drug Discovery, and two senior-authored articles in the New England Journal of Medicine. Dr. Berry has been the principal investigator for numerous medical research programs funded by the National Institutes of Health, the National Cancer Institute, and the National Science Foundation. He is a Fellow of the American Statistical Association and of the Institute of Mathematical Statistics.

IMS Medallion Lecturer

Mary Sara McPeek, Ph.D.

Invited Address: Statistical Challenges in Genetic Association Studies

Common diseases such as asthma, diabetes, and hypertension, which currently account for a large portion of the health care burden, are complex in the sense that they are influenced by many factors, both environmental and genetic. One fundamental problem of interest is to understand what the genetic risk factors are that predispose some people to get a particular complex disease. Technological advances have made it feasible to perform case-control association studies on a genome-wide basis. The observations in these studies can have several sources of dependence, including correlation in the genotypes of nearby markers on a chromosome as well as relatedness of the individuals in the study. How to model the effects of this dependence and how to appropriately take it into account in the analysis of genome-wide association studies present interesting statistical challenges, which will be discussed in this talk along with proposed solutions.

Biography

Mary Sara McPeek is Professor of Statistics, Professor of Human Genetics, Member of the Committee on Genetics, and Senior Fellow in the Computation Institute at the University of Chicago. She received a joint bachelor's degree in Mathematics and master's degree in Statistics from Harvard University. She then earned a Ph.D. in Statistics from the University of California at Berkeley, under the guidance of Professor Terry Speed. While on the faculty in the Department of Statistics at the University of Chicago, she has held visiting positions in the Department of Mathematics at the University of Southern California and the Institute of Pure and Applied Mathematics at UCLA, and has held the Edward Rotan Visiting Professorship at the University of Texas M.D. Anderson Cancer Center. Her other awards include the NIH FIRST Award and the Evelyn Fix Memorial Medal and Citation. Dr. McPeek has served as Associate Editor of Biometrics and currently serves as Associate Editor of Genetics, Statistica Sinica, and Statistical Applications in Genetics and Molecular Biology. She is a member of the Institute of Mathematical Statistics, the International Biometrics Society (ENAR), the American Statistical Association, the American Society of Human Genetics, and the International Genetic Epidemiology Society. She is a reviewer for the National Institutes of Health and the National Science Foundation.

Short Courses

DATE: Sunday, March 16, 2008

Full Day Fee

Members	$250 ($275 after 2/15)
Nonmembers	$270 ($320 after 2/15)

Half Day Fee

Members	$160 ($185 after 2/15)
Nonmembers	$200 ($225 after 2/15)

SC1: Statistical Methods for Genome-wide Association Analysis (Full Day 8:30 A.M. – 5:00 P.M.)

Instructors:
Steven Chanock, Nilanjan Chatterjee and Kevin Jacobs, National Cancer Institute; Peter Kraft, Harvard University

Description:
Identification of large numbers of single nucleotide polymorphisms (SNP) across the human genome and development of technologies for massively multiplex genotyping has made genome wide association studies (GWAS) feasible. A number of GWAS have now provided the "proof of principle" behind the approach and have discovered novel genetic variants associated with complex diseases, such as diabetes, heart disease and cancers.

Analysis of GWAS involving hundreds of thousands of genetic markers poses some unique challenges for statistical researchers. This proposed short course will present the current state of art. The first half of the course will cover basic materials on quality control issues such as genotyping errors and call rates for different genotyping platforms, multi-stage study designs for reducing genotyping cost, principles behind designing replication studies, alternative methods for single-SNP analysis, method for dealing with population stratification and criterion for assessing statistical significance in very high dimensional hypothesis testing problems. The second half of the course will cover more advanced topics, such as methods for fine mapping, including haplotype analysis and methods for exploring gene-gene and gene-environment interactions.

The course will be taught by a core of researchers who has gained rich experience in statistical, genetic, epidemiologic and bioinformatic issues related to GWAS by their active involvement in the design and analysis of the ongoing Cancer Genetics Markers of Susceptibility (CGEMS) project, (http://cgems.cancer.gov) a NCI enterprise initiative for conducting whole genome association studies for prostate and breast cancers. The short course will use data from the CGEMS study for illustrations. A unique feature of the CGEMS study is that both results and data from this study are made publicly available in a continual basis. Thus participants will be able take advantage of their familiarity with the CGEMS data for further training and research purpose.

Pre-requisites:
Participants are expected to have basic knowledge of statistical inference such as hypothesis testing, analysis of contingency tables, regression modeling and Bayesian inference. Participants are also expected to have some familiarity with basic terminologies (e.g. alleles, genotypes, haplotypes, SNPs, linkage disequilibrium etc.) and concepts (e.g. independent assortment, Hardy-Weinberg law, population stratification etc) of population genetics.

Short Courses

SC2: Measurement Error in Nonlinear Models
(Full Day 8:30 A.M. – 5:00 P.M.)

Instructors:
David Ruppert, Cornell University; Ciprian Crainiceanu, Johns Hopkins University; Raymond J. Carroll, Texas A&M University

Description:
Measurement errors, if ignored, can lead to seriously mistaken scientific conclusions. The simplest case is attenuation where the estimated effect of a single error-prone covariate is biased toward a null effect, but more complex and insidious types of biases are possible, especially when there are multiple covariates. Over the past twenty or more years, many statistical methods have been developed for correcting these biases. This course surveys this methodology, with careful attention paid to the assumptions behind the different techniques. Special attention will be given to the practical implementation and available software for measurement error methods that can be applied to nonlinear models, such as logistic regression. We illustrate these techniques with a number of important epidemiological studies where exposures, e.g., to nutrients or radiation in the environment, are measured with sizeable error: Atherosclerosis Risk in Communities (ARIC), Choices for Healthy Outcomes in Caring for ESRD (CHOICE), Observing Protein and Energy Nutrition (OPEN), National Health and Nutrition Examination Survey (NHANES), Nevada Test Site Thyroid Disease Study, and the Framingham Heart Study. The following is a lost of topics covered:

· Examples from nutrition, radiation and more
· Effects of Measurement Error in Linear Models
· Why We Need Special Methods For Measurement Errors
· Measurement Error Examples
· Structure of a Measurement Error Problem
· Classical Error Model in Linear Regression
· Non-differential Error
· Estimating Attenuation
· Berkson and classical measurement error models
· Functional, classical structural, and flexible structural modeling
· Regression calibration, the workhorse method for linear and logistic regression, with illustration in STATA
· Multiplicative error
· SIMEX (Simulation/Extrapolation) with illustration from STATA and R
· Likelihood Methods
· Bayesian Methods, with illustrations using WinBUGS and R2WinBUGS
· Survival analysis

This short course is based on the recently released second edition of the book Measurement Error in Nonlinear Models: A Modern Approach, Carroll RJ, Ruppert D, Stefanski LA, and Crainiceanu C. 2nd Edition, Chapman & Hall CRC Press, 2006, see http://www.stat.tamu.edu/~carroll/eiv.SecondEdition/.

Pre-requisites:
Familiarity with linear models and logistic regression is essential. Background in likelihood methods, applied Bayesian methods and survival analysis will be useful.

SC3 : Statistical Monitoring of Clinical Trials: A Unified Approach (Full Day 8:30 A.M. – 5:00 P.M.)

Instructors:
Michael Proschan, National Institute of Allergy and Infectious Diseases; Janet Wittes, Statistics Collaborative

Description:
Clinical trials are monitored periodically and stopped for efficacy if the treatment arm is shown superior to the control, stopped for safety if there are serious side effects or the treatment is shown worse than control, or stopped for futility if the treatment has no chance of being shown superior by the end of the trial. These different aspects of monitoring require different tools. For proving superiority, we must use boundaries that maintain the correct overall type 1 error rate by accounting for the multiple looks at the data. Boundaries can also be important in safety monitoring, but other factors come into play for safety, such as whether the adverse events were foreseen, and whether they occurred at a substantially higher rate than expected. For futility, the most important tools are conditional and unconditional power. Roughly speaking, these tell us whether we are likely to get a statistically significant result, and whether a null result will be meaningful. This course covers all three aspects of monitoring: efficacy, safety, and futility.

The course begins by showing that the same general framework can be applied to many different types of trials—those with continuous, dichotomous, and survival outcomes. One way to standardize in these disparate clinical trial settings is to compute the z-score. We will show that for large sample sizes, the joint distribution of z-scores over time is the same for the different trials, provided that we define time correctly. Therefore, the same boundaries can be used for the different types of trials. An alternative way to standardize uses the “B-value” B(t), a transformation of the z-score whose expectation is linear in time (t). The slope of the line is the expected z-score at the end of the trial, which is intimately connected with the power of the trial. Therefore, we can determine whether the observed treatment effect is better or worse than expected by comparing the B-values over time to their expected line. We will use this B-value formulation throughout the course.

The course is based on the book Statistical Monitoring of Clinical Trials: A Unified Approach by Proschan, Lan and Wittes (Springer, 2006). Topics include classical monitoring boundaries such as those of Haybittle-Peto, Pocock, and O’Brien-Fleming, the flexible error spending function approach of Lan-DeMets, inference following a group-sequential trial, monitoring for safety, and monitoring for futility using conditional and unconditional power. We will also illustrate how to use free software for monitoring clinical trials.

Pre-requisites:
Inference at the level of Hogg and Craig (1978) or Mood, Graybill, and Boes (1974), and some familiarity with survival methods such as the logrank test.

SC4: Drop-Out in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis
(Full Day 8:30 A.M. – 5:00 P.M.)

Instructors:
Michael Daniels, University of Florida; and Joe Hogan, Brown University

Description
This course provides a survey of modern model-based approaches to handling dropout in longitudinal studies, and illustrates the use of newly-developed methods for sensitivity analysis and incorporation of prior information. The emphasis is on Bayesian approaches but the models and methods discussed can be implemented in non-Bayesian settings as well. The course will begin with a brief review of models for longitudinal data and the basics of Bayesian inference. The second part of the course will focus on dropout. We will discuss formal classifications of the dropout mechanism and describe different classes of models to adjust for biases caused by dropout. We also will discuss the importance of model specification and the logistics of model fitting. Models for both `ignorable’ and `non-ignorable’ dropout will be covered. The final part of the course focuses on nonignorable dropout; we will describe and motivate principles that should guide assessment of sensitivity to missing data assumptions and appropriate use of prior information. These concepts will be demonstrated using three case studies. The WinBUGS software will used throughout the course to illustrate the concepts and models on real data examples.

The course will be based on the book, Daniels, M.J. and Hogan, J.W. (2007) Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall (CRC Press).

Pre-requisites
Attendees should have working knowledge of generalized linear models (e.g., McCullagh and Nelder, 1989) and statistical inference (e.g., Casella and Berger, 1990) at the masters level.

SC5: Modeling Covariance Structures in Mixed Models (Half Day 8:00 A.M. – 12:00 P.M.)

Instructors:
Linda J. Young and Ramon C. Littell
University of Florida

Description:
Mixed models, whether they are general linear mixed models or generalized linear mixed models, have an underlying covariance structure associated with the response variable. This covariance of the responses is generally partitioned into the covariance matrix G associated with the modeled random effects and the covariance structure of the residuals, R. Given a set of data, determining the appropriate structure for these covariance matrices is both important and challenging. This short course will focus on the process of modeling G and R using real data sets. Topics include a review of covariance structures available in SAS, using the estimated covariance matrix to guide in the choice of covariance structure, accounting for spatial variation in either covariates or errors, and incorporating radial smoothing in the analysis. Issues and challenges will be discussed.

Pre-requisites:
The attendees should have experience with general linear mixed models and generalized linear mixed models. Prior use of SAS’s PROC MIXED and/or PROC GLIMMIX is assumed.

Short Courses

SC6: Statistical Computing Using R (with Graphics)
(Half Day: 1:00 P.M. – 5:00 P.M.)

Instructor:
Peter Dalgaard, University of Copenhagen

Description:
The R language and environment (www.r-project.org) is an open source implementation of the S language. It provides a wide range of graphical and statistical techniques, and R packages are increasingly being used as the vehicle for development and dissemination of new methodology.

The short course gives an introduction to R, mostly focusing on its usage for basic biostatistical methodology, but with some illustrations of more advanced usage. The format of the course will be lectures describing the general ideas of the software, alternating with demonstration and detailed explanation of actual R code samples. The following is a list of topics to be covered:

- Basic R features: Vectors, factors, data frames
- Classical tests
- Linear and generalized linear models
- Working with the graphics toolkit
- Elementary R programming
- Nonlinear models
- Generic likelihood methods: mle( )
- Programming examples:
    - simple simulations
    - extracting information from model objects
    - nontrivial data manipulation tasks

Pre-requisites:
A basic knowledge of biostatistical methodology is assumed. Some prior experience with statistical software will be advantageous. If you are completely inexperienced with R, it is recommended to download and install R and go through the sample session contained in Appendix A of the manual "An Introduction to R" (included with the R distribution), see http://www.r-project.org/.

Tutorials

T1: Competing Risks – Some Events are More Important than Others
Date: Monday, March 17
Time: 8:30 A.M. - 10:15 A.M.
Instructor: Melania Pintilie, Ontario Cancer Institute

Description:
In the time to event analysis there is the possibility to observe more than one type of event. A competing risks situation appears when the observation of the event of interest is hindered by the occurrence of another type of event. In the presence of competing risks the probability of the event of interest cannot be estimated using the usual product-limit (Kaplan-Meier) method. Kalbfleisch and Prentice introduced a non-parametric method to estimate the probability of the event of interest, referred as the cumulative incidence function. To facilitate the understanding of these two methods the estimates using the cumulative incidence function will be compared with the estimates obtained from Kaplan-Meier method in theoretical framework as well as through examples. There are two types of hazard that can be modeled, each with its own interpretation. Cox proportional hazards model can be applied for one of the hazards while the second type is modeled using a partial likelihood introduced by Fine and Grey. Although some theoretical details will be given, this course will focus on applied issues. Examples will be shown, mostly drawn from cancer research. The methodology can easily be extended to other areas where competing risks are present. Details on sample size calculation will be given. The software used is R. This tutorial is based on the book Competing Risks –A Practical Perspective, which appeared in 2006 under Wiley's Statistics in Practice series.

Pre-requisites:
It is expected that the participants have basic knowledge of survival techniques. Working knowledge of R would be helpful.

T2: Introduction to Bayesian Analysis Using SAS Software
Date: Monday, March 17
Time: 10:30 A.M. - 12:15 P.M.
Instructor: Fang K. Chen, SAS Institute

Description:
Bayesian methods have become increasingly popular in recent years in a number of different disciplines. This tutorial provides an introduction to Bayesian methods with applications in the areas of the generalized linear model and survival analysis. The first part of the course provides an overview of Bayesian methodology, including motivation and Bayesian inference, as well as computational methods and convergence diagnostics relevant to the SAS implementation. The second part of the course discusses applications using new capabilities in SAS/STAT software in the GENMOD, LIFEREG and PHREG procedures which are based on Gibbs sampling. Examples will include methods such as linear regression, logistic regression, Poisson regression, Cox regression, parametric survival models, and the piecewise exponential model. Note that these enhanced procedures are available as downloads for the 9.1.3 release of SAS software.

Pre-requisites:
A master's level knowledge of statistics is assumed as well as experience with generalized linear models and survival analysis. Familiarity with basic use of the SAS DATA step and SAS statistical procedures is also assumed. Previous exposure to Bayesian methods is useful but not required.

T3: Statistical Analysis of Cost Effectiveness Data
Date: Monday, March 17
Time: 1:45 P.M. - 3:30 P.M.

Instructor: Andrew R. Willan, University of Toronto

Description:
It is becoming increasingly common in randomized controlled trials (RCTs) to collect patient-level cost data along with the measures of effectiveness (health effects). This added dimension to RCTs has motivated the development of new statistical methodology designed to answer questions of health policy in addition to those of purely clinical importance. Initially attention was focussed on estimating the incremental cost-effectiveness ratio (ICER) and calculating the appropriate confidence intervals. The ICER is defined as, where, the difference in mean cost between a new intervention (denoted as T) and standard intervention (denoted as S) and the difference in mean effectiveness between T and S. More recently, due to the acknowledged problems associated with ratio statistics, attention has shifted to making inference about incremental net benefit (INB), defined as, using test of hypothesis and confidence intervals. Incremental net benefit can also be defined in units of effectiveness and referred to as the incremental net health benefit (INHB), defined as. Either definition requires the specification of the willingness-to-pay (WTP) for a unit of effectiveness (denoted as), or at the very least, the analysis must be presented as a function of so that readers can apply the WTP most appropriate for them.

Regardless of whether an ICER or an INB approach is taken the parameters, and the corresponding variances and covariance must be estimated. The procedures used for parameter estimation will depend on (i) whether or not the skewed nature of cost data is accounted for, (ii) whether of not covariates are included, (iii) whether or not random effects, such clinical site or country, are included, and (iv) whether or not the data are censored. The class will provide an overview of the procedures used for the sixteen possible scenarios defined by the four points listed above. In addition, the presentation of cost-effectiveness analyses will be covered, illustrating the connection between the ICER and INB approaches.

The class will cover many of the topics found in Willan AR and Briggs AH (2006) The Statistical Analysis of Cost-effectiveness Data. John Wiley & Sons, Chichester.

Pre-requisites:
A bachelor's degree in statistics or equivalent.

T4: Receiver Operating Characteristic (ROC) Curves for the Uninitiated
Date: Tuesday, March 18
Time: 8:30 A.M. - 10:15 A.M.
Instructor: Mithat Gönen, Memorial Sloan-Kettering Cancer Center

Description:
Receiver Operating Characteristic (ROC) Curves are commonly used in medical diagnostics and, increasingly, in predictive modeling. Despite their increasing popularity in practice, commonly-used statistical packages offer very few standard options to analyze ROC curves. This is a cause of frustration for many biostatisticians who are trying to add ROC curves to their analysis toolkit. This tutorial will present the basics of ROC curves paying close attention to the capabilities of three popular software packages SAS, R and STATA.

The starting point for ROC curves is the definition of sensitivity and specificity. Building on these definitions, the empirical ROC curve will be introduced along with several summary indices such as the area under curve and optimal operating points. Binormal ROC curves will be discussed next, with emphasis on the similarities and distinctions between continuous and ordinal data. If time permits, regression methods for ROC curves will also be covered. Examples are drawn from weather forecasts, credit scoring and medical diagnosis. Attendees will receive complete computer code for all the examples.

Pre-requisites:
The only requirement is a general understanding of statistical principles at the level of a graduate student completing first year at a master's program in statistics or biostatistics. Although it is not essential, it will be very helpful to have data-analytic expertise in one of the three software packages mentioned above.

T5: Analysis of Censored Cost or Health Outcomes Data
Date: Tuesday, March 18
Time: 1:45 P.M. - 3:30 P.M.
Instructors: Hongwei Zhao, University of Rochester; Heejung Bang, Weill Medical College of Cornell University

Abstract:
Medical cost and quality-adjusted lifetime are two common health outcomes data from clinical trials and observational studies. Although these two types of data look seemingly different, they belong to a large class of variables, namely, "mark variable" in statistics (Huang and Louis, 1998) and all mark variables share many of common statistical and methodological properties and can be understood in a unified framework. Just like standard survival data, censoring is an important issue in mark variables accumulated over a long time. Despite the analogy, censoring mechanism in mark variables is considerably different from the traditional paradigm and it is called "informative censoring". It has been a decade since it was shown that use of most standard statistical techniques such as sample mean, t-test, ordinary least squares, Kaplan-Meier estimator, Log-rank test and Cox regression can be invalid (Zhao and Tsiatis 1997; Lin, Feuer, Etzioni & Wax 1997). However, we often find that even experienced researchers still use traditional methods for the analysis of these health outcome data in practice. Key methodological issues to be accounted for are how to handle informative censoring and how to achieve consistent and efficient estimation in various statistical contexts: estimation of mean, median and distribution, two-sample test, regression, and cost-effectiveness analysis.

In this tutorial, we will review novel methods that will provide valid statistical estimation and inference for health outcomes data that have been developed for last 10 years. Unfortunately, not all are easy and user-friendly, and no commercial software is available so far. Therefore, we will suggest some methods as practical solutions for practitioners and health researchers. We will also present the analytic relationships among well known medical cost estimators recently identified (Zhao, Bang, Wang & Pfeifer 2007). An extended concept and application to customer lifetime value, a popularly used metric in marking and business, will be discussed.

Pre-requisites:
Basic knowledge of survival analysis.

Roundtables

R1: Barriers to Producing User Friendly Software for Cutting Edge Methodology
Discussion Leaders:
Cyrus Mehta, Cytel and Ram Tiwari, National Cancer Institute

Important new methodological advances in statistical science rely increasingly on the availablility of reliable, robust and well documented software in order to be useful to the scientific community. It is frequently the case, however, that the developers of these innovative new statistical methods are unable to provide accompanying software that is suitable for general use. The software that they have developed in the course of their research was for their own use only. Thus it works well for the few cases that they have considered, but is often poorly documented, limited in scope, and untested for the general range of problems others are likely to encounter. Not surprisingly the innovative methods that were developed do not get widespread use. This roundtable discussion is intended for researchers who are interested in making a serious commitment to complementing their mathematical contributions with software that meets higher standards than are required for personal use. Is there a way for researchers to develop software that can be used by the broader scientific community without having to meet commercial standards of reliability, robustness and documentation? We wish to hold a roundtable discussion on questions like this, and hear from roundtable participants on what barriers exist (e.g., lack of funds, lack of technical support) and what could do to overcome them.

R2: Careers in a Medical Center Environment
Discussion Leader:
Avital Cnaan, The Children's Hospital of Philadelphia and University of Pennsylvania School of Medicine

Careers in a medical center environment are challenging, rewarding, and very diverse. The areas of application are diverse, from premature infants to aging studies, from statistics of laboratory science, with challenging problems such as various "omics" applications, to large patient record databases with problems of matching records and statistical algorithms to address those, to longitudinal studies with informative patterns of missingness, to studies of rare diseases and small sample problems. The statistician in the medical center contributes to society in an immediate sense when research gets published and affects health care. The statistical methodology problems for the biostatistician to further develop arise naturally in the context of the scientific/medical questions being addressed. From a career perspective, medical centers tend to need statistical staff from entry level statistical programmers to collaborators at the professor level, and thus they provide growth opportunities for young biostatisticians. Opportunities and challenges to biostatisticians in a medical center environment will be discussed.

R3: Writing NIH Statistical Methodology Grants and the NIH Grant Review Process
Discussion Leader:
Marie Davidian, North Carolina State University

Writing grant applications for external funding to support one's methodological research is an important professional growth activity for junior and mid-level researchers, and success in securing such funding provides researchers of all (professional) ages needed resources and peer recognition. This roundtable luncheon will focus on strategies for writing a high-quality NIH grant application to support statistical methodological research and the process by which these grants are submitted and reviewed. Procedures and review criteria used by the NIH panel that reviews many of the methodological grants submitted, the Biostatistical Methods and Research Design (BMRD) study section, will be discussed.

R4: How I Publish without Perishing
Discussion Leader:
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health

During this roundtable we will discuss a wide range of topics related to publishing in statistical journals, indeed all journals. Topic will include identifying your audience, choosing a target journal, structuring your manuscript, principal components of effective writing, showing respect for the journal and the referees, and dealing with journal decisions. Participants should be ready to discuss their successes and frustrations. All will leave the roundtable well nourished with improved understanding on how to maximize publication success and to use rejections to produce subsequent success.

R5: Co-Development of Biomarkers and Treatments
Discussion Leader:
Estelle Russek-Cohen, U.S. Food and Drug Administration

In trying to move personalized medicine forward, there is a need for creative designs to evaluate biomarkers and decide when the results of biomarker evaluation can improve the selection of treatments for individual patients. When one of the treatments under consideration is novel and the biomarker is also novel, the challenges are even greater. Studies with longer term outcomes such as overall survival may call for a different strategy than studies that provide more immediate feedback. Since there is no one design that can fit all cases, in this roundtable we will discuss creative solutions to design problems in the co-development of biomarkers and treatments.

Roundtables

R6: Data Monitoring Committees: Current Issues

Discussion Leader:
Susan Ellenberg, University of Pennsylvania

Data Monitoring Committees (DMCs) or Data and Safety Monitoring Boards (DSMBs) are increasingly employed in clinical trials sponsored both by industry and government. The role of these committees may vary greatly, depending on the sponsor and the type of trial. The purpose of this roundtable is to discuss the various models that have arisen for DMC operation in different settings, and the advantages and disadvantages of different approaches. Some specific issues that might be addressed: the role of DMCs in early phase trials; the role of sponsors and study investigators in the DMC process; conflict of interest considerations for DMC members; the role of DMC members in the reporting of trial results.

R7: On Creating and Operating an On-Campus Statistical Consulting Center

Discussion Leaders:
Randall Reiger and Bob Gallop, West Chester University

Having recently created a statistical consulting service on-campus, the facilitators of this roundtable would like to share their experiences and swap ideas with others about such an endeavor. Among other issues, we will discuss the following:

- How to get started on creating a statistical consulting center on campus.
- How to get faculty and students involved.
- Ideas for financial infrastructure of consulting center: fees, grants, and service.
- Defining a mission.
- Involving faculty from across campus.
- How to promote services.
- How to create ties with outside companies.

We hope that the discussion and exchange of ideas will serve to enhance the delivery of such a service on campus, and provide advancement in the efficacy of such services.

R8: Should Statistical Software Certification Programs be Integrated into the Academic Classroom Environment?

Discussion Leaders:
Monica Jackson and Jun Lu, American University

Many graduate programs require courses in statistical software. However, offering software certification programs as part of these courses remains controversial. Employers in the IT industry like to hire students with validated skills from passing software certification exams. However, some statisticians in academia feel that it is not appropriate to offer industry certifications in the classroom since it promotes a company's product. There are also issues as to who should pay for the certification—the students or the university? In the Fall of 2007, the discussion leaders experimented with the idea of incorporating the SAS certificate program in the statistical software course at American University. The goal of this curriculum development study was to include the certification exam into classroom teaching environment. The approach was received with mixed reviews. What is the best way to effectively prepare students so that statistical software courses can benefit them in their future career and the academic community? This roundtable will discuss alternative approaches for teaching statistical software as well as debate if industrial certificate exams should be a part of statistics education in academia. We welcome statisticians from industry and academia to join in this discussion.

R9: Development and Uses of Microsimulation Models in Health Care Policy Issues

Discussion Leader:
Carolyn Rutter, Group Health Cooperative

Microsimulation models are characterized by simulation of individual event histories for an idealized population or cohort of interest. They have been increasingly used to inform health care policy issues in terms of extrapolating the results of RCT's to different settings (especially screening RCT's), cost effectiveness, and understanding the impact of disease control interventions (prevention, screening, and treatment) on national trends. At this roundtable, we will discuss challenges to the credibility of this type of modeling, and strategies to overcome these challenges. Depending on the interests at the table, various aspects of microsimulation modeling would be discussed, including strategies for model building, parameter calibration, model validation, model application/interpretation, and presentation of model results in the health care policy forum. The organizers of this roundtable have developed a natural history colorectal cancer policy model, and has been involved in a cooperative group of cancer modelers who have employed a comparative modeling approach.

R10: Careers at the National Institutes of Health

Discussion Leader:
Dean Follman, National Institute of Allergy and Infectious Diseases

Broadly speaking, statisticians at NIH have four main functions: conducting research on statistical methodology, participation and oversight of large collaborative medical studies, small-scale collaborations with individual medical researchers, and promoting opportunities for research in the extramural statistical community. Problems are varied and important. Examples include development of HIV vaccines, estimating trends in cancer rates, design and analysis of bone marrow transplantation studies, and statistical methods for recurrent events. Numerous opportunities exist to collaborate with first-rate medical professionals on their studies and conduct research on statistical methodology usually inspired by these studies. NIH is a natural home for research statisticians who enjoy developing new methods based on important applied problems. In this roundtable we will discuss the various opportunities for statisticians at NIH, and practical issues of how to make contacts and learn about and apply for positions.

R11: Climate Change Investigations

Discussion Leader:
Timothy G. Gregoire, Yale University

Climate change is a "hot" topic in the popular and even scientific press. It has been the focus of a few sessions at recent JSMs but heretofore has been a rather quiet issue within the statistics community. This roundtable is being proposed to provide a venue for discussion among those conducting research that directly or indirectly deals with climate-change issues. It is open, also, to those who simply have an interest in climate change and wish to learn more about current statistical investigations into it. I can describe my own work on the tundra of the north slope of Alaska, and others will be welcomed to describe their current and past research that bears on this issue, too.

R12: Using Prior Information or Adaptive Designs in Bayesian Clinical Trials: The Medical Device Experience

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

The FDA's Center for Devices and Radiological Health has played a pioneering role among regulatory bodies in its encouragement of companies who would like to propose studies with Bayesian designs or analyses. In May of 2005 FDA issued a draft guidance on the use of Bayesian statistics in medical device clinical trials. One of the opportunities on the FDA Critical Path list is the Use of Prior Experience or Accumulated Information in Trial Design. Participants at this roundtable are welcome to bring examples of successful approaches to these novel designs, as well as their opinions of the advantages and disadvantages of Bayesian clinical trials.

R13: Writing Collaborative Grants

Discussion Leader:
Lisa M. LaVange, University of North Carolina at Chapel Hill

A statistician collaborating on a grant proposal has opportunities to contribute to the science of the proposed research and to learn about topics of interest in other fields of study. The role of the statistician should be established early in the collaboration and may be that of a staff or faculty member on the project team. An important aspect of the collaboration is that the statistician and other investigators understand the role as defined and the accompanying expectations in terms of level of involvement and voice in decision making. Beginning the collaboration early in the life of the project will help ensure that the statistician can provide the most benefit in terms of study design and analysis planning. This involvement often turns into an educational opportunity for the investigators. Involvement beyond the grant submission can be equally important, and clarification about what is expected in terms of data handling and statistical programming, in addition to providing statistical expertise and data analysis, is essential. Roundtable participants will discuss what has worked well and what has not in terms of building strong and effective collaborations during the grant writing stage, as well as tips for successful partnerships after award of the grant.

R14: Careers at the U.S. Food and Drug Administration

Discussion Leader:
Steve Wilson, U.S. Food and Drug Administration

The Food and Drug Administration, with five Centers to choose from (Biologics, Devices, Drugs, Food and Toxicological Research), is a great place for a career in the quantitative sciences. FDA statisticians work in a highly dynamic, scientific environment with a mission to make timely and complex decisions regarding issues that advance public health and serve to speed innovations. The challenges/opportunities are many: from adaptive clinical trial designs to risk assessment to Bayesian statistics to pharmaco*nomics to standards for electronic health records.

With so much to do and so many opportunities, how do you make the choices that are right for you? Participants in this round table will have the chance to learn about the FDA and to think about potential paths to successful careers with the Agency.

R15: Diagnostics for Mixed Models

Discussion Leader:
Geert Verbeke, Biostatistical Centre, K.U.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. Participants at this roundtable will discuss what is currently available and in which directions more work is needed.

Saturday, March 15

9:00 a.m. – 9:00 p.m.	**Workshop for Junior Researchers** *(Washington A)*
3:00 p.m. – 5:30 p.m.	**Conference Registration** *(Regency Foyer)*

Sunday, March 16

7:30 a.m.-6:30 p.m.	**Conference Registration** *(Regency Foyer)*
8:00 a.m.-12:00 p.m.	**Short Courses** SC5: Introduction Modeling Covariance Structures in Mixed Models *(Potomac 5/6)*
8:30 a.m.-5:00 p.m.	**Short Courses** SC1: Statistical Methods for Genome-wide Association Analysis *(Regency A/B)* SC2: Measurement Error in Nonlinear Models *(Regency C)* SC3: Statistical Monitoring of Clinical Trials: A Unified Approach *(Regency D)* SC4: Drop-Out in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis *(Regency 3/4)*
11:00 a.m.-4:00 p.m.	**Fostering Diversity Workshop** *(Washington A)*
1:00 p.m.-5:00 p.m.	**Short Course** SC6: Statistical Computing Using R (with Graphics) *(Potomac 5/6)*
3:00 p.m.-5:00 p.m.	**Exhibits Open** *(Regency Foyer)*
4:30 p.m.-6:30 p.m.	**ENAR Executive Committee Meeting (Closed)** *(Kennedy Room)*
4:30 p.m.-6:30 p.m.	**Placement Service Opens** *(Jefferson Room)*
7:30 p.m.-8:00 p.m.	**New Member Reception** *(Regency Ballroom)*
8:00 p.m.-11:00 p.m.	**Social Mixer and Poster Presentation Sessions** *(Regency Ballroom)* 1. Posters: Epidemiologic Methods and Survey Research 2. Posters: Longitudinal and Categorical Data 3. Posters: Survival Analysis 4. Posters: Clinical Trials: General 5. Posters: Clinical Trials: Adaptive Design and Adaptive Randomization 6. Posters: Applied Data Analysis 7. Posters: Latent Variables and Missing Data 8. Posters: Statistical Models and Methods 9. Posters: Statistical Genetics 10. Posters: Genomics and Proteomics 11. Posters: Microarray Analysis

Monday, March 17

7:30 a.m.-8:30 a.m.	**Student Breakfast** *(Tidewater Room)*
7:30 a.m.-5:00 p.m.	**Conference Registration** *(Regency Foyer)*
7:30 a.m.-5:00 p.m.	**Speaker Ready Room** *(Prince William Room)*
9:30 a.m.-5:00 p.m.	**Placement Service** *(Jefferson Room)*
8:30 a.m.-5:30 p.m.	**Exhibits Open** *(Regency Foyer)*
8:30 a.m.-10:15 a.m.	**Tutorial** **T1: Competing Risks – Some Events are More Important than Others** *(Regency E)*

Monday, March 17
continued

Scientific Program
12. Statistical Monitoring of Clinical Trials: Where the Rubber Meets the Road *(Regency A)*
13. Statistical Issues in Genomic Studies in Population Sciences *(Regency B)*
14. Hierarchical Modeling in Environmental Exposure and Toxicological Risk Assessment *(Regency C)*
15. Model Validation and Model Selection in Longitudinal Data Analysis *(Regency D)*
16. New Statistical Methods in Diagnostic Medicine *(Washington A)*
17. Regularization and Hierarchical Modeling: Two Philosophies for Variable Selection *(Washington B)*
18. Contributed Papers: Methods for Variable Selection and Model Building *(Potomac 1)*
19. Contributed Papers: Research Methods *(Potomac 2)*
20. Contributed Papers: Genomics and Microarray Analyses *(Potomac 3)*
21. Contributed Papers: Epidemiologic Methods *(Potomac 4)*
22. Contributed Papers: Bayesian Methods *(Potomac 5)*
23. Contributed Papers: Latent variables and Structural Equations Modeling *(Potomac 6)*

10:15 a.m.—10:30 a.m. **Refreshment Break and Visit the Exhibitors** *(Regency Foyer)*

10:30 a.m.—12:15 p.m. **T2: Introduction to Bayesian Analysis Using SAS Software** *(Regency E)*
Scientific Program
24. Non-inferiority Clinical Trials - Fixed Margin or No Fixed Margin *(Regency A)*
25. Model Selection for High Dimensional Data with Practical Solutions *(Regency B)*
26. Spatial and Spatio-Temporal Latent Structure Modeling *(Regency C)*
27. New Advances in Survival Analysis for Large Dimensional Biomedical Data *(Regency D)*
28. On the Utility of Deterministic Models for Causal Effects *(Washington A)*
29. Estimating the Distribution of Usual Intakes of Nutrients and Foods, and Relating Usual Intake to Health Parameters in a National Health Survey *(Washingon B)*
30. Contributed Papers: Multiple Testing *(Potomac 1)*
31. Contributed Papers: Biomarkers *(Potomac 2)*
32. Contributed Papers: Joint Models- Survival and Longitudinal *(Potomac 3)*
33. Contributed Papers: Data Mining and Machine Learning *(Potomac 4)*
34. Contributed Papers: Genetic Epidemiology- Associations *(Potomac 5)*
35. Contributed Papers: Generalized Linear Models *(Potomac 6)*

12:15 p.m.—1:30 p.m. **Roundtable Luncheons (Registration Required)** *(Tidewater Room)*

12:30 p.m.—4:30 p.m. **Regional Advisory Board (RAB) Luncheon Meeting (By Invitation Only)** *(Arlington Room)*

1:45 p.m.—3:30 p.m. **Tutorial**
T3: Statistical Analysis of Cost Effectiveness Data *(Regency E)*
Scientific Program
36. Statistical Issues in Analysis of Genomics Data *(Regency A)*
37. Recent Developments in Non-smooth Estimating Functions for Censored Data *(Regency B)*
38. Bayesian Methods in Epidemiology *(Regency C)*
39. Longitudinal Data Analysis in the Presence of Mortality *(Regency D)*
40. Improving Measurement in Profiling Providers of Healthcare *(Washington A)*
41. Bayesian Variable Selection with High Dimensional Covariate Data *(Washington B)*
42. Contributed Papers: Clustered and Hierarchical Models *(Potomac 1)*
43. Contributed Papers: Statistical Genetics *(Potomac 2)*
44. Contributed Papers: Clinical Trial Adaptive Design and Randomization *(Potomac 3)*
45. Contributed Papers: Genome-Wide Association Studies *(Potomac 4)*
46. Contributed Papers: Spatial Modeling Applications *(Potomac 5)*
47. Contributed Papers: Missing or Incomplete Data *(Potomac 6)*

3:30 p.m.—3:45 p.m. **Refreshment Break and Visit the Exhibitors** *(Regency Foyer)*

3:45 p.m.—5:30 p.m. **Scientific Program**
48. FDA Advisory Committees: A Statistician’s Role in Deciding Public Policy (A Panel Discussion) *(Regency A)*
49. New Directions in Safety Planning and Analysis for Clinical Development *(Regency B)*
50. Validation of Genomic Classifiers *(Regency C)*

51. Methods for Analyzing Multi-reader Receiver Operating Characteristic (ROC) Data *(Regency D)*
52. Recent Advances in the Modeling of Competing Risks *(Washington A)*
53. Analysis of Marked Point Patterns with Spatial and Nonspatial Covariate Information *(Washington B)*
54. Contributed Papers: Statistical Inference *(Potomac 1)*
55. Contributed Papers: Statistical Methods- General *(Potomac 2)*
56. Contributed Papers: Clinical Trials Design *(Potomac 3)*
57. Contributed Papers: Statistical Methods for Genomics *(Potomac 4)*
58. Contributed Papers: Survival Analysis - Nonparametric Methods *(Potomac 5)*
59. Contributed Papers: Functional Data Analysis *(Potomac 6)*

6:00 p.m.—7:30 p.m. **President's Reception (By Invitation Only)** *(Tidewater Room)*

Tuesday, March 18

7:30 a.m.—5:00 p.m. **Conference Registration** *(Regency Foyer)*

7:30 a.m. – 5:00 p.m. **Speaker Ready Room** *(Prince William Room)*

9:30 a.m.—3:30 p.m. **Placement Service** *(Jefferson Room)*

8:30 a.m.—5:30 p.m. **Exhibits Open** *(Regency Foyer)*

8:30 a.m.—10:15 a.m. **Tutorial**
T4: Receiver Operating Characteristic (ROC) Curves for the Uninitiated *(Conference Theater)*
Scientific Program
60.Adaptive Designs: Perspectives from Academia, Industry and Regulatory *(Regency A)*
61. The Education of Clinical Trial Statisticians: Do We Need to Change with the Times? *(Regency B)*
62. Critical Design and Statistical Considerations with Use of (Genomic) Biomarker for Prediction or Optimizing Therapy *(Regency C)*
63. Advanced Semiparametric Modeling of Genetics/Genomic Data *(Regency D)*
64. The Use of Survey Samples in Epidemiological Follow-Up Studies *(Washington A)*
65. Analysis of Recurrent Marker Data *(Washington B)*
66. Contributed Papers: Joint Modeling *(Potomac 1)*
67. Contributed Papers: Applications in Causal Inference *(Potomac 2)*
68. Contributed Papers: Proteomics and Genomics *(Potomac 3)*
69. Contributed Papers: Sample Size *(Potomac 4)*
70. Contributed Papers: Imaging Analysis *(Potomac 5)*
71. Contributed Papers: Survival Analysis - Dependent Censoring *(Potomac 6)*

10:15 a.m.—10:30 a.m. **Refreshment Break and Visit the Exhibitors** *(Regency Foyer)*

10:30 a.m.—12:15 p.m. **72. Presidential Invited Address** *(Regency Ballroom)*

12:30 p.m.—4:30 p.m. **Regional Committee Luncheon Meeting (By Invitation Only)** *(Arlington Room)*

1:45 p.m.—3:30 p.m. **Tutorial**
T5: Analysis of Censored Cost or Health Outcomes Data *(Conference Theater)*
Scientific Program
73. Joint Models for Surrogate and Final Outcomes *(Regency A)*
74. Statistical Analysis of Large-Scale Environmental Datasets *(Regency B)*
75. Variable Selection and Dimension Reduction in Genomics *(Regency C)*
76. Targeting Treatment Efficacy: Flexible Modeling to Clinical Trial Design *(Regency D)*
77. Statistical Learning in Biological Signals and Images *(Washington A)*
78. Nonparametric Regression for Survival Analysis *(Washington B)*
79. Contributed Papers: CGH Array and Copy Number *(Potomac 1)*
80. Contributed Papers: Biopharmaceutical Studies *(Potomac 2)*
81. Contributed Papers: Power Analysis *(Potomac 3)*
82. Contributed Papers: Survival Analysis - Cure Rate Models and Recurrent Events *(Potomac 4)*
83. Contributed Papers: Experimental Design *(Potomac 5)*
84. Contributed Papers: ROC/ Diagnostic Methods *(Potomac 6)*

3:30 p.m.—3:45 p.m. **Refreshment Break and Visit the Exhibitors**

Tuesday, March 18
continued

p.m. **Scientific Program**
85. IMS Medallion Lecture Statistical Challenges in Genetic Association Studies *(Regency A)*
86. On Futility Calculations In Randomized Clinical Trials *(Regency B)*
87. Statistical Methods for Detecting Copy Number Variation *(Regency C)*
88. Challenges in Large, Simple Trials *(Regency D)*
89. Multistate Approach Towards Modeling Complex Biomedical Data *(Washington A)*
90. Informing Health Policy Decisions: Best Test Strategies for Colorectal Cancer Screening *(Washington B)*
91. Contributed Papers: Quantitative Trait Loci Mapping *(Potomac 1)*
92. Contributed Papers: Applied Data Analysis *(Potomac 2)*
93. Contributed Papers: Longitudinal Models- Discrete *(Potomac 3)*
94. Contributed Papers: Survival Analysis - Competing Risks *(Potomac 4)*
95. Contributed Papers: Clinical Trials – General *(Potomac 5)*
96. Contributed Papers: Methods in Spatial Modeling *(Potomac 6)*

5:30 p.m.—6:30 p.m. **ENAR Business Meeting (Open to all ENAR members)** *(Potomac 1)*

6:30 p.m.—9:30 p.m. **Tuesday Night Event - Dinner at La Tasca (Registration Required)**
(Buses begin departing from the hotel at 6:15 p.m.)

Wednesday, March 19

7:30 a.m.—9:00 a.m. **Planning Committee Breakfast Meeting (By Invitation Only)** *(Kennedy Room)*

8:00 a.m.—12:30 p.m. **Conference Registration** *(Regency Foyer)*

8:00 a.m. – 12:00 noon **Speaker Ready Room** *(Prince William Room)*

8:00 a.m.—12:00 p.m. **Exhibits Open** *(Regency Foyer)*

8:30 a.m.—10:15 a.m. **Scientific Program**
97. Collaboration in HIV Research: Two Case Studies *(Regency A)*
98. Joint Modeling Approaches for Longitudinal Data under Complex Study Designs *(Regency B)*
99. Multi-task Learning for Borrowing Information from Disparate Data Sources *(Regency C)*
100. Dynamic Treatment Regimes: Practice and Theory *(Regency D)*
101. Text Data Mining *(Washington A)*
102. Nonparametric Bayes: The Practical use for Genomic Data *(Washington B)*
103. Contributed Papers: Methods in Causal Inference *(Potomac 1)*
104. Contributed Papers: Bioassay and Cancer Applications *(Potomac 2)*
105. Contributed Papers: Missing Data and Imputation *(Potomac 3)*
106. Contributed Papers: Longitudinal Models- Continuous *(Potomac 4)*
107. Contributed Papers: Bayesian and Multi-Level Survival Analysis *(Potomac 5)*
108. Contributed Papeers: Multiple Testing in Genomics *(Potomac 6)*

10:15 a.m.—10:30 a.m. **Refreshment Break and Visit the Exhibitors** *(Regency Foyer)*

10:30 a.m.—12:15 p.m. **Scientific Program**
109. Advances in the Design and Analysis of Randomized Clinical Trials *(Regency A)*
110. Recent Advances in Graphs/Graphical Models for Genetic Network Analysis *(Regency B)*
111. New Statistical Methods for Biomedical Imaging Data *(Regency C)*
112. Statistical Challenges in Genomewide Association Studies *(Regency D)*
113. Recent Advances in Analyzing Biomarker Data with Limits of Detection *(Washington A)*
114. Biological Applications of Machine Learning *(Washington B)*
115. Contributed Papers: Health Services and Policy Research *(Potomac 1)*
116. Contributed Papers: Measurement Error *(Potomac 2)*
117. Contributed Papers: Microarray Data Analysis *(Potomac 3)*
118. Contributed Papers: Survival Analysis Methods and Applications *(Potomac 4)*
119. Contributed Papers: Variable Selection and Model Building – Applications *(Potomac 5)*
120. Contributed Papers: Nonparametric Methods *(Potomac 6)*

Scientific Program

Poster Presentations

Sunday March 16

7:30 p.m. – 8:00 p.m. New Member Reception
(Regency Ballroom)

8:00 p.m. – 11:00 p.m. *(Regency Ballroom)*

1. POSTERS: EPIDEMIOLOGIC METHODS AND SURVEY RESEARCH
Sponsors: ENAR, ASA Section on Statistics in Epidemiology, ASA Section on Risk Analysis and ASA Survey Research Methods Section

1a. Extreme Verification Bias in Paired Continuous Tests May Mask the Magnitude of the Difference Between the Diagnostic Accuracies of Screening Modalities
Deborah H. Glueck*, Molly M. Lamb, Colin O'Donnell, University of Colorado, Keith E. Muller, University of Florida, John M. Lewin, Diversified Radiology of Colorado

1b. Case-Control Study of Lung-Cancer Risk from Residential Radon Exposure in Worcester County, Massachusetts
Richard E. Thompson*, Johns Hopkins University, Donald F. Nelson, Worcester Polytechnic Institute, Joel H. Popkin and Zenaida Popkin, St. Vincent Hospital and Fallon Clinic, Worcester Medical Center

1c. Effect of BMI on Lifetime Risk of Diabetes for U.S. Adults
James P. Boyle*, Centers for Disease Control and Prevention

1d. Variable Selection for Multiply-Imputed Large Data Set in Dioxin Exposure Study
Qixuan Chen*, James M. Lepkowski, Brenda Gillespie, and David H. Garabrant, University of Michigan

1e. A Latent Variable Model for an Epidemiological Study with Multiple Correlated Exposures
Abbie Stokes-Riner* and Sally W. Thurston, University of Rochester, Jason Roy, Geisinger Center for Health Research

1f. Parameter Estimation on a Partially Grouped Sample
Sergey Tarima*, Medical College of Wisconsin

1g. Bayesian Hierarchical Methods for Small Area Estimation: Diabetes Prevalence by U.S. Counties
Theodore J. Thompson*, Betsy L. Gunnels and James P. Boyle, Centers for Disease Control and Prevention

1h. Shift in Age Distribution of Pertussis Infant Mortality to Very Young Infants, United States, 1933-2004
Andrew L. Baughman*, Tracy Pondo and Margaret M. Cortese, Centers for Disease Control and Prevention

2. POSTERS: LONGITUDINAL AND CATEGORICAL DATA
Sponsors: ENAR, ASA Section on Statistics in Epidemiology and ASA Biometrics Section

2a. Process Modeling for Ordered Categorical Data
Simone Gray* and Alan Gelfand, Duke University

2b. Tests for Zero Inflation in a Bivariate Zero-Inflated Poisson Model
Byoung Cheol Jung*, University of Seoul, JungBok Lee and Seo Hoon Jin, Korea University

2c. Estimation of Causal Effects in Longitudinal Observational Studies
Elizabeth Johnson*, Constantine Frangakis and Scott L. Zeger, Johns Hopkins University

2d. Selecting the Working Correlation Structure By a New Generalized AIC Index for Longitudinal Data
Jiawei Liu*, Georgia State University, Bruce G. Lindsay, Pennsylvania State University, Wei-Lun Lin, Georgia State University

2e. Diagnostics for the Covariance Structure of Nonlinear Mixed Effects Models for Balanced Longitudinal Data
Karen Chiswell*, GlaxoSmithKline, John F. Monahan, North Carolina State University

2f. Estimating Sample Size for Misclassified Binary Response With Covariate Measurement Error
Dunlei Cheng*, Institute for Health Care Research and Improvement, Baylor Health Care System

2g. Spatial and Functional Modeling of Variablity in Tooth-Level Periodontal Outcomes
Thomas M. Braun*, University of Michigan

3. POSTERS: SURVIVAL ANALYSIS
Sponsors: ASA Health Policy Statistics Section, ASA Biometrics Section and ASA Biopharmaceutical Section

3a. A Nonlinear Modeling Framework for Analyzing Time-to-event Outcomes From Complex Multi-Stage Samples
Scott W. Keith* and David B. Allison, University of Alabama at Birmingham

3b. A New Distribution for Cumulative Incidence Functions
Sarah R. Haile* and Jong-Hyeon Jeong, University of Pittsburgh

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Scientific Program Poster Presentations

Sunday, March 16 *continued*

3c. Outlier Detection and Goodness of Fit for Recurrent Event Models with Frailty
Jonathan T. Quiton*, Western Kentucky University, Edsel A. Pena, University of South Carolina

3d. Survival Analysis Based on Quantile Regression Models
Limin Peng* and Yijian Huang, Emory University

3e. Minimization-based Interval Estimation of Hazard Ratio Parameters
Kenichi Yoshimura*, National Cancer Center-Japan

3f. Survival Analysis for Multivariate Failure Time Data
Zugui Zhang*, Paul Kolm, Edward F. Ewen, Claudine Jurkovitz and James Bowen, Christiana Care Health System, Joseph Jackson, Bristol-Myers Squibb, Dogan Fidan, Sanofi-Aventis Research-France, William Weintraub, Christiana Care Health System

3g. Testing for Shape Restricted Hazard Function Using Resampling Techniques
Desale Habtzghi*, Georgia College and State University, Somnath Datta, University of Louisville, Mary Meyer, Colorado State University

3h. Semiparametric Methods for the Analysis of Clustered Survival Data from Case-cohort Studies
Hui Zhang*, Douglas E. Schaubel and Jack D. Kalbfleisch, University of Michigan

4. POSTERS: CLINICAL TRIALS: GENERAL
Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology

4a.Evaluation of Outcomes for Stroke: An Illustration of the Sliding Dichotomy
Sharon D. Yeatts*, Yuko Y. Palesch and Scott W. Miller, Medical University of South Carolina

4b. Effects of Heterogeneity on Simon's Two-stage Optimal Design
Rebecca B. McNeil*, Beth K. Rush, Thomas G. Brott, and James F. Meschia, Mayo Clinic

4c. Sensitivity and Specificity of Immunological Correlates of Protection
Yunjie Chen*, Virginia Polytechnic Institute and State University, Andrew J. Dunning, Sanofi Pasteur

4d. A Comparison of the Power of Several Tests for Detecting Qualitative Treatment by Covariate Interactions in Clinical Trials
Scott W. Miller*, Yuko Y. Palesch, Renee H. Martin and Peng Huang, Medical University of South Carolina

4e. On Tests for Qualitative Interaction
Robert A. Parker*, Amgen Inc.

4f. Statistical Assessment of Medication Adherence Data: A Technique to Analyze the J-Shaped Curve
Jeffrey M. Rohay*, Pinney Associates, Gary M. Marsh, Stewart Anderson, Vincent C. Arena, Ada Youk and Jacqueline Dunbar-Jacob, University of Pittsburgh

4g. Bioequivalence and Pharmacokinetics With the use of WinNonlin
Anna M. Maeser*, University of North Carolina-Wilmington and Biostudy Solutions, LLC

4h.Crossover Designs Under Subject Dropout
Shi Zhao* and Dibyen Majumdar, University of Illinois-Chicago

5. POSTERS: CLINICAL TRIALS: ADAPTIVE DESIGN AND ADAPTIVE RANDOMIZATION
Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology

5a. GLMIP 1.0: SAS/IML Software for Planning Repeated Measures Internal Pilots
Keith E. Muller*, University of Florida, Christopher S. Coffey, University of Alabama at Birmingham, John A. Kairalla, University of Florida, Jacqueline L. Johnson, Novartis

5b. A Two-Stage Procedure For Factorial Designs Utilizing A Formal Futility Rule
Nitin K. Nair* and Christopher S. Coffey, University of Alabama at Birmingham

5c. Estimation and Hypothesis Testing Properties for Outcome-based Adaptive Randomization in Clinical Trials with Binary Outcomes --- A Simulation Study Comparing Urn Models and Sequential Estimation Methods
Xuemin Gu* and J. Jack Lee, M. D. Anderson Cancer Center

5d. Can Response-adaptive Randomization Benefit Your Trial?
Amy S. Nowacki* and Wenle Zhao, Medical University of South Carolina

5e. Optimal Adaptive Group Sequential Design for Phase II Clinical Trials: A Bayesian Decision Theoretic Approach
Yiyi Chen* and Brian J. Smith, University of Iowa

6. POSTERS: APPLIED DATA ANALYSIS
Sponsors: ENAR and ASA Health Policy Statistics Section

6a. Bimodality and Variances: Bump Hunting in Motor Control of Parkinson's Disease
Sue Leurgans*, Rush University Medical Center, Julie Robichaud, David Vaillancournt and Daniel Corcos, University of Illinois-Chicago

6b. Order-preserving Dimension Reduction Procedure for the Dominance of the Two Mean Curves with Application of Tidal Volume Curves
Sang Han Lee*, Nathan Kline Institute, Johan Lim, Yonsei University, Marina Vannucci, Rice University, Eva Petkova, New York University, Maurice Preter and Donald K. Klein, Columbia University

6c. Adjustment of Systematic Measurement Error in Human Body Surface Area Measured by Three Dimensional Scan
JungBok Lee*, Korea University, Young Ju Kim, Kangwon University, Dongdeuk Jang, Bong Hyun Nam and Eunhee Kim, Korea National Institute of Toxicological Research

6d. Stationary State of the Population After a Screening Intervention: An M(t)/G/Inf network Model
Shih-Yuan Lee* and Alexander Tsodikov, University of Michigan

6e. A Generalized Growth Mixture Modeling Approach to Predict Institutionalization of People with Alzheimer's Disease (AD) from Anxiety and Depression Trajectories of Caregivers
Song Zhang*, Sati Mazumdar, Steven H. Belle and Richard Schulz, University of Pittsburgh

6f. Ecological Inference in Meta-Analysis
Mireya Diaz*, Case Western Reserve University

6g. Hierarchical Modeling of Removal Experiments to Estimate Catchability of Blue Crab in Virginia
Mary Christman, Xiaobo Li* and Thomas Bohrmann, University of Florida

7. POSTERS: LATENT VARIABLES AND MISSING DATA
Sponsors: ENAR and ASA Survey Research Methods Section

7a. Bayesian Latent Variable Modeling of Outcomes for Bipolar Disorder
Brian Neelon*, A. James O'Malley and Sharon-Lise Normand, Harvard University

7b. Using Concordance Correlation and Multiple Imputation to Combine Two Different Assessments of the Same Construct
Christopher J. Swearingen*, Medical University of South Carolina

7c. Longitudinal Aspects of Nonresponse in the Survey of Industrial Research and Development
Adriana Pérez*, University of Lousiville

7d. The Use of Sample Weights in Hot Deck Imputation
Rebecca R. Andridge* and Roderick J. Little, University of Michigan

7e. Using a Mixture Model for Multiple Imputation in the Presence of Outliers
Michael R. Elliott*, University of Michigan

7f. Estimation in Hierarchical Models with Incomplete Data
Yong Zhang* and Trivellore E. Raghunathan, University of Michigan

7g. Modeling Sensitivities and Specificities for Longitudinal Data: Application to Psychosocial Research
Qin Yu*, Wan Tang, Yan Ma and Xin Tu, University of Rochester

8. POSTERS: STATISTICAL MODELS AND METHODS
Sponsors: ENAR, ASA Section on Risk Analysis and ASA Biometrics Section

8a. Growth Rate Estimates Assuming a Weibull Growth Curve
Alaina M. Houmard*, University of North Carolina-Wilmington

8b. A Comparison Study of Global Models and Confidence Intervals Under the Exponential Growth Model
Lifang Du*, University of North Carolina-Wilmington

8c. Using a Stochastic Search in a Bayesian Hierarchical Regression Model
Qijun Fang*, University of North Carolina-Wilmington

8d. Linear Regression Models for Symbolic Interval-valued Data and The Confidence Interval
Wei Xu* and Lynne Billard, University of Georgia

8e. Bayesian ROC Curve Estimation Under Binormality Using a Partial Likelihood Based on Ranks
Jiezhun Gu*, Duke Clinical Research Institute, Subhashis Ghosal, North Carolina State University

8f. The Test of Stochastic Ordering to Detect Treatment Related Trend with Clustered Discrete Data
Kyeongmi Cheon*, University of Memphis, Aniko Szabo, Medical College of Wisconsin, Ebenezer O. George, University of Memphis

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Scientific Program Poster Presentations

Sunday, March 16 *continued*

8g. An Alternative Method for Analyzing 2 x c Tables
Julia L. Sharp*, Clemson University, John J. Borkowski, Montana State University

8h. Model-Based Canonical Correlation Analysis for Functional Data
Hyejin Shin*, Auburn University, Seokho Lee, Texas A&M University

9. POSTERS: STATISTICAL GENETICS
Sponsors: ENAR and ASA Biopharmaceutical Section

9a. Multiple Imputation for Family Association Studies
Miguel A. Padilla, Howard Wiener, Donald Rubin and Hemant K. Tiwari*, University of Alabama at Birmingham

9b. Imputation of Missing Genotypes: An Empirical Evaluation
Zhenming Zhao* and Paola Sebastiani, Boston University

9c. Population Genetics of Bluefish
Christina M. Stuart*, University of North Carolina-Wilmington

9d. A Computational Procedure for Identifying Estrogen Receptor Binding Sites in Human Osteosarcoma Cells Using High Resolution Tiling Arrays
Hui Tang*, Mayo Clinic

9e. Genetic Association Studies Using Samples Ascertained on the Basis of a Correlated Trait
Genevieve M. Monsees* and Peter Kraft, Harvard University

9f. Allelic Based Gene-Gene Interaction Associated with Quantitative Traits
Jeesun Jung* and Bin Sun, Indiana University, Deukwoo Kwon, National Cancer Institute, Daniel L. Koller and Tatiana M. Foroud, Indiana University

9g. Association Studies of Case-Control Data with Genotyping Uncertainty
Youfang Liu* and Jung-Ying Tzeng, North Carolina State University

9h. Rapid Inference of Haplotypes in Pedigrees
Xiaohua Gong* and Zhao-bang Zeng, North Carolina State University

9i. Mutation Detection for Virus Data
Hongyuan Cao* and Xingye Qiao, University of North Carolina-Chapel Hill

9j. An Empirical Bayesian Method to Correct for Winner's Curse in Genetic Association Studies
Rui Xiao* and Michael Boehnke, University of Michigan

9k. Linkage Disequilibrium Filter in Genome Wide Association Studies
Nadia Timofeev* and Paola Sebastiani, Boston University

10. POSTERS: GENOMICS AND PROTEOMICS
Sponsors: ENAR, ASA Biometrics Section and ASA Biopharmaceutical Section

10a. Bootstrap Aggregation for Ordinal Response Prediction in High-throughput Genomic Datasets
Kellie J. Archer*, Virginia Commonwealth University

10b. Machine Learning Algorithms for Genetic Association Studies
Bareng Aletta S. Nonyane* and Andrea S. Foulkes, University of Massachusetts

10c. A Global Test Procedure with Applications to Gene Expression Profiling Studies
Chien-Ju Lin*, Ching-Wei Chang and James J. Chen, U.S. Food and Drug Administration

10d. An Approach to Detect Statistical Interactors in Gene Networks
Alina Andrei* and Christina M. Kendziorski, University of Wisconsin-Madison

10e. Multivariate Models to Detect Genomic Signatures for a Class of Drugs
Brooke L. Fridley*, Greg Jenkins, Daniel Schaid, Liewei Wang and Fang Li, Mayo Clinic

10f. Systematic Quantification of the Effects of Population Structure on Genome-wide Association Studies
Hongyan Xu* and Varghese George, Medical College of Georgia

10g. Estimation of Genetic Variance Explained for Individual Quantitative Trait Loci
William Bridges*, Clemson University, Steve Knapp, University of Georgia

11. POSTERS: MICROARRAY ANALYSIS
Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Biometrics Section

11a. Modified Linear Discriminant Analysis Approaches for Classification of High-dimensional Microarray Data
Ping Xu*, Guy N. Brock and Rudolph S. Parrish, University of Louisville

Sunday, March 16 *continued*

11b. Exploring the Breast Cancer Gene Expression Data: Identification and Analysis of Functional ERalpha Regulatory Network in Silico
Li-yu D. Liu*, Mei-Ju M. Chen, Ming-Shian Tsai, Cho-Han S. Lee and Chien-yu Chen, National Taiwan University, Tzu L. Phang, University of Colorado, Li-Yun Chang, Wen-hung Kuo, Hsiao-Lin Hwa and Huang-Chun Lien, National Taiwan University

11c. Adjusting for Covariates in the Analysis of Microarray and Other High Dimensional Biological Data
Lang Chen*, T. Mark Beasley, Christopher S. Coffey and Grier P. Page, University of Alabama at Birmingham

11d. A Censored Beta Model for Estimating the Proportion of True Null Hypotheses in a Microarray Experiment
Anastasios Markitsis* and Yinglei Lai, The George Washington University

11e. Hierarchical Mixture Method with Data-adaptive Dimension Reduction for Assessing Differential Expression in Two-sample Microarray Experiments with Small Sample Sizes
Sang-Hoon Cho*, University of Wisconsin - Madison

11f. Realistic Simulation of AffyMetrix Gene Expression Arrays
Andrew C. Hardin*, Southern Methodist University

11g. Bayesian Analysis of Microarray Experiments with Multiple Sources of Variation
Cumhur Y. Demirkale*, Dan Nettleton and Tapabrata Maiti, Iowa State University

11h. A Flexible Model Selection Algorithm for the Cox Model with High-Dimensional Data
Alexander T. Pearson* and Derick R. Peterson, University of Rochester

Monday, March 17

8:30 a.m. - 10:15 a.m.

12. STATISTICAL MONITORING OF CLINICAL TRIALS: WHERE THE RUBBER MEETS THE ROAD *(Regency A)*
Sponsors: ENAR, ASA Section on Statistics in Epidemiology and ASA Biopharmaceutical Section
Organizer: Craig B. Borkowf, Centers for Disease Control and Prevention
Chair: Laura Lee Johnson, National Center for Complementary and Alternative Medicine

8:30 **Disobeying Rules: Unplanned Early Stopping of a Non-inferiority Trial**
Janet T. Wittes*, Statistics Collaborative

8:55 **A Constrained Boundaries Approach for Flexibly Monitoring Time to Event Outcomes**
Sean Brummel and Daniel L. Gillen*, University of California-Irvine

9:20 **Spending Functions and the Updating of Information Fractions**
Michael A. Proschan*, National Institute of Allergy and Infectious Diseases

9:45 **A Conditional Power Approach to the Evaluation of Predictive Power**
Kuang-Kuo G. Lan* and Peter Hu, Johnson & Johnson, Michael A. Proschan, National Institute of Allergy and Infectious Diseases, National Institutes of Health

10:10 **Floor Discussion**

13. STATISTICAL ISSUES IN GENOMIC STUDIES IN POPULATION SCIENCES *(Regency B)*
Sponsor: IMS and ASA Biometrics Section
Organizer: Xihong Lin, Harvard University
Chair: Xihong Lin, Harvard University

8:30 **Gene-Environment Interaction Studies**
Raymond J. Carroll*, Texas A&M University, Nilanjan Chatterjee, National Cancer Institute, Yi-Hau Chen, Academia Sinica, Bhramar Mukherjee, University of Michigan

9:00 **Powerful Multilocus Association Testing of Complex Traits**
Michael P. Epstein*, Emory University

9:30 **Incorporating Prior Biological Knowledge in Genome-Wide Association Studies**
Hongyu Zhao*, Yale University

10:00 **Floor Discussion**

14. HIERARCHICAL MODELING IN ENVIRONMENTAL EXPOSURE AND TOXICOLOGICAL RISK ASSESSMENT *(Regency C)*
Sponsors: ASA Section on Statistics and the Environment, ASA Section on Risk Analysis and ASA Section on Statistics in Epidemiology
Organizer: Michael Pennell, University of Rhode Island
Chair: Michael Pennell, University of Rhode Island

8:30 **Bayesian Hierarchical Regression for Model Selection and Clustering**
Amy H. Herring*, The University of North Carolina at Chapel Hill, David B. Dunson, National Institute of Environmental Health Sciences, Stephanie M. Engel, Mount Sinai School of Medicine

9:00 **Regional Spatial Modeling of Arsenic in Environmental Media: Implications for Human Exposure Assessment**
Catherine A. Calder* and Peter F. Craigmile, The Ohio State University, Hongfei Li, IBM Corporation, Rajib Paul, The Ohio State University, Jian Zhang, Freddie Mac

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

9:30 **Using Hierarchical PK/PD Models for Probabilistic Dose-Response Assessment**
Ralph L. Kodell*, University of Arkansas for Medical Sciences

10:00 **Floor Discussion**

15. MODEL VALIDATION AND MODEL SELECTION IN LONGITUDINAL DATA ANALYSIS *(Regency D)*

Sponsors: ENAR and ASA Biometrics Section
Organizer: Peter X.K. Song, University of Waterloo
Chair: Annie Qu, Oregon State University

8:30 **Quadratic Inference Functions in Marginal Models for Longitudinal Data**
Peter X K Song*, University of Waterloo

8:55 **Reducing the Bias of Between- / Within- cluster Covariate Methods when Data are Missing at Random**
John M. Neuhaus*, University of California-San Francisco, Charles E. McCulloch, University of Chicago

9:20 **Regularization Parameter Selection for Penalized Likelihood Variable Selection Procedures**
Runze Li*, Penn State University

9:45 **A BIC Criterion for Longitudinal Data Model Selection**
Lan Wang*, University of Minnesota, Annie Qu, Oregon State University

10:10 **Floor Discussion**

16. NEW STATISTICAL METHODS IN DIAGNOSTIC MEDICINE *(Washington A)*

Sponsors: ASA Health Policy Statistics Section and ENAR
Organizer: Yueh-Yun Chi, University of Florida
Chair: Yueh-Yun Chi, University of Florida

8:30 **Bootstrap and Empirical Likelihood-based Nonparametric Inference For the Difference between Two Partial AUCs**
Gengsheng Qin* and Yan Yuan, Georgia State University, Xiao-Hua Zhou, University of Washington

9:00 **Estimating Diagnostic Accuracy from Designs with no Gold Standard, Partial Gold Standard, or Imperfect Gold Standard Evaluation**
Paul S. Albert*, National Cancer Institute

9:30 **Semiparametric Least Squares ROC Analysis of Correlated Biomarker Data**
Liansheng Tang*, George Mason University, Xiao-Hua Zhou, University of Washington

10:00 **Discussant:** Constantine Gatsonis, Brown University

17. REGULARIZATION AND HIERARCHICAL MODELING: TWO PHILOSOPHIES FOR VARIABLE SELECTION

(Washington B)
Sponsors: ENAR, ASA Section on Statistics and the Environment and ASA Biometrics Section
Organizer: Samiran Ghosh, Indiana University, Purdue University-Indianapolis
Chair: Dipankar Bandyopadhyay, Medical University of South Carolina

8:30 **Model Selection and Diagnostics for Genetic Markers**
Dipak K. Dey*, University of Connecticut, Feng Guo, Virginia Tech University, Kent E. Holsinger, University of Connecticut

9:00 **Detecting Differentially Expressed Genes via Multilevel Nonlinear Mixture Dirichlet Models for High-Dimensional EST Data**
Fang Yu, University of Nebraska Medical Center, Ming-Hui Chen* and Lynn Kuo, University of Connecticut, Peng Huang, Medical University of South Carolina, Wanling Yang, The University of Hong Kong

9:30 **Variable Selection via Adaptive Elastic Net and its Consistency Property**
Samiran Ghosh*, Indiana University, Purdue University-Indianapolis

10:00 **Floor Discussion**

18. CONTRIBUTED PAPERS:METHODS FOR VARIABLE SELECTION AND MODEL BUILDING *(Potomac 1)*

Sponsor: ENAR
Chair: Jane F. Pendergast, University of Iowa

8:30 **Bayesian Variable Selection Under Heredity**
Woncheol Jang*, University of Georgia, Johan Lim, Yonsei University

8:45 **A Cp Statistic for Fixed Effects Variable Selection in the Linear Mixed Model for Longitudinal Data**
Anita A. Abraham* and Lloyd J. Edwards, University of North Carolina at Chapel Hill

9:00 **A Bayesian Approach to Effect Estimation Accounting for Adjustment Uncertainty**
Chi Wang*, Giovanni Parmigiani, Ciprian Crainiceanu and Francesca Dominici, Johns Hopkins University

9:15 **Nonparametric Bayes Conditional Distribution Modeling with Variable Selection**
Yeonseung Chung*, University of North Carolina, David B. Dunson, National Institute of Environmental Health Sciences

9:30 **The Adaptive Dantzig Selector for Variable Selection**
Lee Dicker* and Xihong Lin, Harvard University

9:45 **Bivariate Dimension Reduction and Variable Selection When n<<p**
Lexin Li, North Carolina State University, Xuerong Wen*, University of Missouri-Rolla

10:00 **A Link-Free Method for Testing the Significance of Predictors**
Peng Zeng*, Auburn University

19. CONTRIBUTED PAPERS: RESEARCH METHODS
(Potomac 2)
Sponsors: ENAR and ASA Section on Statistics in Epidemiology
Chair: Sung-joon Min, University of Colorado Health Sciences Center

8:30 **Outcome-dependent Sampling in the Survival Analysis Setting**
Denise Esserman*, University of North Carolina at Chapel Hill, Yanyan Liu, Wuhan University of Peoples Republic of China, Haibo Zhou, University of North Carolina at Chapel Hill

8:45 **Marginal Hazards Model for Case-Cohort Studies with Multiple Disease Outcomes**
Sangwook Kang*, University of Georgia, Jianwen Cai, University of North Carolina at Chapel Hill

9:00 **Estimating Family Relationships using DNA Fingerprints within the NHANES-III Household Survey**
Hormuzd A. Katki*, National Cancer Institute, Christopher Sanders, National Center for Health Statistics, Barry I. Graubard, National Cancer Institute, Andrew Bergen, SRI International

9:15 **Flexible Modelling of Age-Dependent Infectious Disease Parameters Using Seroprevalence Data in Combination With Social Contact Data**
Marc Aerts*, Niel Hens, Nele Goeyvaerts, Kaatje Bollaerts, Ziv Shkedy and Christel Faes, Hasselt University-Belgium, Benson Ogunjimi, Olivier Lejeune, Pierre Van Damme and Philippe Beutels, University of Antwerp-Belgium

9:30 **Benchmark Analysis for Two Predictor Variables**
Roland C. Deutsch*, University of North Carolina-Greensboro, John M. Grego and Brian T. Habing, University of South Carolina, Walter W. Piegorsch, University of Arizona

9:45 **Quantifying Treatment Effect when Flexibly Modeling Individual Change in a Nonlinear Mixed Effects Model**
Robert J. Gallop*, West Chester University

10:00 **Covariate-adjusted Putative Placebo Analysis in Active-controlled Clinical Trials**
Zhiwei Zhang*, U.S. Food and Drug Administration

20. CONTRIBUTED PAPERS: GENOMICS AND MICROARRAY ANALYSES *(Potomac 3)*
Sponsors: ENAR and ASA Biometrics Section
Chair: Yu Shyr, Vanderbilt University

8:30 **Inferring the True Correlation in Cross-species Microarray Data**
Sunghee Oh and George C. Tseng*, University of Pittsburgh

8:45 **A Linear Mixed Effects Clustering Model for Multi-species Time Course Gene Expression Data**
Kevin H. Eng*, Sunduz Keles and Grace Wahba, University of Wisconsin - Madison

9:00 **Reconstruction of Genetic Association Networks from Microarray Data: A Partial Least Squares Approach**
Vasyl Pihur*, Somnath Datta and Susmita Datta, University of Louisville

9:15 **Modeling Spatial Correlation in Gene Regulation**
Guanghua Xiao*, University of Texas, Southwestern Medical Center, Cavan Reilly and Arkady B. Khodursky, University of Minnesota

9:30 **A Visualization of Normality Transformations**
Glen Satten, Centers for Disease Control and Prevention, Somnath Datta, Bart Brown and Guy Brock*, University of Louisville

9:45 **Statistical Test for the ChIP-chip Tiling Arrays Without Replication**
Youngchul Kim* and Taesung Park, Seoul National University, Seungyeoun Lee, Sejong University, Jae K. Lee, University of Virginia

10:00 **A Meta-Analysis Approach for Gene Association Network Reconstruction**
YounJeong Choi* and Christina M. Kendziorski, University of Wisconsin-Madison

21. CONTRIBUTED PAPERS:EPIDEMIOLOGIC METHODS
(Potomac 4)
Sponsors: ASA Section on Statistics in Epidemiology, ASA Section on Risk Analysis and ASA Section on Statistics and the Environment
Chair: Jesse A. Berlin, Johnson & Johnson Pharmaceutical Research and Development

8:30 **Estimating Disease Prevalence Indirectly Using Incidence and Survival Data**
Kent R. Bailey*, Mayo Clinic

8:45 **Inference for Multinomial Parameters and Odds Ratio Under Order Restrictions**
Broderick O. Oluyede*, Georgia Southern University, Mavis Pararai, Indiana University of Pennsylvania

9:00 **Estimation of Incidence and Remission Rates From Cross-Sectional Survey Data**
Jason Roy* and Walter Stewart, Geisinger Center for Health Research

9:15 **Bayesian Inference for Non-Isotropic Changes in Risk about a Fixed Point**
Ronald E. Gangnon*, University of Wisconsin, Jane A. McElroy, University of Missouri

9:30 **Using Matching Estimators to Explore the Potential Influence of Unmeasured Smoothly Time-Varying Confounders**
Yun Lu* and Scott L. Zeger, Johns Hopkins University

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

9:45 **Identifying Effect Modifiers in Air Pollution Time-series Studies using a Two-stage Analysis**
Sandrah P. Eckel* and Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health

10:00 **An Algorithm for Optimal Tapered Matching, With Application to Disparities in Survival**
Shoshana R. Daniel*, Katrina Armstrong, Jeffrey H. Silber and Paul R. Rosenbaum, Univeristy of Pennsylvania

22. CONTRIBUTED PAPERS: BAYESIAN METHODS

(Potomac 5)
Sponsors: ENAR, ASA Section on Statistics and the Environment, and ASA Biometrics Section
Chair: Timothy D. Johnson, University of Michigan

8:30 **Hierarchical Bayes Estimation For Bivariate Binary Data with Applications to Small Area Estimation**
Malay Ghosh, University of Florida, Ananya Roy*, University of Nebraska-Lincoln, Ming-Hui Chen, University of Connecticut, Myron Katzoff and Van L. Parsons, National Center for Heath Statistics

8:45 **Bayesian Assessment of Hierarchical Models**
Ying Yuan* and Valen Johnson, University of Texas MD Anderson Cancer Center

9:00 **Joint Effect from Environmental and Behavior Risk Factors to Cancer Death Rates using Bayesian Multi-level Modeling**
Hongmei Zhang*, University of South Carolina

9:15 **Bayesian Generalized Product Partition Model**
Ju-Hyun Park*, University of North Carolina at Chapel Hill, David B. Dunson, National Institute of Environmental Health Sciences

9:30 **Bayesian Mixture Labelling by Highest Posterior Density**
Weixin Yao*, Kansas State University, Bruce G. Lindsay, The Pennsylvania State University

9:45 **A Practical Procedure to Find Matching Priors for Frequentist Inference**
Juan Zhang* and John E. Kolassa, Rutgers University

10:00 **How to Choose Hyperparameter Values in Normal Models**
Susan Alber* and J. Jack Lee, University of Texas, MD Anderson Cancer Center

23. CONTRIBUTED PAPERS: LATENT VARIABLES AND STRUCTURAL EQUATIONS MODELING *(Potomac 6)*

Sponsor: ENAR
Chair: Knashawn Morales, University of Pennsylvania

8:30 **Exploratory Structural Equation Modeling**
Bengt Muthen*, University of California Los Angeles, Tihomir Asparouhov, Mplus

8:45 **Incorporating Structural Equation Modeling to Fecundability models**
Sung Duk Kim* and Rajeshwari Sundaram, National Institute of Child Health and Human Development

9:00 **Structural Equation Models with Functional Latent Variables**
Mingan Yang*, David B. Dunson and Pablo Nepomnaschy, National Institute of Environmental Health Sciences-NIH

9:15 **Multilevel Latent Class Models with Dirichlet Mixing Distribution**
Chongzhi Di* and Karen Bandeen-Roche, Johns Hopkins University

9:30 **Hierarchical Mixed Membership Models for Honey Bee Genomes**
Tanzy M. Love*, Carnegie Mellon University

9:45 **Confounding Effects in Genetic Analysis**
Mariza de Andrade*, Brooke L. Fridley and Stephen T. Turner, Mayo Clinic

10:00 **A Genomic Imprinting Test for Ordinal Traits in Pedigree Data**
Rui Feng*, University of Alabama at Birmingham, Heping Zhang, Yale University

Monday, March 17

10:15 a.m.-10:30 a.m. **Refreshment Break and Visit the Exhibitors**

10:30 a.m.-12:15 p.m.

24. NON-INFERIORITY CLINICAL TRIALS - FIXED MARGIN OR NO FIXED MARGIN *(Regency A)*

Sponsors: ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology
Organizer: Sue-Jane Wang, U.S. Food and Drug Administration
Chair: H.M. James Hung, U.S. Food and Drug Administration

10:30 **Controlling the Type 1 Error Rate in the Presence of Departures From the Assumptions of Assay Sensitivity and Constancy**
Steven Snapinn* and Qi Jiang, Amgen

11:00 **Assessing the Evidence for Non-Inferiority: A Likelihood Approach**
Sue-Jane Wang, U.S. Food and Drug Administration, Jeffrey Blume*, Brown University

11:30 **Empirical Data on the Choice of Irrelevance Margins Results from a Systematic Review**
Stefan Lange* and Guido Skipka, Institute for Quality and Efficiency in Health Care-Germany

12:00 **Discussant:**
Robert O'Neill, U.S. Food and Drug Administration

25. MODEL SELECTION FOR HIGH DIMENSIONAL DATA WITH PRACTICAL SOLUTIONS *(Regency B)*

Sponsors: ENAR and ASA Biometrics Section
Organizer: Annie Qu, Oregon State University
Chair: Annie Qu, Oregon State University

10:30 **Spline Estimation of Single Index Models**
Lijian Yang*, Michigan State University, Li Wang, University of Georgia

10:55 **Parsimonious Models for Correlation Matrices using Ordered Partial Correlations**
Michael J. Daniels*, University of Florida, Mohsen Pourahmadi, Northern Illinois University

11:20 **Semiparametric Modeling of Clustered Biomedical Data**
Naisyin Wang*, Texas A&M University, Annie Qu, Oregan State University

11:45 **Generalization Error Estimation in Semisupervised Learning**
Junhui Wang*, Columbia University, Xiaotong Shen, University of Minnesota

12:10 **Floor Discussion**

26. SPATIAL AND SPATIO-TEMPORAL LATENT STRUCTURE MODELING *(Regency C)*

Sponsors: ENAR and ASA Section on Statistics and the Environment
Organizer: Andrew Lawson, University of South Carolina
Chair: Linda Young, University of Florida

10:30 **A Spatial Latent Class Model for Multivariate Spatial Data**
Melanie M. Wall*, University of Minnesota

11:00 **Spatial Dynamic Factor Analysis**
Hedibert F. Lopes, University of Chicago, Esther Salazar and Dani Gamerman*, Universidade Federal do Rio de Janeiro

11:30 **Mixture- and GAM-based Models for Space-time Latent Structure in Health Data**
Bo Cai, Andrew B. Lawson* and Kun Huang, University of South Carolina

12:00 **Floor Discussion**

27. NEW ADVANCES IN SURVIVAL ANALYSIS FOR LARGE DIMENSIONAL BIOMEDICAL DATA *(Regency D)*

Sponsors: IMS and ASA Biometrics Section
Organizer: Yi Li, Dana-Farber Cancer Institute
Chair: Yi Li, Dana-Farber Cancer Institute

10:30 **Inverse Regression for Censored Data**
Nivedita Nadkarni and Michael R. Kosorok*, University of North Carolina-Chapel Hill

10:55 **Analyzing Event Time Data in the Presence of High-dimensional Confounders**
Donglin Zeng* and Danyu Lin, University of North Carolina

11:20 **Survival Analysis of Case-Control Family Data with General Semi-Parametric Shared Frailty Model and Missing Genetic Information**
Anna Graber* and Malka Gorfine, Technion - Israel Institute of Technology, Li Hsu, Fred Hutchinson Cancer Research Center

11:45 **Survival Analysis With Large Dimensional Covariates: An Application In Microarray Studies**
David Engler*, Brigham Young University, Yi Li, Dana-Farber Cancer Institute

12:10 **Floor Discussion**

28. ON THE UTILITY OF DETERMINISTIC MODELS FOR CAUSAL EFFECTS *(Washington A)*

Sponsor: ENAR
Organizer: Marshall Joffe, University of Pennsylvania School of Medicine
Chair: Marshall Joffe, University of Pennsylvania School of Medicine

10:30 **Causal Inference for Continuous Time Processes Observed Only at Discrete Times**
Mingyuan Zhang*, Marshall M. Joffe and Dylan Small, University of Pennsylvania

10:55 **Comparison of Deterministic versus Stochastic Monotonicity Assumptions for Instrumental Variables Method**
Dylan S. Small*, University of Pennsylvania

11:20 **Mimicking Counterfactual Outcomes to Avoid Deterministic Treatment Effects**
Judith J. Lok*, Harvard School of Public Health

11:45 **Causal Models for the Effects of Weight Gain on Mortality**
James M. Robins*, Harvard University

12:10 **Floor Discussion**

29. ESTIMATING THE DISTRIBUTION OF USUAL INTAKES OF NUTRIENTS AND FOODS, AND RELATING USUAL INTAKE TO HEALTH PARAMETERS IN A NATIONAL HEALTH SURVEY *(Washington B)*

Sponsors: ENAR, ASA Section on Statistics in Epidemiology and ASA Survey Research Methods Section
Organizer: Janet A. Tooze, Wake Forest University School of Medicine
Chair: Phillip S. Kott, U.S. Department of Agriculture, National Agricultural Statistical Service

10:30 **Challenges in the Estimation of Usual Intake of Foods and Nutrients**
Patricia M. Guenther*, U.S. Department of Agriculture, Dennis W. Buckman, Information Management Services, Inc., Raymond J. Carroll, Texas A&M University, Kevin W. Dodd, National Cancer Institute, Laurence S. Freedman, Gertner Institute for Epidemiology and Health Policy Research, Victor Kipnis, Susan M. Krebs-Smith, Douglas Midthune and Amy F. Subar, National Cancer Institute, Janet A. Tooze, Wake Forest University School of Medicine

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

10:45 **A Two-part Mixed Effects Model with Correlated Random Effects to Model Food and Nutrient Intake**
Janet A. Tooze*, Wake Forest University School of Medicine, Doug Midthune and Kevin W. Dodd, National Cancer Institute, Laurence S. Freedman, Gertner Institute for Epidemiology and Health Policy Research, Susan M. Krebs-Smith and Amy F. Subar, National Cancer Institute, Patricia M. Guenther, U.S. Department of Agriculture, Raymond J. Carroll, Texas A&M University, Victor Kipnis, National Cancer Institute

11:00 **Estimation of Usual Food and Nutrient Intake Distributions in the National Health and Nutrition Examination Survey (NHANES)**
Kevin W. Dodd*, National Cancer Institute, Dennis W. Buckman, Information Management Services, Inc., Raymond J. Carroll, Texas A&M University, Laurence S. Freedman, Gertner Institute for Epidemiology and Health Policy Research, Patricia M. Guenther, U.S. Department of Agriculture, Victor Kipnis, Susan M. Krebs-Smith, Douglas Midthune and Amy F. Subar, National Cancer Institute, Janet A. Tooze, Wake Forest University School of Medicine

11:30 **Modeling the Relationship Between Usual Food Intake and Health Parameters in the National Health and Nutrition Examination Survey (NHANES)**
Laurence S. Freedman*, Gertner Institute for Epidemiology and Health Policy Research, Douglas Midthune, National Cancer Institute, Dennis W. Buckman, Information Management Services, Inc., Kevin W. Dodd, National Cancer Institute, Raymond J. Carroll, Texas A&M University, Janet A. Tooze, Wake Forest University School of Medicine, Patricia M. Guenther U.S. Department of Agriculture, Susan M. Krebs-Smith, Amy F. Subar and Victor Kipnis, National Cancer Institute

12:00 **Discussant**: Lester (Randi) Curtin, National Center for Health Statistics, Centers for Disease Control and Prevention

30. CONTRIBUTED PAPERS: MULTIPLE TESTING

(Potomac 1)
Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Jimin Choi, H. Lee Moffitt Cancer Center & Research Institute

10:30 **General Gatekeeping Procedures with Logical Restrictions**
Alex Dmitrienko, Eli Lilly & Company, Lingyun Liu and Ajit C. Tamhane*, Northwestern University

10:45 **A Unified Approach for Constructing a Closed Multiple Testing Procedure for a Fixed Sequence of Families with Multiple Null Hypotheses**
Hanjoo Kim*, University of Pennsylvania School of Medicine, Richard A. Entsuah, Wyeth Research, Justine Shults, University of Pennsylvania School of Medicine

11:00 **Extension of Piegorsch and Casella Simultaneous Intervals to Generalized Linear Models and Functions of Their Parameters**
Amy E. Wagler* and Melinda McCann, Oklahoma State University

11:15 **A Bayesian Approach to Large Scale Simultaneous Inference**
Bing Han*, RAND Corporation, Steven F. Arnold and Naomi Altman, Penn State Univeristy

11:30 **Considering P-Value Dependence in a Stepwise Multiplicity Adjustment Method**
Richard E. Blakesley*, Sati Mazumdar and Patricia R. Houck, University of Pittsburgh

11:45 **On Consonance of Closed Testing in Combination Drug Efficacy Trials**
Julia Soulakova*, University of Nebraska

12:00 **Fast FSR and Inference in Regression**
Dennis D. Boos* and Leonard A. Stefanski, North Carolina State University

31. CONTRIBUTED PAPERS: BIOMARKERS *(Potomac 2)*

Sponsors: ENAR and ASA Biometrics Section
Chair: Tomasz Burzykowski, University of Hasselt, Belgium

10:30 **Identifying High-dimensional Biomarkers for Personalized Medicine via Variable Importance Ranking**
Hojin Moon*, California State University-Long Beach, Songjoon Baek, U.S. Food and Drug Administration, Hongshik Ahn, Stony Brook University, Ralph L. Kodell, University of Arkansas for Medical Sciences, Chien-Ju Lin and James J. Chen, U.S. Food and Drug Administration

10:45 **Convex Hull Ensemble Method for Class Prediction with Application to Personalized Medicine**
Songjoon Baek*, U.S. Food and Drug Administration, Ralph L. Kodell, University of Arkansas for Medical Sciences, Hojin Moon, California State University, James J. Chen, U.S. Food and Drug Administration

11:00 **Dynamic Optimal Strategy for Monitoring Disease Recurrence**
Hong Li*, Brown University

11:15 **Identifying Genes That Respond to Abiotic Stress**
Haiyan Wang*, Kansas State University

11:30 **Enhanced Endpoint Analysis Using Auxiliary Information in Clinical Trials**
Linda Sun* and Cong Chen, Merck & Company

11:45 **A Study of Logic Regression with Application to Bladder Cancer**
Bethany J. Wolf*, Omar Moussa, James Klein Dennis K. Watson and Elizabeth H. Slate, Medical University of South Carolina

12:00 **Use of a Pseudo Maximum Likelihood Estimator to Simplify Computations for a Multivariate Left-Censored Longitudinal Model**
Ghideon S. Ghebregiorgis* and Lisa Weissfeld, University of Pittsburgh

32. CONTRIBUTED PAPERS:JOINT MODELS- SURVIVAL AND LONGITUDINAL *(Potomac 3)*

Sponsors: ENAR, ASA Biometrics Section and ASA Biopharmaceutical Section
Chair: Kaushik Ghosh, University of Nevada- Las Vegas

10:30 **Joint Model of Longitudinal Process and State-change Process**
Caitlin Ravichandran*, McLean Hospital and Harvard University, Victor DeGruttola, Harvard University

10:45 **Joint Modeling of Longitudinal Outcomes and Event Time Data in a Rheumatoid Arthritis Study**
Li Zhu*, University of California-Davis, Juan Li and Eric Chi, Amgen Inc.

11:00 **Fully Exponential Laplace Approximations for the Joint Modeling of Survival and Longitudinal Data**
Dimitris Rizopoulos*, Geert Verbeke and Emmanuel Lesaffre, Catholic University, Leuven-Belgium

11:15 **Joint Analysis of Generalized Longitudinal Measurements and Survival Data with Multiple Failure Types**
Ning Li*, Robert Elashoff and Gang Li, University of California Los Angeles

11:30 **Prediction of Event Risk Using Joint Models for Longitudinal Measurements and Event Times**
Nicholas J. Salkowski* and Melanie M. Wall, University of Minnesota

11:45 **Simultaneous Modeling of Longitudinal Data and Dropout Process in Clinical Studies**
Qinfang Xiang*, Endo Pharmaceuticals

12:00 **Bayeisan Joint Analysis of Longitudinal Measurements and Competing Risks Failure Time Data via Modeling of Multivariate Random Effects Covariances**
Xin Huang*, Gang Li and Robert M. Elashoff, Univeristy of California, Los Angeles, Jianxin Pan, University of Manchester-UK

33. CONTRIBUTED PAPERS:DATA MINING AND MACHINE LEARNING *(Potomac 4)*

Sponsor: ENAR
Chair: Yoonkyung Lee, Ohio State University

10:30 **Carrying Prediction Models Across Microarray Data Sets Generated by Different Labs and Different Platforms**
Chunrong Cheng* and George C. Tseng, University of Pittsburgh

10:45 **Variable Selection in Penalized Model-based Clustering via Regularization on Grouped Parameters**
Benhuai Xie*, Wei Pan and Xiaotong Shen, University of Minnesota

11:00 **Principal Component Analysis for Interval-valued Symbolic Data**
Jennifer Le-Rademacher*, University of Georgia

11:15 **LASSO-Patternsearch Algorithm**
Weiliang Shi*, Grace Wahba, Stephen Wright, Kristine Lee, Ronald Klein and Barbara Klein, University of Wisconsin-Madison

11:30 **Identifying Represenative Trees in Random Forest for Survival Data**
Mousumi Banerjee, Ying Ding* and Anne-Michelle Noone, University of Michigan

11:45 **Data Mining for Medical Research: Identifying Relationships That Cannot Be Identified In Any Other Way**
Shenghan Lai*, Johns Hopkins University

12:00 **Sparse Distance Weighted Discrimination**
Lingsong Zhang* and Xihong Lin, Harvard University

34. CONTRIBUTED PAPERS:GENETIC EPIDEMIOLOGY-ASSOCIATIONS *(Potomac 5)*

Sponsor: ENAR and ASA Biopharmaceutical Section
Chair: Hongzhe Li, University of Pennsylvania

10:30 **Estimating Odds Ratios in Genome Scans: An Approximate Conditional Likelihood Approach**
Arpita Ghosh*, Fei Zou and Fred A. Wright, The University of North Carolina at Chapel Hill

10:45 **Ranges of Measures of Association for Pedigree Binary Variables**
Yihao Deng*, Indiana University-Purdue University-Fort Wayne, N. Rao Chaganty, Old Dominion University

11:00 **Score Statistics for Family-based Association Mapping of Quantitative Traits**
Samsiddhi Bhattacharjee* and Eleanor Feingold, University of Pittsburgh

11:15 **A Semiparametric Association Test in Structured Populations**
Meijuan Li* and Tim Hanson, University of Minnesota

11:30 **A Bayesian Semiparametric Framework for Genetic Association Studies in the Presence of Population Structure**
Nicholas M. Pajewski* and Purushottam W. Laud, Medical College of Wisconsin

11:45 **A Joint Test for Haplotype-based Association in Case-Control Studies**
Tao Wang*, Howard Jacob, Soumitra Ghosh and Xujing Wang, Medical College of Wisconsin, Zhao-Bang Zeng, North Carolina State University

12:00 **Efficient Multivariate Polygenic and Association Analysis**
Wei-Min Chen*, University of Virginia

35. CONTRIBUTED PAPERS:GENERALIZED LINEAR MODELS *(Potomac 6)*

Sponsors: ENAR and ASA Biometrics Section
Chair: Dirk F. Moore, University of Medicine and Dentistry of New Jersey

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Scientific Program

Monday, March 17 *continued*

10:30 **Pseudo-Score Inference for Parameters in Discrete Statistical Models**
Alan Agresti*, University of Florida, Euijung Ryu, Mayo Clinic

10:45 **Simultaneous Confidence Intervals for Comparing Binomial Parameters**
Alan Agresti, University of Florida, Matilde Bini and Bruno Bertaccini, University of Florence-Italy, Euijung Ryu*, Mayo Clinic

11:00 **Bayesian Generalized Linear Models for Genome-wide QTL Analysis**
Nengjun Yi* and Samprit Banerjee, University of Alabama at Birmingham

11:15 **Regression Splines for Threshold Selection with Application to a Random Effects Logistic Dose-response Model**
Daniel L. Hunt*, St. Jude Children's Research Hospital, Chin-Shang Li, University of California-Davis

11:30 **Diagnosis of Model Misspecification For Random Effects in Generalized Linear Mixed Models**
Xianzheng Huang*, University of South Carolina

11:45 **Estimation of Treatment Effects by Combining Two or More Design Plans**
Yvonne M. Zubovic* and Chand K. Chauhan, Indiana University Purdue University-Fort Wayne

12:00 **Approaches for Estimating and Testing Conditional Correlations**
Xueya Cai*, Gregory, E. Wilding and Alan Hutson, SUNY-Buffalo

Monday, March 17

12:15-1:30 p.m. **Roundtable Luncheons**
(Registration Required)

1:45-3:30 p.m.

36. STATISTICAL ISSUES IN ANALYSIS OF GENOMICS DATA *(Regency A)*

Sponsors: ASA Biopharmaceutical Section and ENAR
Organizers: Boris Zaslavsky and Jawahar Tiwari, U.S. Food & Drug Administration
Chair: Jawahar Tiwari, U.S. Food and Drug Administration

1:45 **Overcoming Adverse Effects of Correlations in Microarray Data Analysis**
Lev Klebanov, Charles University-Czech Republic, Andrei Yakovlev*, University of Rochester

2:15 **A Gene Expression Barcode for Microarray Data**
Rafael A. Irizarry* and Michael J. Zilliox, Johns Hopkins University

2:45 **A Systematic Framework of Whole-genome SNP Screening Seeking a Predictive Genomic Biomarker for Potential Treatment Individualization**
Sue-Jane Wang*, U.S. Food and Drug Administration

3:15 **Discussant:** Xing Qui, University of Rochester

37. RECENT DEVELOPMENTS IN NON-SMOOTH ESTIMATING FUNCTIONS FOR CENSORED DATA *(Regency B)*

Sponsors: ENAR and ASA Biometrics Section
Organizer: Menggang Yu, Indiana University
Chair: Bin Nan, University of Michigan

1:45 **Estimation in the Semiparametric Accelerated Failure Time Model with Missing Data**
Bin Nan and John D. Kalbfleisch*, University of Michigan, Menggang Yu, Indiana University

2:10 **Accelerated Recurrence Time Models**
Yijian Huang* and Limin Peng, Emory University

2:35 **On Combining Multiple Estimating Equations**
Lu Tian*, Northwestern University

3:00 **Variance Estimation in Censored Linear Regression**
Zhezhen Jin*, Columbia University

3:25 **Floor Discussion**

38. BAYESIAN METHODS IN EPIDEMIOLOGY *(Regency C)*

Sponsors: ASA Section on Statistics in Epidemiology and ENAR
Organizer: Bhramar Mukherjee, University of Michigan
Chair: Samiran Sinha, Texas A&M University

1:45 **Novel Bayesian Approaches to Model-robust Inference**
Kenneth M. Rice*, Adam Szpiro and Thomas Lumley, University of Washington

2:10 **Bayesian Modeling of Complex Traits**
Paola Sebastiani*, Boston University

2:35 **A Bayesian Approach for Estimating the Effect of Exposure to Air Pollution on Cardiovascular Disease Risk Factors**
Trivellore E. Raghunathan* and Yun Bai, University of Michigan

3:00 **Empirical Bayes-Type Shrinkage Estimation in Genetic Epidemioloy**
Bhramar Mukherjee*, University of Michigan, Nilanjan Chatterjee, National Cancer Institute

3:25 **Floor Discusssion**

39. LONGITUDINAL DATA ANALYSIS IN THE PRESENCE OF MORTALITY *(Regency D)*

Sponsor: ENAR, ASA Section on Statistics in Epidemiology and ASA Biometrics Section
Organizer: Daniel Scharfstein, Johns Hopkins University
Chair: Daniel Scharfstein, Johns Hopkins University

1:45 **Analysis of Longitudinal Data Subject to Truncation due to Death**
Patrick J. Heagerty*, University of Washington

2:15 **Identification and Estimation of the Survivor Average Causal Effect**
Brian L. Egleston*, Fox Chase Cancer Center

2:45 **Using an Intervention-Based Framework to Address Input Data Missing Due to Death**
Constantine E. Frangakis*, Johns Hopkins University, Donald B. Rubin, Harvard University, Ming-Wen An and Ellen MacKenzie, Johns Hopkins University

3:15 **Floor Discussion**

40. IMPROVING MEASUREMENT IN PROFILING PROVIDERS OF HEALTHCARE *(Washington A)*

Sponsors: ASA Health Policy Statistics Section and ENAR
Organizer: Marc Elliott, RAND Corporation
Chair: Yulei He, Harvard University

1:45 **Findings from the HCAHPS Mode Experiment**
Marc N. Elliott*, RAND Corporation, Elizabeth Goldstein and William G. Lehrman, Centers for Medicare and Medicaid Services, Alan M. Zaslavsky, Harvard University, Katrin Hambarsoomians, RAND Corporation, Mary Anne Hope and Laura A. Giordano, Health Services Advisory Group

2:10 **Ensuring the Reliability of Consumer Assessments of Clinicians and Groups in the CAHPS 4.0 Clinician and Group Survey**
Dana Safran*, Tufts University, Marc N. Elliott, Julie Brown and Ron D. Hays, RAND Corporation

2:35 **Improving Subgroup Comparisons of Consumer Reports by Adjusting for Differences in Extreme Response Tendency**
Amelia Haviland*, Marc N. Elliott and Katrin Hambarsoomians, RAND Corporation

3:00 **Optimal Survey Design When Nonrespondents are Subsampled for the Follow-Up**
A. James O'Malley* and Alan M. Zaslavsky, Harvard University

3:25 **Floor Discussion**

41. BAYESIAN VARIABLE SELECTION WITH HIGH DIMENSIONAL COVARIATE DATA *(Washington B)*

Sponsors: IMS and ASA Biometrics Section
Organizer: Joseph Ibrahim, University of North Carolina at Chapel Hill
Chair: Debajyoti Sinha, Medical University of South Carolina

1:45 **A Stochastic Partitioning Method to Associate High-dimensional Datasets**
Stefano Monni, University of Pennsylvania, Mahlet G. Tadesse*, Georgetown University

2:15 **Bayesian Variable Selection and Monte Carlo Computation for High Dimensional Data in Regression Models**
Faming Liang, Texas A&M University, Ming Hui Chen, University of Connecticut, Joseph G. Ibrahim, University of North Carolina at Chapel Hill, Mayetri Gupta*, Boston University

2:45 **Variable Selection via a Bayesian Ensemble**
Hugh A. Chipman, Acadia University, Edward I George*, University of Pennsylvania, Robert E. McCulloch, University of Chicago

3:15 **Floor Discussion**

42. CONTRIBUTED PAPERS: CLUSTERED AND HIERARCHICAL MODELS *(Potomac 1)*

Sponsors: ASA Section on Statistics and the Environment, ASA Section on Statistics in Epidemiology, and ASA Biometrics Section
Chair: Hernando Ombao, Brown University

1:45 **Vaccine Efficacy Trials using Stepped Wedge Design**
JoAnna Scott*, University of Washington

2:00 **A General Class of Agreement Coefficients for Categorical Responses**
Wei Zhang* and Vernon M. Chinchilli, The Pennsylvania State University

2:15 **Comparison of Parametric, Nonparametric and Smooth Nonparametric Mixed Effect Models**
Tihomir Asparouhov*, Mplus, Bengt Muthen, University fo California Los Angeles

2:30 **Analysis of Group Randomized Trials with Multiple Binary Endpoints and Small Number of Groups**
Ji-Hyun Lee* and Michael, J. Schell, H. Lee Moffitt Cancer Center & Research Institute, Richard Roetzheim, University of South Florida

2:45 **Hierarchical Spatial Modeling of Additive and Dominance Genetic Variance for Large Spatial Trial Datasets**
Andrew O. Finley*, Michigan State University, Sudipto Banerjee, University of Minnesota, Patrik Waldmann, and Tore Ericsson, Swedish University of Agricultural Sciences

3:00 **Bayesian Groundwater Contamination Model**
Yongsung Joo*, University of Florida, Keunbaik Lee and Donald Mercante, Louisiana State University

3:15 **A Positive Stable Frailty Model for Clustered Failure Time Data with Covariate Dependent Frailty**
Dandan Liu*, John D. Kalbfleisch and Doug E. Schaubel, University of Michigan

43. CONTRIBUTED PAPERS:STATISTICAL GENETICS *(Potomac 2)*

Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Guoqing Diao, George Mason University

1:45 **Binary Trait Mapping in Experimental Crosses with Selective Eenotyping**
Ani Manichaikul*, Johns Hopkins University, Karl W. Broman, University of Wisconsin-Madison

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

2:00 **A Time-Dependent Poisson Random Field Model for Polymorphism Within and Between Two Related Species**
Amei Amei*, University of Nevada Las Vegas, Stanley Sawyer, Washington University in St. Louis

2:15 **An Incomplete-Data Quasi-likelihood Framework with Application to Genetic Association Studies on Related Individuals**
Zuoheng Wang* and Mary Sara McPeek, University of Chicago

2:30 **An Epistatic Model for Mapping Phenotypic Plasticity of a Count Trait**
Arthur Berg*, Derek Drost, Evandro Novaes, Matias Kirst and Rongling Wu, University of Florida

2:45 **Genotyping Error Detection in Samples of Unrelated Individuals**
Nianjun Liu*, University of Alabama at Birmigham, Dabao Zhang, Purdue University, Hongyu Zhao, Yale University

3:00 **Genetic Mapping by Minimizing Integrated Square Errors**
Song Wu*, Guifang Fu, Yunmei Chen and Rongling Wu, University of Florida

3:15 **Multilocus Estimation of the Recombination Fraction, Outcrossing Rate and Linkage Disequilibrium in Wild Populations**
Wei Hou* and Jiahan Li, University of Florida, Kun Han, Zhejiang Forestry University, Song Wu, University of Florida, Yanchun Li, Zhejiang Forestry University, Rongling Wu, University of Florida

44. CONTRIBUTED PAPERS:CLINICAL TRIAL ADAPTIVE DESIGN AND RANDOMIZATION *(Potomac 3)*
Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Thomas Braun, University of Michigan

1:45 **An Efficient Clinical Trial Design Investigating Treatment by Covariate Interaction**
Ayanbola O. Elegbe*, Bristol Myers Squibb, David T. Redden, University of Alabama at Birmingham

2:00 **Estimation of Treatment Difference in Proportions in Clinical Trials With Blinded Sample Size Re-estimation**
Xiaohui Luo*,Merck & Co., Peng-Liang Zhao, Sanofi-Aventis

2:15 **Nonlinear Bayesian Prediction Model**
Haoda Fu*, and David Manner, Eli Lilly & Company

2:30 **Internal Pilot with Interim Analysis for Single Degree of Freedom Hypothesis Tests**
John A. Kairalla* and Keith E. Muller, University of Florida, Christopher S. Coffey, University of Alabama at Birmingham

2:45 **Optimal and Adaptive Designs in Dose Ranging Studies Based on Both Efficacy and Safety Responses**
Olga V. Marchenko*, i3 Statprobe

3:00 **Adaptive Designs for Dose Finding Based on the Emax Model**
S. Krishna Padmanabhan*, Wyeth, Francis Hsuan, Temple University, Vladimir Dragalin, Wyeth

3:15 **Optimal Adaptive Designs for Binary Response Trials**
Youngsook Jeon* and Feifang Hu, University of Virginia

45. CONTRIBUTED PAPERS:GENOME-WIDE ASSOCIATION STUDIES *(Potomac 4)*
Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Eleanor Feingold, University of Pittsburgh

1:45 **A Bayesian Approach for Incorporating Prior Knowledge in Genome-wide Association Studies**
Haojun Ouyang* and Jung-Ying Tzeng, North Carolina State University

2:00 **Stepwise Forward Multiple Regression for Complex Traits in High Density Genome-wide Association Studies**
Xiangjun Gu*, Christopher I. Amos and Gary Rosner, M.D. Anderson Cancer Center, Mary Relling, St. Jude Children's Research Hospital, Ralph F. Frankowski, University of Texas School of Public Health at Houston

2:15 **Probability of Detecting Disease-Associated SNPs in Case-Control Genome-Wide Association Studies**
Ruth Pfeiffer* and Mitchell Gail, National Cancer Institute

2:30 **Shrinkage Estimation for Robust and Efficient Screening of single SNP Association from Case-Control Genome-wide Association Studies**
Sheng Luo*, Johns Hopkins University, Nilanjan Chatterjee, National Cancer Institute, Bhramar Mukherjee, University of Michigan

2:45 **Genome-wide Association Studies with Related Individuals**
Weihua Guan*, Liming Liang, Michael Boehnke and Gonçalo R. Abecasis, University of Michigan

3:00 **A Mixture of Expert Model for Population Stratification in Genomic Association Study**
Yulan Liang*, University at Buffalo, The State University of New York

3:15 **A Hierarchical Bayesian Model for Genome-wide Association Studies of SNPs with Missingness**
Zhen Li* and George Casella, University of Florida

46. CONTRIBUTED PAPERS:SPATIAL MODELING APPLICATIONS *(Potomac 5)*
Sponsors: ENAR and ASA Section on Statistics in Defense and National Security
Chair: Andrew Lawson, University of South Carolina

1:45 **Repeated Measures Methodology for Spatial Cluster Detection while Accounting for Moving Locations**
Andrea J. Cook*, University of Washington, Diane R. Gold, Brigham and Women's Hospital and Harvard University, Yi Li, Harvard University and The Dana Farber Cancer Institute

2:00 **Seasonal/Regional Effect Estimates for PM2.5 and Hospital Admissions Rates**
Keita Ebisu* and Michelle L. Bell, Yale University, Roger D. Peng and Francesca Dominici, Johns Hopkins University

2:15 **Models for Spatial Bivariate Binary Data**
Petrutza C. Caragea and Emily J. Berg*, Iowa State University

2:30 **Weighted Normal Spatial Scan Statistic for Heterogeneous Population Data**
Lan Huang*, Information Management Services, Inc., Ram C. Tiwari, National Cancer Institute, Zhaohui Zou, Information Management Services, Inc., Martin Kulldorff, Harvard University and Harvard Pilgrim Health Care, Eric J. Feuer, National Cancer Institute

2:45 **Spatio-Temporal Model for Irregularly Spaced Aerosol Optical Depth Data**
Jacob J. Oleson* and Naresh Kumar, University of Iowa

3:00 **Introducing the S-value: An Exploratory Tool for Detecting Spatial Dependence on a Lattice**
Petrutza C. Caragea* and Mark S. Kaiser, Iowa State University

3:15 **Spatial Processes with Stochastic Heteroscedasticity**
Wenying Huang*, Ke Wang and Frank J. Breidt, Colorado State University, Richard A. Davis, Columbia University

47. CONTRIBUTED PAPERS:MISSING OR INCOMPLETE DATA *(Potomac 6)*
Sponsors: ASA Section on Statistics in Defense and National Security, ASA Survey Research Methods Section, ASA Section on Statistics in Epidemiology and ASA Biometrics Section
Chair: David Todem, Michigan State University

1:45 **Dynamic Graphics and Sensitivity Analysis for Dropout Data**
Edward C. Chao*, Data Numerica Institute

2:00 **New Methods for Estimating Stage Distribution in Cancer Registry Data**
Guoliang Tian* and Ming T. Tian, University of Maryland

2:15 **Weighted Estimating Equations for Longitudinal Studies with Death and Non-monotone Missing Time-dependent Covariates and Outcomes**
Michelle Shardell* and Ram R. Miller, University of Maryland

2:30 **Nonparametric Regression with Missing Outcomes Using Weighted Kernel Estimating Equations**
Lu Wang*, Xihong Lin, and Andrea Rotnitzky, Harvard University

2:45 **Marginalized Semi-Parametric Shared Parameter Models for Incomplete Ordinal Responses**
Roula Tsonaka*, Dimitris Rizopoulos, Geert Verbeke and Emmanuel Lesaffre, Catholic University-Belgium

3:00 **Information Attainable in Some Randomly Incomplete Multivariate Response Models**
Tejas A. Desai*, The Indian Institute of Management at Ahmedabad, Pranab K. Sen, The University of North Carolina at Chapel Hill

3:15 **When NMAR is Almost MAR**
Yan Zhou*, Roderick Little and John D. Kalbfleisch, University of Michigan

Monday, March 17

3:30—3:45 p.m. **Refreshment Break and Visit the Exhibitors**

3:45—5:30 p.m.

48. FDA ADVISORY COMMITTEES: A STATISTICIAN'S ROLE IN DECIDING PUBLIC POLICY (A PANEL DISCUSSION) *(Regency A)*
Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology
Organizer: Alvin Van Orden, U.S. Food and Drug Administration
Chair: Scott Evans, Harvard University

An Introduction to FDA Advisory Committees
Gregory Campbell*, U. S. Food and Drug Administration
Role of an Industry Statistician in an FDA Advisory Committee Meeting
Frank W. Rockhold*, GlaxoSmithKline R&D
A Statistician's Role in Food and Drug Administration Advisory Committees
David A. Schoenfeld*, Massachusetts General Hospital
Seeking Advice from Statisticians: A Clinician's Perspective
Celia M. Witten*, U.S. Food and Drug Administration

49. NEW DIRECTIONS IN SAFETY PLANNING AND ANALYSIS FOR CLINICAL DEVELOPMENT *(Regency B)*
Sponsors: ASA Biopharmaceutical Section and ENAR
Organizers: Brenda Crowe, Eli Lilly & Co. and Eileen King, Procter & Gamble Co.
Chair: Eileen King, Procter & Gamble Co.

3:45 **Regulatory Perspectives on Planning for Pre-marketing Safety**
George C. Rochester*, U.S. Food and Drug Administration

4:10 **Planning for Meta-analysis**
Jesse A. Berlin*, Johnson & Johnson Pharmaceutical Research and Development

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

4:35 **Analysis of Clinical Adverse Event Data Using False Discovery Rate Methods**
Devan V. Mehrotra* and Joseph F. Heyse, Merck Research Laboratories

5:00 **Detecting Safety Signals in Clinical Trials: A Bayesian Perspective**
H. Amy Xia* and Haijun Ma, Amgen, Inc.

5:25 **Floor Discussion**

50. VALIDATION OF GENOMIC CLASSIFIERS *(Regency C)*

Sponsors: ENAR and ASA Biopharmaceutical Section
Organizer: Gene Pennello, U.S. Food and Drug Administration
Chair: Lakshmi Vishnuvajjala, U.S. Food and Drug Administration

3:45 **The Microarray Quality Control Consortium: Experiences in the Development and Validation of Genomic Predictive Models for Clinical and Toxicogenomic Data Sets**
Russell D. Wolfinger*, SAS Institute Inc.

4:10 **Developing and Validating Genomic Classifiers**
Kevin K. Dobbin*, National Cancer Institute

4:35 **Gene Expression Profiling Devices for Cancer Prognosis: FDA Clearance Process**
Reena Philip*, U.S. Food and Drug Administration

5:00 **Drug-Diagnostic Co-development: Simultaneous Validation of a New Therapeutic and a Pharmacogenetic Diagnostic Biomarker for Selecting Patients Most Likely Benefit from It**
Gene A. Pennello*, U.S. Food and Drug Administration

5:25 **Floor Discussion**

51. METHODS FOR ANALYZING MULTI-READER RECEIVER OPERATING CHARACTERISTIC (ROC) DATA *(Regency D)*

Sponsors: ENAR, ASA Section on Statistics in Epidemiology and ASA Biometrics Section
Organizer: Stephen Hillis, Iowa City VA Medical Center
Chair: Michael Berbaum, University of Chicago-Illinois

3:45 **Using Marginal ANOVA Models to Motivate, Generalize, and Derive Properties for the Obuchowski-Rockette Procedure for Multi-Reader ROC Data Analysis**
Stephen L. Hillis*, Iowa City VA Medical Center

4:15 **Ensemble Variance for Binormal Model of MRMC AUC**
Brandon D. Gallas*, U.S. Food and Drug Administration

4:45 **Comparisons of Methods for Analysis of Multi-reader Multi-modality ROC Studies**
Xiao-Hua A. Zhou*, University of Washington, VA Puget Sound Health Care System

5:15 **Floor Discussion**

52. RECENT ADVANCES IN THE MODELING OF COMPETING RISKS *(Washington A)*

Sponsors: ENAR, ASA Biometrics Section and ASA Section on Defense and National Security
Organizer: Abdus Wahed, University of Pittsburgh
Chair: Abdus Wahed, University of Pittsburgh

3:45 **Cause-Specific Relative Risk Models for Estimating Absolute Risk**
Mitchell H. Gail*, National Cancer Institute

4:15 **Parametric Inference on Competing Risks Data**
Jong-Hyeon Jeong*, University of Pittsburgh

4:45 **Quantile Inference for Competing Risks Data**
Jason P. Fine*, University of Wisconsin, Madison

5:15 **Discussant:**
John P. Klein, Medical College of Wisconsin

53. ANALYSIS OF MARKED POINT PATTERNS WITH SPATIAL AND NONSPATIAL COVARIATE INFORMATION *(Washington B)*

Sponsors: IMS, ASA Section on Statistics in Defense and National Security and ASA Section on Statistics in Epidemiology
Organizers: Bradley P. Carlin and Sudipto Banerjee, University of Minnesota
Chair: Alan F. Gelfand, Duke University

3:45 **Analysis of Marked Point Patterns with Spatial and Nonspatial Covariate Information**
Shengde Liang and Bradley P. Carlin*, University of Minnesota, Alan E. Gelfand, Duke University

4:15 **Bayesian Spatial Scan Statistic Adjusted for Overdispersion and Spatial Correlation.**
Deepak Agarwal*, Yahoo! Research

4:45 **Estimating Functions for Inhomogeneous Spatial Point Processes with Incomplete Covariate Data**
Rasmus P. Waagepetersen*, Aalborg University

5:15 **Discussant:**
Sudipto Banerjee, University of Minnesota

54. CONTRIBUTED PAPERS:STATISTICAL INFERENCE *(Potomac I)*

Sponsors: ENAR and ASA Section on Statistics in Epidemiology
Chair: Pamela Shaw, National Institute of Allergy and Infectious Diseases

3:45 **Approximate P-Values From Monte Carlo Hypothesis Testing**
Allyson M. Abrams*, Martin Kulldorff and Ken P. Kleinman, Harvard Pilgrim Health Care and Harvard University

4:00 **An Application of Exact Tests using Two Correlated Binomial Variables in Clinical Trials**
Jihnhee Yu*, University at Buffalo, James L. Kepner, American Cancer Society

4:15 **Improving Efficiency of Inferences in Randomized Clinical Trials using Auxiliary Covariates**
Min Zhang*, Anastasios, A. Tsiatis and Marie Davidian, North Carolina State University

4:30 **Inference Following an Adaptive Design - Practical Issues Arising From a Large Study in Oncology**
Lothar T. Tremmel*, Cephalon

4:45 **Modeling Infectivity Rates and Attack Windows for Two Viruses**
Jian Wu, A. John Bailer* and Stephen E. Wright, Miami University

5:00 **Comparing Statistical Interval Estimates**
Michelle Quinlan*, University of Nebraska-Lincoln, James Schwenke, Boehringer Ingelheim Pharmaceuticals, Inc., Walt Stroup, University of Nebraska-Lincoln

5:15 **Exact Bayesian Inference in 2 by 2 Contingency Tables**
Yong Chen* and Sining Chen, Johns Hopkins University, Haitao Chu, The University of North Carolina at Chapel Hill

55. CONTRIBUTED PAPERS:STATISTICAL METHODS- GENERAL *(Potomac 2)*

Sponsor: ENAR
Chair: Sumithra Mandrekar, Mayo Clinic

3:45 **A Novel Moment-based Dimension Reduction Approach in Multivariate Regression**
Jae Keun Yoo*, University of Louisville

4:00 **One-Sided Coverage Intervals for a Proportion Estimated from a Stratified Simple Random Sample**
Phillip S. Kott*, National Agricultural Statistics Service, U.S. Department of Agriculture

4:15 **Robust Estimation of the Spectral Envelope**
Mark A. Gamalo*, University of Missouri - Kansas City

4:30 **Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models**
Sonja Greven*, Ludwig-Maximilians-Universität Munich, Ciprian M. Crainiceanu, Johns Hopkins University, Helmut Küchenhoff, Ludwig-Maximilians-Universität Munich, Annette Peters, GSF-National Research Center for Environment and Health

4:45 **One-Sided Tests and Confidence Bounds for the Difference between Probabilities for Matched Pairs Dichotomous Data**
Donald J. Schuirmann*, U.S. Food and Drug Administration

5:00 **Spatial Models with Applications in Computer Experiments**
Ke Wang*, Wenying Huang and Frank J. Breidt, Colorado State University, Richard A. Davis, University of Columbia

5:15 **Graphical Tool for Bayesian Network Reconstruction**
Peter Salzman* and Anthony Almudevar, University of Rochester

56. CONTRIBUTED PAPERS:CLINICAL TRIALS DESIGN

(Potomac 3)
Sponsors: ENAR, ASA Section on Statistics in Epidemiology and ASA Biopharmaceutical Section
Chair: Rickey Carter, Medical University of South Carolina

3:45 **Design and Interim Monitoring of a Clinical Trial based on the Gompetz Distribution**
Arzu Onar*, Robert P. Sanders, Amar Gajjar and James M. Boyett, St Jude Children's Research Hospital

4:00 **Optimal Cost-effective Design Strategies in Late-stage Clinical Trials**
Cong Chen* and Robert A. Beckman, Merck & Co., Inc.

4:15 **Resolving a Clinical Trial When Accrual has Slowed or Stopped**
David N. Stivers* and Scott M. Berry, Berry Consultants, Donald A. Berry, University of Texas, M. D. Anderson Cancer Center

4:30 **The Influence of Different Designs for Survival Trials on the Type I Error Rate of the Logrank Statistic**
Jitendra Ganju* and Julia Ma, Amgen, Inc.

4:45 **Optimal Enrichment Strategies for the Randomized Discontinuation Design**
Peter Müller, Gary L. Rosner and Lorenzo Trippa*, The University of Texas, M.D. Anderson Cancer Center

5:00 **Crossover Designs for Mortality Trials**
Martha C. Nason* and Dean Follmann, National Institute of Allergy and Infectious Diseases

5:15 **A Comparison of Test Statistics for the Sequential Parallel Design**
Xiaohong Huang, Sanofi-Aventis Corporation, Roy N. Tamura*, Eli Lilly and Company

57. CONTRIBUTED PAPERS:STATISTICAL METHODS FOR GENOMICS *(Potomac 4)*

Sponsors: ENAR, ASA Biometrics Section and ASA Biopharmaceutical Section
Chair: Alina Andrei, University of Wisconsin-Madison

3:45 **Harnessing Naturally Randomized Transcription to Infer Causal Regulatory Relationships Among Genes**
Lin S. Chen*, Frank Emmert-Streib and John D. Storey, University of Washington

4:00 **Incorporating Gene Networks into Statistical Tests for Genomic Data**
Peng Wei* and Wei Pan, University of Minnesota

4:15 **Prediction of Survival Time using Gene Expression Profiles of Multiple Myeloma Patients**
Jimin Choi*, Jimmy Fulp and Dan Sullivan, H. Lee Moffitt Cancer Center & Research Institute, Choongrak Kim, Pusan National University

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Monday, March 17 *continued*

4:30 **Analysis of Gene Sets Based on the Underlying Regulatory Network**
Ali Shojaie* and George Michailidis, University of Michigan

4:45 **A Model-based Approach for Combining Heterogeneous High Throughput Genomic Data to Improve Detection of Regulatory Motifs**
Heejung Shim*, Adam Hinz and Sunduz Keles, University of Wisconsin-Madison

5:00 **Nonparametric Meta-analysis for Identifying Signature Genes in the Integration of Multiple Genomic Studies**
Jia Li* and George C. Tseng, University of Pittsburgh

5:15 **A Genome-wide Study of Nucleosome Free Region in Yeast**
Wei Sun*, Wei Xie, Feng Xu, Michael Grunstein and Ker-Chau Li, University of North Carolina

58. CONTRIBUTED PAPERS:SURVIVAL ANALYSIS - NONPARAMETRIC METHODS *(Potomac 5)*

Sponsors: ENAR and ASA Biometrics Section
Chair: Suddhasatta Acharyya, Brown University

3:45 **Bezier Curve Smoothing of the Kaplan-Meier Estimator**
Choongrak Kim*, Eunyoung Yun and Mina Baek, Pusan National University-Korea

4:00 **Nonparametric Estimation of State Occupation, Entry and Exit Times with Multistate Current Status Data**
Ling Lan* and Somnath Datta, University of Louisville

4:15 **Using the Seminonparametric Density to Estimate Survival Functions in the presence of Censored Data**
Kirsten Doehler*, University of North Carolina at Greensboro, Marie Davidian, North Carolina State University

4:30 **Estimation for Censored Heteroscedastic Linear Regression**
Xuewen Lu* and Zhulin He, University of Calgary

4:45 **Local Linear Estimation of Conditional Hazard Function**
Jinmi Kim*, Hyunmi Hwang and Choongrak Kim, Pusan National University-Korea

5:00 **Examining Model Fit for Penalized Splines: A Simulation Study**
Elizabeth J. Malloy*, American University, Donna Spiegelman, Harvard University, Ellen A. Eisen, Harvard University and University of California-Berkeley

5:15 **Inference for a Change-point Cox Model with Current Status Data**
Rui Song* and Michael R. Kosorok, University of North Carolina, Shuangge Ma, Yale University

59. CONTRIBUTED PAPERS:FUNCTIONAL DATA ANALYSIS *(Potomac 6)*

Sponsor: ENAR
Chair: Brisa Sanchez, University of Michigan

3:45 **Functional Mixed Registration Models**
Donatello Telesca*, University of Texas, M.D. Anderson Cancer Center, Lurdes YT Inoue, University of Washington

4:00 **Principal Differential Analysis: Estimating Coefficients and Linearly Independent Solutions of a Linear Differential Operator with Covariates for Functional Data**
Seoweon Jin*, Southern Methodist University, Joan G. Staniswalis, University of Texas at El Paso, Anu Sharma, University of Colorado at Boulder

4:15 **Identifying Temporally Differentially Expressed Genes through Functional Principal Components Analysis**
Xueli Liu* and Mark C.K. Yang, University of Florida

4:30 **Functional Sliced Inverse Regression in Lipoprotein Profile Data**
Yehua Li*, University of Georgia, Tailen Hsing, University of Michigan

4:45 **A Generalized Akaike's Information Criterion for Selecting Penalty Function in Spatially Adaptive Smoothing Splines**
Ziyue Liu* and Wensheng Guo, University of Pennsylvania

5:00 **Bootstrapping Sums of Independent but not Identically Distributed Continuous Processes with Applications to Functional Data**
Chung Chang* and Robert T. Ogden, Columbia University

5:15 **Directed Spatio-temporal Monitoring of Disease Incidence Rates**
Dan J. Spitzner*, University of Virginia, Brooke Marshall, Virginia Tech University

Tuesday, March 18

8:30—10:15 a.m.

60. ADAPTIVE DESIGNS: PERSPECTIVES FROM ACADEMIA, INDUSTRY AND REGULATORY *(Regency A)*

Sponsors: ASA Biopharmaceutical Section and ENAR
Organizers: Ghanshyam Gupta and Weishi Yuan, U.S. Food & Drug Administration
Chair: Aiyi Liu, National Institute of Child Health and Human Development

8:30 **Adaptive Designs: Why, How and When?**
Christopher Jennison*, University of Bath-U.K.

9:00 **Design and Implementation of Adaptive Trials: Experiences of an Industry Consultant**
Cyrus R. Mehta*, Cytel Inc.

9:30 **Adaptability of Clinical Trial Design: A Larger Context**
H.M. James Hung*, U.S. Food and Drug Administration
10:00 **Discussant:** Ralph D'Agostino, Boston University

61. THE EDUCATION OF CLINICAL TRIAL STATISTICIANS: DO WE NEED TO CHANGE WITH THE TIMES? *(Regency B)*

Sponsors: ASA Section on Teaching Statistics in the Health Sciences, ASA Statistical Education Section, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology
Organizer/Chair: Scott Evans, Harvard School of Public Health
Panelists: David L. DeMets, University of Wisconsin
Naitee Ting, Pfizer Inc., Global R&D
Karl E. Peace, Georgia Southern University
Walter Offen, Eli Lilly & Company
Marvin Zelen, Harvard University

62. CRITICAL DESIGN AND STATISTICAL CONSIDERATIONS WITH USE OF (GENOMIC) BIOMARKER FOR PREDICTION OR OPTIMIZING THERAPY *(Regency C)*

Sponsors: ENAR and ASA Biophamaceutical Section
Organizer: Sue-Jane Wang, U.S. Food and Drug Administration
Chair: Sue-Jane Wang, U.S. Food and Drug Administration

8:30 **Prediction of Clinical Outcomes with High Dimensional Markers**
Tianxi Cai*, Harvard University, Lu Tian and Jie Huang, Northwestern University, Samuel McDaniel and Rebecca Betensky, Harvard University
8:55 **Use of Clinical Reclassification to Assess Risk Prediction Models**
Nancy R. Cook*, Brigham & Women's Hospital and Harvard University
9:20 **Uncertainties in the Multiple-Biomarker Classifier Problem and a Proposed Strategy for Study Design and Analysis**
Robert F. Wagner* and Weijie Chen, U.S. Food and Drug Administration, Waleed A. Yousef, Helwan University-Cairo, Egypt
9:45 **PREDICT-1: The First Powered, Prospective, Randomised Trial of Pharmacogenetic Screening to Reduce Drug Adverse Events**
Sara H. Hughes*, GlaxoSmithKline
10:10 **Floor Discussion**

63. ADVANCED SEMIPARAMETRIC MODELING OF GENETICS/GENOMIC DATA *(Regency D)*

Sponsors: IMS, ASA Biometrics Section and ASA Biopharmaceutical Section
Organizer: Naisyin Wang, Texas A&M University
Chair: Naisyin Wang, Texas A&M University

8:30 **Semiparametric Methods for Case-Control Association Studies**
Danyu Lin*, University of North Carolina
8:55 **Biomarker Discovery for Genomics and Proteomics Data Using Functional Mixed Models**
Jeffrey S. Morris*, The University of Texas-M.D. Anderson Cancer Center
9:20 **Testing the Significance of Cell-Cycle Patterns in Time-Course Microarray Data**
Guei-Feng Tsai, Center for Drug Evaluation-Taipei, Taiwan , Annie Qu*, Oregon State University
9:45 **A Forest-based Approach to Identifying Gene and Gene-gene Interactions**
Xiang Chen, Ching-Ti Liu, Meizhuo Zhang and Heping Zhang*, Yale University
10:10 **Floor Discussion**

64. THE USE OF SURVEY SAMPLES IN EPIDEMIOLOGICAL FOLLOW-UP STUDIES *(Washington A)*

Sponsors: ENAR, ASA Section on Statistics in Epidemiology and ASA Survey Research Methods Section
Organizer: Barry Graubard, National Cancer Institute
Chair: Barry Graubard, National Cancer Institute

8:30 **The Hispanic Community Health Study**
Lisa LaVange, University of North Carolina Chapel Hill
8:55 **The NCS: Establishment and Protection of the Inferential Base**
Jonas H. Ellenberg*, University of Pennsylvania
9:20 **Discussant:**
Roderick J. Little, University of Michigan
9:35 **Discussant:**
Sholom Wacholder, National Cancer Institute
9:50 **Discussant:**
William Kalsbeek, University of North Carolina
10:05 **Floor Discussion**

65. ANALYSIS OF RECURRENT MARKER DATA *(Washington B)*

Sponsors: ENAR and ASA Biometrics Section
Organizer: Lei Liu, University of Virginia
Chair: Xuelin Huang, M.D. Anderson Cancer Center

8:30 **Forward and Backward Recurrent Marker Processes**
Mei-Cheng Wang*, Johns Hopkins University
9:00 **Analysis of Longitudinal Data in the Presence of Informative Observational Times**
Lei Liu*, University of Virginia, Xuelin Huang, M.D. Anderson Cancer Center, John O' Quigley, University of Virginia
9:30 **Regression Analysis of Longitudinal Data with Dependent Observation Process**
(Tony) Jianguo Sun* and Liuquan Sun, University of Missouri, Dandan Liu, University of Michigan
10:00 **Floor Discussion**

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Scientific Program

Tuesday, March 18 *continued*

66. CONTRIBUTED PAPERS: JOINT MODELING

(Potomac I)
Sponsor: ENAR
Chair: Ji-Hyun Lee, H. Lee Moffitt Cancer Center & Research Institute

8:30 **Predicting Outcomes from Later-Stage Longitudinal Trials with Potential Withdrawals Using Findings from Early Stage Trials : A Case History**
A. Lawrence Gould* and Kenneth Liu, Merck Research Laboratories

8:45 **A Bayesian Approach to Clustering Trajectories in the Presence of Multiple Changepoints**
Pulak Ghosh, Georgia State University, Kaushik Ghosh*, University of Nevada-Las Vegas, Ram C. Tiwari, National Cancer Institute

9:00 **Hierarchical models for spatially-referenced longitudinal data with applications to crop growth**
Haowen Cai*, Sudipto Banerjee, Ryan T. Thelemann and Gregg A. Johnson, University of Minnesota

9:15 **Incorporating Libido into Human Fecundability Models**
Kirsten Lum*, National Institutes of Health and American University, Rajeshwari Sundaram and Germaine Buck Louis, National Institutes of Health

9:30 **Joint Modeling of Cluster Size and Clustered Failure Times**
Hanna Yoo*, Yang-Jin Kim, Jae Won Lee and Shin-Jae Lee, Korea University

9:45 **Semiparametric Analysis of Panel Count Data with Correlated Observation and Follow-up Times**
Xin He*, The Ohio State University, Xingwei Tong, Beijing Normal University-Peoples Republic of China, Jianguo Sun, University of Missouri

10:00 **Latent Class Transition Models for Chronic Disability Survey Data with Censoring and Staggered Entry**
Toby A. White*, University of Washington

67. CONTRIBUTED PAPERS:APPLICATIONS IN CAUSAL INFERENCE *(Potomac 2)*

Sponsors: ENAR, ASA Health Policy Statistics Section and ASA Section on Statistics in Epidemiology
Chair: Lan Huang, National Cancer Institute

8:30 **Identification of Causal Treatment Effects using Reference Stratification**
Booil Jo*, Stanford University

8:45 **Estimation and Inferenece for the Causal Effect of Receiving Treatment on a Multinominal Outcome**
Jing Cheng*, University of Florida College of Medicine

9:00 **Estimation of Treatment Efficacy in Randomized Clinical Trials Which Involve Non-trial Departures**
Tingting Ge* and Stanley P. Azen, University of Southern California

9:15 **Bounds on ATE and Unmeasured Confounding Bias in Observational Studies**
Tao Liu*, Joseph W. Hogan and Allison K. DeLong, Brown University

9:30 **Direct and Indirect Effects for Clustered and Longitudinal Data**
Tyler J. VanderWeele*, University of Chicago

9:45 **Correcting for Survivor Treatment-Selection Bias with a Structural Failure Time Model:Survival in Oscar Award Winning Performers**
Xu Han*, Dylan Small, Dean Foster and Vishal Patel, University of Pennsylvania

10:00 **A Systematic Review of Propensity-score Matching in the Medical Literature from 1996 to 2003**
Peter C. Austin*, Institute for Clinical Evaluative Sciences

68. CONTRIBUTED PAPERS:PROTEOMICS AND GENOMICS *(Potomac 3)*

Sponsors: ENAR and ASA Biometrics Section
Chair: Mimi Y. Kim, Albert Einstein College of Medicine

8:30 **Finding the Optimum Number of Probes to Interrogate Transcripts Using Affymetrix GeneChip Technology**
Fenghai Duan*, University of Nebraska

8:45 **Comparison of Resampling Methods on Validation of the Procedure for Constructing a Survival Prediction Model for Proteomic Studies**
Heidi Chen* and Ming Li, Vanderbilt University, Dean Billheimer, University of Utah, Yu Shyr, Vanderbilt University

9:00 **Quality Assessment and Reproducibility Analysis of Mass Spectrometry Data**
Shuo Chen*, Emory University

9:15 **A Novel Wavelet-based Approach for the Pre-processing of Mass Spectrometry Data**
Deukwoo Kwon*, National Cancer Institute, Joon Jin Song, University of Arkansas, Jaesik Jeong, Texas A&M University, Ruth M. Pfeiffer, National Cancer Institute, Marina Vannucci, Rice Univeristy

9:30 **Testing for Treatment Effects on Gene Ontology**
Taewon Lee* and Varsha G. Desai, National Center for Toxicological Research, Robert R. Delongchamp, University of Arkansas, Cruz Velasco, Louisiana State University

9:45 **A Model for the Analysis of Proteolytic 18O Stable-isotope Labeled Peptides in MALDI-TOF Mass-spectra**
Dirk Valkenborg and Tomasz Burzykowski*, Hasselt University-Belgium

10:00 **Background Correction, Normalization and Summaries of Bead Level Expression Data from Illumina Beadarray**
10:00 Yang Xie*, Guanghua Xiao, Lianghao Ding, University of Texas, Jeff Allen, Southern Methodist University, Michael D. Story, University of Texas

69. CONTRIBUTED PAPERS:SAMPLE SIZE *(Potomac 4)*
Sponsors: ENAR, ASA Section on Statistics in Epidemiology, ASA Biopharmaceutical Section
Chair: Mary Putt, University of Pennsylvania

8:30 **Sample Sizes for Pilot Studies**
Rickey E. Carter* and Robert F. Woolson, Medical University of South Carolina

8:45 **Evaluation of Ten Events Per Covariate Recommendation for Cox Proportional Hazards Regression Models**
Mehmet Kocak* and Arzu Onar, St. Jude Children's Research Hospital, Seok P. Wong, University of Memphis

9:00 **Design and Sample Size for Evaluating Combinations of Drugs of Linear and Loglinear Dose Response Curves**
Hongbin Fang* and Guo-Liang Tian, University of Maryland, Wei Li, University of Tennessee at Memphis, Ming Tan, University of Maryland

9:15 **Sample Size Formula for Time-to-Event Data with Incomplete Event Adjudication Using Generalized Logrank Statistic**
Shanhong Guan*, Merck & Co., Michael Kosorok, University of North Carolina-Chapel Hill, Tom Cook, University of Wisconsin-Madison

9:30 **Inverse-Probability-Weighting Based Sample Size Formula for Two-Stage Adaptive Treatment Strategies**
Abdus S. Wahed* and Wentao Feng, University of Pittsburgh

9:45 **Sample Size Calculation for a Mixture of Discrete and Continuous Endpoints:An Application of the Poisson-Binomial Distribution**
Yolanda Munoz Maldonado*, Sarah M. Baraniuk and Lemuel A. Moye, University of Texas-Houston, Schoo of Public Health

10:00 **Sample Size Calculation for Thorough QT/QTc Study**
Zhaoling Meng*, Xun Chen, Robert Kringle and Peng-Liang Zhao, Sanofi Aventis

70. CONTRIBUTED PAPERS:IMAGING ANALYSIS
(Potomac 5)
Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Mulugeta Gebregziabher, Medical University of South Carolina

8:30 **Correlation Coefficient from Data with Missing Covariates: Method Comparison**
Dana L. Tudorascu* and Lisa Weissfeld, University of Pittsburgh

8:45 **Adjusted Exponentially Tilted Empirical Likelihood with Applications to Neuroimaging Data**
Xiaoyan Shi*, Hongtu Zhu, Joseph G. Ibrahim and Martin Styner, Unversity of North Carolina at Chapel Hill

9:00 **FDR Thresholding in a Neuroimaging Context**
Lynn E. Eberly*, University of Minnesota, Brian S. Caffo, Johns Hopkins University

9:15 **A Bayesian Image Analysis of the Change in Tumor/Brain Contrast Uptake Induced by Radiation**
Xiaoxi Zhang*, Timothy D. Johnson and Roderick J. A. Little, University of Michigan

9:30 **Bayesian Spatial Modeling of fMRI data:A Multiple-Subject Analysis**
Lei Xu*, Timothy D. Johnson and Thomas E. Nichols, University of Michigan

9:45 **A Nonparametric Mixture Model for The fMRI Visual Field Map**
Raymond G. Hoffmann*, Nicholas M. Pajewski and Edward A. DeYoe, Medical College of Wisconsin

10:00 **Using Scientific and Statistical Sufficiency to Lift the Curse of Dimensionality in Imaging**
Yueh-Yun Chi* and Keith E. Muller, University of Florida

71. CONTRIBUTED PAPERS:SURVIVAL ANALYSIS - DEPENDENT CENSORING *(Potomac 6)*
Sponsors: ENAR and ASA Biometrics Section
Chair: Bo Cai, University of South Carolina

8:30 **'Smooth' Inference for Bivariate Survival Functions with Arbitrarily Censored Data**
Lihua Tang* and Marie Davidian, North Carolina State University

8:45 **A Class of Semiparametric Mixture Cure Survival Models with Dependent Censoring**
Megan Othus* and Yi Li, Harvard University, Ram Tiwari, National Cancer Institute

9:00 **Weighted Likelihood Method for Grouped Survival Data in Case-Cohort Studies, with Application to HIV Vaccine Trials**
Zhiguo Li*, University of Michigan, Peter Gilbert, Fred Hutchinson Cancer Research Center, Bin Nan, University of Michigan

9:15 **Analysis of Time-to-Event Outcomes in Neonatal Clinical Trials with Twin Births**
Michele L. Shaffer*, Penn State College of Medicine, Sasiprapa Hiriote, The Pennsylvania State University

9:30 **Semiparametric Copula Regression Model for Censored Lifetime Medical Cost**
Jing Qian* and Yijian Huang, Emory University

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Tuesday, March 18 *continued*

9:45 **Resource-Use Analyses with Death as a Confounding Factor**
Benjamin L. Trzaskoma* and Amy C. Rundle Genentech, David R Nelson, Eli Lilly & Company, Fang Xie, Novartis Vaccines-Genentech

10:00 **A Mixture-Model Approach to Bivariate Hybrid Censored Survival Data**
Suhong Zhang*, Ying J. Zhang, Kathryn Chaloner, Jack T. Stapleton, University of Iowa

Tuesday, March 18

10:15-10:30 a.m. **Refreshment Break and Visit the Exhibitors**

10:30 a.m.-12:15 p.m.

72. PRESIDENTIAL INVITED ADDRESS *(Regency Ballroom)*
Sponsor: ENAR
Organizer/Chair: Eric Feuer, National Cancer Institute

10:30 **Introduction**: Eric Feuer, National Cancer Institute

10:35 **Distinguished Student Paper Awards**
Jane F. Pendergast, University of Iowa

10:45 **Biostatistics, Science, and the Public Eye**
Donald A. Berry, University of Texas, M.D. Anderson Cancer Center

Tuesday, March 18

1:45-3:30 p.m.

73. JOINT MODELS FOR SURROGATE AND FINAL OUTCOMES *(Regency A)*
Sponsors: ASA Biometrics Section, ASA Section on Statistics in Epidemiology and ASA Biopharmaceutical Section
Organizer: Michael Elliott, University of Michigan
Chair: Michael Elliott, University of Michigan

1:45 **Assessing Surrogacy in Clinical Trials Using Counterfactual Models**
Yun Li*, Jeremy MG Taylor and Michael R. Elliott, University of Michigan

2:15 **The Meta-analytic Framework for the Evaluation of Surrogate Endpoints in Clinical Trials**
Geert Molenberghs*, Hasselt University-Belgium

2:45 **A Unified Framework for Surrogate Outcomes and Markers**
Marshall M. Joffe*, University of Pennsylvania, Tom Greene, University of Utah

3:15 **Discussant:**
Donald Rubin, Harvard University

74. STATISTICAL ANALYSIS OF LARGE-SCALE ENVIRONMENTAL DATASETS *(Regency B)*
Sponsors: ASA Section on Statistics and the Environment, ASA Section on Statistics in Defense and National Security and ASA Section on Statistics in Epidemiology
Organizer: Catherine A. Calder, The Ohio State University
Chair: Catherine A. Calder, The Ohio State University

1:45 **Cloud Height Estimation Based on Multi-Angle Satellite (MISR) Images**
Ethan B. Anderes*, University of California at Berkeley

2:10 **Nonstationary Covariance Models for Global Data**
Mikyoung Jun*, Texas A&M University, Michael Stein, University of Chicago

2:35 **Integrating Satellite and Monitoring Data to Retrospectively Estimate Monthly PM2.5 Concentrations in the Eastern United States**
Christopher J. Paciorek* and Yang Liu, Harvard University

3:00 **Dimension Reduction Approaches for Analyzing Large Spatial Datasets**
Alan E. Gelfand*, Duke University

3:25 **Floor Discussion**

75. VARIABLE SELECTION AND DIMENSION REDUCTION IN GENOMICS *(Regency C)*
Sponsors: ENAR, ASA Biometrics Section
Organizer: Sunduz Keles, University of Wisconsin-Madison
Chair: Sunduz Keles, University of Wisconsin-Madison

1:45 **Clustering with Multiple Distance Metrics - Mixture Models with Profile Transformations.**
Rebecka J. Jornsten*, Rutgers University

2:15 **Sparse Partial Least Squares Regression with an Application to Genome Scale Transcription Factor Analysis**
Hyonho Chun* and Sunduz Keles, University of Wisconsin-Madison

2:45 **Variable Selection for Varying-Coefficients Models with Applications to Analysis of Genomic Data**
Hongzhe Li* and Lifeng Wang, University of Pennsylvania

3:15 **Floor Discussion**

76. TARGETING TREATMENT EFFICACY: FLEXIBLE MODELING TO CLINICAL TRIAL DESIGN *(Regency D)*
Sponsors: IMS and ASA Biopharmaceutical Section
Organizer: Michael LeBlanc, Fred Hutchinson Cancer Research Center
Chair: Michael LeBlanc, Fred Hutchinson Cancer Research Center

1:45 **A Novel Data Driven Approach to Stage Grouping of Esophageal Cancer**
Hemant Ishwaran*, Cleveland Clinic

2:15 **Causal Subgroup Modeling in Cancer Clinical Trials**
Mary W. Redman* and Michael L. LeBlanc, Fred Hutchinson Cancer Research Center

2:45 **Moving from Correlative Clinical Science to Predictive Medicine in Clinical Trial Designs**
Richard M. Simon*, National Cancer Institute
3:15 **Floor Discussion**

77. STATISTICAL LEARNING IN BIOLOGICAL SIGNALS AND IMAGES *(Washington A)*
Sponsors: ENAR, ASA Biometrics Section and ASA Biopharmaceutical Section
Organizer: Hernando Ombao, Brown University
Chair: Hernando Omba, Brown University

1:45 **Detecting Cognitive Fatigue via Mixtures of Autoregressive Models**
Raquel Prado*, University of California-Santa Cruz
2:10 **Classification of Biomedical Signals via Multiscale Local Stationarity**
Piotr Z. Fryzlewicz*, University of Bristol, Hernando Ombao, Brown University
2:35 **Semiparametric Logistic Regression for Genetic Pathway Data: Kernel Machines and Generalized Linear Mixed Models**
Dawei Liu*, Brown University, Xihong Lin, Harvard University, Debashis Ghosh, The Pennsylvania State University
3:00 **PfCluster and FCA, and their Applications to Profile Data Analysis**
Jiayang Sun*, Case Western Reserve University, Yaomin Xu, Case Western Reserve University and Cleveland Clinic Foundation
3:25 **Floor Discussion**

78. NONPARAMETRIC REGRESSION FOR SURVIVAL ANALYSIS *(Washington B)*
Sponsors: ENAR and ASA Biometrics Section
Organizer: Zhangsheng Yu, The Ohio State University
Chair: Sijian Wang, University of Michigan

1:45 **Partially Linear Hazard Regression with Varying-coefficients**
Jianwen Cai*, University of North Carolina at Chapel Hill, Jianqing Fan, Princeton University, Jiancheng Jiang, University of North Carolina at Charlotte, Haibo Zhou, University of North Carolina at Chapel Hill
2:15 **A Partial Linear Semiparametric Additive Risks Model for Two-Stage Design Survival Studies**
Gang Li, University of California,-Los Angeles, Tong Tong Wu*, University of Maryland, College Park
2:45 **Threshold Regression Models with Semiparametric Covariate Function for Survival Data Analysis**
Zhangsheng Yu* and Mei-Ling Ting Lee, The Ohio State University
3:15 **Floor Discussion**

79. CONTRIBUTED PAPERS:CGH ARRAY AND COPY NUMBER *(Potomac 1)*
Sponsor: ENAR
Chair: Mingyao Li, University of Pennsylvania

1:45 **DNA Copy Number Abnormal Detection Based on Decomposition of Intensity Multimodal Distribution**
Aixiang Jiang*, Jirong Long, Wei Zheng and Yu Shyr, Vanderbilt University
2:00 **A Bayesian Change-point Algorithm for the Analysis of SNP-data**
Fridtjof Thomas*, University of Texas, Stanley Pounds, St. Jude Children's Research Hospital
2:15 **Tolerance Intervals from Probe-specific Mixed-models to Detect Gains and Loses using Multiplex Ligation-dependent Probe Amplification (MLPA)**
Juan R. Gonzalez*, Center for Research in Environmental Epidemiology (CREAL)-Barcelona, Spain, Josep L. Carrasco, University of Barcelona-Spain, Lluis Armengol, Center for Genomic Regulation,-Barcelona, Spain, Yutaka Yasui, University of Alberta-Canada
2:30 **Pathway Screening Analysis of Genomic Copy Number Change Data**
Stanley B. Pounds*, St. Jude Children's Research Hospital
2:45 **A Latent Class Model with Hidden Markov Dependence for Array CGH Data**
Stacia M. DeSantis*, Medical University of South Carolina, E.A. Houseman, University of Massachusetts-Lowell, Brent A. Coull, Harvard University, David N. Louis and Gayatry Mohapatra, Massachusetts General Hospital, Rebecca A. Betensky, Harvard University
3:00 **A Hybrid Method for Genomic Alteration Detection in CGH Microarray Data Analysis**
Ao Yuan, National Human Genome Center-Howard University, Wenqing He* and Juan Xiong, University of Western Ontario
3:15 **Floor Discussion**

80. CONTRIBUTED PAPERS:BIOPHARMACEUTICAL STUDIES *(Potomac 2)*
Sponsor: ENAR, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology
Chair: Sung Duk Kim, National Institute of Child Health and Human Development

1:45 **Design and Analysis of Alcohol/Benzo Interaction Studies**
Daowen Zhang*, North Carolina State University, Hui Quan and Zhaoling Meng, Sanofi-Aventis
2:00 **Assessing Drug Interaction when Data Collected at Fixed Rays**
Maiying Kong*, University of Louisville, J. Jack Lee, M.D. Anderson Cancer Center, University of Texas at Houston

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

2:15 **Quantifying Combination Drug Synergy using Nonlinear Blending**
John J. Peterson*, GlaxoSmithKline Pharmaceuticals, R&D

2:30 **Assessing Association in a Stratified Experiment**
Jason Liao* and Daniel Holder, Merck Research Laboratories

2:45 **An Algorithm with Genetic Information for Warfarin Dosing**
Kerrie Nelson*, University of South Carolina, David A. Schoenfeld, Harvard University

3:00 **Semiparametric Bayesian Inference for Phage Display Experiments**
Luis G. Leon-Novelo*, Peter Mueller, Kim-Anh Do, Renata Pasqualini, Wadih Arap, Mikhail Kolonin and Jessica Sun, University of Texas, M.D. Anderson Cancer Center

3:15 **Comparing Randomly Right Censored Samples Based on Overlap of Distributions**
Michael J. Dallas, Merck & Co., Inc.

81. CONTRIBUTED PAPERS:POWER ANALYSIS

(Potomac 3)
Sponsor: ENAR and ASA Biopharmaceutical Section
Chair: Sharon Yeatts, Medical University of South Carolina

1:45 **Detecting Qualitative Interaction among Subgroups of a Clinical Trial: A Bayesian Approach**
Emine O. Bayman*, The University of Iowa and Uludag University-Turkey, Kathryn Chaloner, The University of Iowa

2:00 **Selection of the Clinically Relevant Difference for Sample Size Determination in a Clinical Trial by Matching Bayesian and Frequentist Procedures**
Maria M. Ciarleglio*, V.A. Cooperative Studies Program

2:15 **Two-Stage Clinical Trial Designs for Multiple Endpoints with Correlated Observations**
Fei Ye* and Yu Shyr, Vanderbilt University

2:30 **Comparison of Two-Phase Analyses for Case-Control Genetic**
Gang Zheng, National Heart, Lung and Blood Institute, Mark J. Meyer*, American University, Wentian Li, Feinstein Institute for Medical Research, Manhasset, Yaning Yang*, University of Science and Technology of China, Hefei Anhui, China

2:45 **Borrowing Strength in A Simple Hierarchical Model**
Xuefeng Li*, U.S. Food and Drug Administration

3:00 **An Efficient Group Size Ratio in Trials with Multiple Dose Groups Versus a Common Control**
Jianliang Zhang*, MedImmune, Inc

3:15 **Statistical Methods for Active Extension Trials**
Zonghui Hu* and Dean Follmann, National Institute of Allergy and Infectious Diseases

82. CONTRIBUTED PAPERS:SURVIVAL ANALYSIS - CURE RATE MODELS AND RECURRENT EVENTS *(Potomac 4)*

Sponsors: ASA Section on Statistics in Epidemiology, ASA Biometrics Section and ASA Biopharmaceutical Section
Chair: Susan Murray, University of Michigan

1:45 **Conditional Mean Gap Time Estimation with Recurrent Events**
Adin-Cristian Andrei*, University of Wisconsin-Madison

2:00 **Semiparametric Cure Rate Models with Random Effects**
Guoqing Diao*, George Mason University, Guosheng Yin, The University of Texas-M. D. Anderson Cancer Center

2:15 **Optimal Goodness-of-Fit Tests For Recurrent Event Data**
Russell S. Stocker*, Mississippi State University, Edsel A. Pena, University of South Carolina

2:30 **An Accelerated Failure Time Cure Model for Time-to-Event Data with Masked Cause of Failure**
Jing J. Zhang*, Harvard University, Molin Wang, Harvard University and Dana-Farber Cancer Institute

2:45 **A Bayesian Analysis of Recurrent Events Data with Dependent Termination:An Application to Heart Transplant Problem**
Bichun Ouyang*, Medical University of South Carolina, Debajyoti Sinha, Florida State University, Joseph G. Ibrahim, University of North Carolina

3:00 **Flexible Modeling of Additive Treatment Effects on the Recurrent Event Mean in the Presence of a Terminating Event**
Qing Pan*, George Washington University, Douglas E. Schaubel, University of Michigan

3:15 **Nonparametric Method of Mixture Model with Prevalent Sampling**
Yu-Jen Cheng*, Mei-Cheng Wang and Noya Galai, Johns Hopkins University

83. CONTRIBUTED PAPERS:EXPERIMENTAL DESIGN

(Potomac 5)
Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Martha Nason, National Institute of Allergy and Infectious Diseases

1:45 **Optimal and Efficient Designs for Gompertz Regression Models**
Gang Li*, GlaxoSmithKline, Dibyen Majumdar, University of Illinois at Chicago

2:00 **Is the Continual Reassessment Method a Better Phase I Design? A Comprehensive Comparison with the Standard 3+3 Dose Escalation Scheme**
Alexia Iasonos* and Elyn Riedel, Memorial Sloan Kettering Cancer Center

2:15 **Three-Dose-Cohort Designs in Cancer Phase I Trials**
Bo Huang* and Rick Chappell, University of Wisconsin-Madison

2:30 **Internal Pilot Designs for Observational Studies**
Matthew J. Gurka*, University of Virginia, Christopher S. Coffey, University of Alabama at Birmingham

2:45 **Bayesian Adaptively Randomized Clinical Trials in the Presence of Patient Heterogeneity**
J. Kyle Wathen*, Mark F. Munsell and Marcos J. de Lima, University of Texas-M.D. Anderson Cancer Center

3:00 **Rank-based Analysis of Crossover Trials for Classic and Novel Designs**
Mary Putt*, University of Pennsylvania

3:15 **Implementing Optimal Allocation for Sequential Continuous Responses with Multiple treatments**
Hongjian Zhu* and Feifang Hu, University of Virginia

84. CONTRIBUTED PAPERS:ROC/ DIAGNOSTIC METHODS *(Potomac 6)*
Sponsors: ENAR and ASA Biometrics Section
Chair: Kyungsook Kim, U.S. Food and Drug Administration

1:45 **Guessing Free-response Process and Empirical FROC Curve**
Andriy Bandos*, Howard E. Rockette, David Gur and Tao Song, University of Pittsburgh

2:00 **ROC Based Utility Function Maximization for Feature Selection and Classification**
Zhenqiu Liu*, University of Maryland Medicine

2:15 **Assessing Agreement Among Acupuncturists in TEAMSI Trial**
Anna TR Legedza*, Vertex Pharmaceuticals Inc., Roger B. Davis, Beth Israel Deaconess Medical Center and Harvard University

2:30 **Estimate Benefits Due to Fecal Occult Blood Test for Colorectal Cancer Screening**
Dongfeng Wu*, University of Louisville, Diane Erwin, Information Management Services, Inc., Gary L. Rosner and Lyle D. Broemeling, University of Texas, M.D. Anderson Cancer Center

2:45 **Improving the Efficiency of Testing Predictive Values using Auxiliary Covariate**
Yoonjin Cho*, North Carolina State University, Kosinski Andrezj, Duke univeristy

3:00 **A Transformation-invariant Monotone Smoothing of Receiver Operating Characteristic Curves**
Liansheng Tang, George Mason University, Pang Du*, Virginia Tech University

3:15 **A Generalized Nonparametric Approach of Comparing FROC Systems**
Tao Song*, Andriy Bandos, Howard Rockette and David Gur, University of Pittsburgh

Tuesday, March 18

3:30—3:45 p.m **Refreshment Break and Visit the Exhibitors**

3:45—5:30 p.m

85. IMS MEDALLION LECTURE *(Regency A)*
Sponsor: IMS
Organizer: Ram Tiwari, National Cancer Institute
Chair: Ram Tiwari, National Cancer Institute

3:45 **Statistical Challenges in Genetic Association Studies**
Mary Sara McPeek, University of Chicago

86. ON FUTILITY CALCULATIONS IN RANDOMIZED CLINICAL TRIALS *(Regency B)*
Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology
Organizer: Janet Wittes, Statistics Collaborative
Chair: Judith Goldberg, New York University

3:45 **Practical Aspects of Some Common Futility Rules**
Boris Freidlin*, National Cancer Institute

4:15 **Formulating and Selecting Futility Rules in a Long-Term Trial**
Christy Chuang-Stein*, Michael Brown and Wayne Ewy, Pfizer Inc., Cathie Spino, University of Michigan

4:45 **Conditional Power Considerations in the Design of a Phase 3 Microbicide Trial in Africa**
Jennifer Schumi*, Statistics Collaborative, Inc., Zeda Rosenberg, International Partnership for Microbicides, Stephanie Dickinson, Indiana Statistical Consulting Center, Janet Wittes, Statistics Collaborative, Inc.

5:15 **Discussant:**
Donald A. Berry, University of Texas, M.D. Anderson Cancer Center

87. STATISTICAL METHODS FOR DETECTING COPY NUMBER VARIATION *(Regency C)*
Sponsor: ENAR
Organizer: Mahlet Tadesse, Georgetown University
Chair: Mahlet Tadesse, Georgetown Universtiy

3:45 **Detection of Copy Number Variations From High-density SNP Arrays: An Integrated Bayesian Hidden Markov Model Approach Incorporating Pedigree Information**
Zhen Chen*, University of Pennsylvania, Mahlet Tadesse, Georgetown University, Kai Wang and Mingyao Li, University of Pennsylvania

4:15 **Estimating Genome-wide Copy Number Using Allele Specific Mixture Models**
Shin Lin*, Benilton Caravalho, Wenyi Wang, Aravinda Chakravarti, and Rafael Irizarry, Johns Hopkins University

4:45 **Circular Binary Segmentation for the Analysis of Array CGH Data**
Venkatraman E. Seshan*, Columbia University, Adam Olshen, Memorial Sloan Kettering Cancer Center

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

5:15 **Discussant:**
Mingyao Li, University of Pennsylvania

88. CHALLENGES IN LARGE, SIMPLE TRIALS *(Regency D)*

Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology
Organizer: Jesse A. Berlin, Johnson & Johnson Pharmaceutical Research and Development
Chair: Jesse A. Berlin, Johnson & Johnson Pharmaceutical Research and Development

3:45 **Introduction to Large, Simple Trials: What are They and When Should They be Done (or Not Done)?**
Julie E. Buring*, Harvard University

4:15 **Practical Challenges to Keeping the Simple in Large, Simple Studies**
James D. Neaton*, University of Minnesota

4:45 **On Sequential Monitoring and Inference of Safety Endpoints**
Qing Liu*, Johnson & Johnson Pharmaceutical Research and Development

5:15 **Discussant:**
Susan Ellenberg, University of Pennsylvania

89. MULTISTATE APPROACH TOWARDS MODELING COMPLEX BIOMEDICAL DATA *(Washington A)*

Sponsors ENAR, ASA Biometrics Section and ASA Biopharmaceutical Section
Organizer: Rajeshwari Sundaram, National Institute of Child Health and Human Development
Chair: Rajeshwari Sundaram, National Institute of Child Health and Human Development

3:45 **Regression Analysis for Multistate Models Based on a Pseudo-value Approach, with Applications to Bone Marrow Transplantation Studies**
John P. Klein*, Medical College of Wisconsin

4:15 **Nonparametric Estimation of State Waiting Time Distributions in a Markov Multistate Model**
Somnath Datta* and Ling Lan, University of Louisville, Rajeshwari Sundaram, National Institute of Child Health and Development

4:45 **Joint Modeling of HIV Mutations and Failure Time Data**
Chengcheng Hu*, Harvard University

5:15 **Floor Discussion**

90. INFORMING HEALTH POLICY DECISIONS: BEST TEST STRATEGIES FOR COLORECTAL CANCER SCREENING *(Washington B)*

Sponsors: ENAR, ASA Health Policy Statistics Section and ASA Section on Statistics in Epidemiology
Organizer: Ann G. Zauber, Memorial Sloan-Kettering Cancer Center
Chair: Sujata Patil, Memorial Sloan-Kettering Cancer Center

3:45 **Microsimulation Modeling of Colorectal Cancer to Inform Screening Guidelines**
Ann G. Zauber*, Memorial Sloan-Kettering Cancer Center, Iris Lansdorp-Vogelaar and Janneke Wilschut, Erasmus University-The Netherlands, Amy Kundsen, Massachusetts General Hospital, Marjolein van Ballegooijen, Erasmus University-The Netherlands, Martin Brown, National Cancer Institute, Karen M. Kuntz, University of Minnesota

4:10 **The Methodology and Advantages of Comparative Modeling in Microsimulation Analyses**
Iris Lansdorp-Vogelaar*, Erasmus University-The Netherlands, Ann G. Zauber, Memorial Sloan-Kettering Cancer Center, Janneke Wilschut Erasmus University-The Netherlands, Amy Knudsen, Massachusetts General Hospital, Marjolei van Ballegooijen, Erasmus University- The Netherlands, and Karen M. Kuntz, University of Minnesota

4:35 **Comparative Effectiveness of Colorectal Cancer Screening Strategies**
Karen M. Kuntz*, University of Minnesota, Iris Lansdorp-Vogelaar and Janneke Wilschut, Erasmus University-The Netherlands, Amy B. Knudsen, Massachusetts General Hospital, Marjolein van Ballegooijen, Erasmus University-The Netherlands, Ann G. Zauber, Memorial Sloan-Kettering Cancer Center

5:00 **Potential Impact of Microsimulation Modeling of Best Test Scenarios for Colorectal Cancer Screening on Health Policy Recommendations**
Martin L. Brown*, National Cancer Institute

5:25 **Floor Discussion**

91. CONTRIBUTED PAPERS:QUANTITATIVE TRAIT LOCI MAPPING *(Potomac 1)*

Sponsor: ENAR
Chair: Heping Zhang, Yale University

3:45 **Asymptotic Test of Mixture Model and its Applications to QTL Interval Mapping**
Dong-Yun Kim*, Virginia Tech University, Yuehua Cui, Michigan State University

4:00 **Bayesian Semiparametric Multiple Quantitative Trait Loci (QTL) Mapping**
Fei Zou*, University of North Carolina at Chapel Hill, Fuxia Cheng, Illinois State University, Haibo Zhou, University of North Carolina at Chapel Hill, Ina Hoeschele, Virginia Tech University, Hanwen Huang, University of North Carolina at Chapel Hill

4:15 **Genetic Mapping of Quantitative Trait Loci in Autotetraploids**
Jiahan Li* and Rongling Wu, University of Florida

4:30 **Modeling a Non-separable Covariance Structure in High-dimensional Functional Mapping of Quantitative Traits**
John Stephen Yap* and Rongling Wu, University of Florida

4:45 **Mapping Quantitative Trait Nucleotides Encoding Complex Diseases in a Natural Population with Family Structure**
Qin Li* and Arthur Berg, Rongling Wu, University of Florida

5:00 **Bayesian QTL Mapping for Multiple Traits**
Samprit Banerjee * and Nengjun Yi, University of Alabama at Birmingham

5:15 **A Statistical Model for Characterizing cis- and trans-acting Regulation by eQTL**
Yao Li*, Jianhan Li and Ronglin Wu, University of Florida

92. CONTRIBUTED PAPERS: APPLIED DATA ANALYSIS

(Potomac 2)
Sponsors: ENAR, ASA Section on Teaching Statistics in the Health Sciences, ASA Statistical Education and ASA Section on Statistics in Epidemiology
Chair: Marc Aerts, University of Hasselt, Belgium

3:45 **Changing Approaches of Prosecutors Towards Juvenile Repeated Sex-offenders: A Bayesian Evaluation**
Dipankar Bandyopadhyay*, Medical University of South Carolina, Debajyoti Sinha, Florida State University, Stuart Lipsitz, Brigham and Women's Hospital, Elizabeth Letourneau, Medical University of South Carolina

4:00 **Variance Estimation in Regression Models**
Eugenio Andraca-Carrera* and Bahjat F. Qaqish, University of North Carolina at Chapel Hill

4:15 **Limiting the Impact of Influential Observations in Linear Regression Via Weight Functions**
Tamekia L. Jones* and David T. Redden, University of Alabama at Birmingham

4:30 **Genotype Adjusted Familial Correlation Analysis Using Three Generalized Estimating Equations**
Hye-Seung Lee*, University of South Florida, Myunghee Cho Paik and Joseph H. Lee, Columbia University

4:45 **Directional Dependence of Truncation Invariant FGM Copula Functions**
Yoon-Sung Jung*, Kansas State University, Jong-Min Kim and Engin A. Sungur, University of Minesota at Morris

5:00 **Practical Application of Statistics in Clinical Studies**
Jay Mandrekar*, Mayo Clinic College of Medicine

5:15 **Experiences in a Masters-Level Course in Biostatistical Consulting**
Stephen W. Looney* and Jennifer L. Waller, Medical College of Georgia

93. CONTRIBUTED PAPERS: LONGITUDINAL MODELS-DISCRETE *(Potomac 3)*

Sponsors: ENAR and ASA Biometrics Section
Chair: Kathleen Wannemuehler, Centers for Disease Control and Prevention

3:45 **Analysis of the Second Longitudinal Study of Aging**
Hyokyoung Hong*, University of Illinois at Urbana-Champaign

4:00 **Marginalized Models for Longitudinal Count Data**
Keunbaik Lee*, Louisiana State University-Health Science Center, Michael Daniels and Yongsung Joo, University of Florida

4:15 **Likelihood Analysis of Joint Marginal and Conditional Models for Longitudinal Categorical Data**
Baojiang Chen*, Grace Y. Yi and Richard J. Cook, University of Waterloo

4:30 **Integer-valued Autoregressive Models for Analysis of Longitudinal Count Data**
Mohamed Alosh*, U.S. Food and Drug Administration

4:45 **A Non-homogeneous Markov Process Model for Alzheimer's Disease Progression**
Rebecca A. Hubbard* and Xiao-Hua Zhou, University of Washington

5:00 **Semiparametric Models for Bivariate Panel Count Data**
Li-Yin Lee* and KyungMann Kim, University of Wisconsin

5:15 **Testing the Effect of an Exposure on Longitudinal Binary Outcomes when Events Are Rare**
Xiaonan Xue*, Mimi Y. Kim, Tao Wang and Howard D. Strickler, Albert Einstein College of Medicine

94. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS - COMPETING RISKS *(Potomac 4)*

Sponsors: ENAR, ASA Section on Risk Analysis and ASA Biometrics Section
Chair: Adin-Cristian Andrei, University of Wisconsin-Madison

3:45 **Regression Analysis for Bivariate Failure Time Associations in the Presence of a Competing Risk**
Jing Ning*, M.D. Anderson Cancer Center, Karen Bandeen-Roche, Johns Hopkins University

4:00 **Classification Trees for Survival Data with Competing Risks**
Fiona M. Callaghan*, Chung-Chou H. Chang, University of Pittsburgh

4:15 **Multi-cancer Emergence Rate Estimation with Scheduled Diagnoses: A Transition Probability Approach**
Junfeng Liu* and Weichung Joe Shih, University of Medicine and Dentistry of New Jersey and Cancer Institute of New Jersey

4:30 **Multiple Imputation Inspired by Pseudo-Value Approach for Censored Survival Data**
Lyrica X. Liu*, Susan Murray and Alex Tsodikov, University of Michigan

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Tuesday, March 18 *continued*

4:45 **Testing and Estimation of Time-Varying Cause-Specific Hazard Ratios with Covariate Adjustment**
Yanqing Sun, University of North Carolina at Charlotte, Seunggeun Hyun*, University of South Carolina Upstate, Peter Gilbert, Fred Hutchinson Cancer Research Center and University of Washington

5:00 **Bayesian Semi-parametric Regression Under Competing Risks Setting**
Xiaolin Fan*, and Purushottam W. Laud, Medical College of Wisconsin

5:15 **Nonparametric Estimation of Cause-specific Cross Hazard Ratio with Bivariate Competing Risks Data**
Yu Cheng*, University of Pittsburgh, Jason P. Fine, University of Wisconsin-Madison

95. CONTRIBUTED PAPERS:CLINICAL TRIALS - GENERAL *(Potomac 5)*

Sponsors: ASA Health Policy Statistics Section, ASA Section on Statistics in Epidemiology and ASA Biopharmaceutical Section
Chair: Valerie Durkalski, Medical University of South Carolina

3:45 **Predicting the Duration of a Sequential Trial from Interim Blinded Data**
W. J. Hall*, University of Rochester

4:00 **On Non-inferiority Assessment from Binary Data in the Combination of 2 by 2 Tables**
Kallappa M. Koti*, U.S. Food and Drug Administration

4:15 **A Survey of the Likelihood Approach to Bioequivalence**
Leena Choi*, Vanderbilt University, Brian Caffo and Charles Rohde, Johns Hopkins University

4:30 **Bayesian Screening for Pharmacogenetic Effects in Clinical Trials**
Mengye Guo* and Daniel F. Heitjan, University of Pennsylvania

4:45 **Statistical Modeling and Graphical Analysis of Safety Data in Clinical Trials**
Michael O'Connell*, Insightful, Dawn Woodard, Duke University

5:00 **Assessment of Key Factors for Overall Health Related Quality of Life**
Peng Zhao*, University of Southern California

5:15 **Benefit in Progression-free Survival (PFS) with a Genasense-dacarbazine (Genasense-DTIC) Regimen in Advanced Melanoma: A Case Study on Assessment symmetry**
Richard Kay*, R.K. Statistics-United Kingdom, Erard Gilles and Jane Wu, Genta Incorporated, Alexander M.M. Eggermont, Erasmus University Medical Center-The Netherlands

96. CONTRIBUTED PAPERS:METHODS IN SPATIAL MODELING *(Potomac 6)*

Sponsors: ENAR, ASA Section on Statistics and the Environment and ASA Section on Statistics in Defense and National Security
Chair: Ronald Gangnon, University of Wisconsin-Madison

3:45 **Hierarchical Models for Large Spatially-referenced Binary Data with an Application to Survival of Tropical Tree Seedlings**
Sang Mee Lee*, Sudipto, Banerjee and Liza Comita, University of Minnesota

4:00 **Ramps: An R Package for Unified Geostatistical Modeling of Complex Spatiotemporal Data**
Brian J. Smith*, University of Iowa, Jun Yan, University of Connecticut, Mary K. Cowles, University of Iowa

4:15 **Smoothed ANOVA with Spatial Effects as a Competitor to MCAR in Multivariate Spatial Smoothing**
Yufen Zhang*, James S. Hodges and Sudipto Banerjee, University of Minnesota

4:30 **Mining Edge Effects in Areally Referenced Spatial Data: A Bayesian Model Choice Approach**
Pei Li*, Sudipto Banerjee and Alexander M. McBean, University of Minnesota

4:45 **Nonstationary Spatial Gaussian Markov Random Field Priors**
Yu Yue* and Paul L. Speckman, University of Missouri-Columbia

5:00 **Logistic Joinpoint Regression Models in Cohort Studies**
Ryan Gill* and Grzegorz Rempala, University of Louisville

5:15 **Consistent Nonparametric Intensity Estimation for Inhomogeneous Spatial Point Processes**
Yongtao Guan*, Yale University

Wednesday, March 19

8:30—10:15 a.m.

97. COLLABORATION IN HIV RESEARCH: TWO CASE STUDIES *(Regency A)*

Sponsors: ENAR, ASA Section on Statistics in Epidemiology and ASA Biopharmaceutical Section
Organizers: Misrak Gezmu, National Institute of Allergy and Infectious Diseases & Marie Davidian, North Carolina State University
Chair: Misrak Gezmu, National Institute of Allergy and Infectious Diseases

8:30 **Using Mathematical-Statistical Modeling to Inform the Design of HIV Treatment Strategies and Clinical Trials**
Eric S. Rosenberg*, Clinician/Scientist, Massachusetts General Hospital and Harvard University, Marie Davidian*, Statistician, North Carolina State University

9:20 **Modeling and Estimation of Kinetic Parameters and Replicative Fitness of HIV-1 from Flow-Cytometry-Based Growth Competition Experiments**
Carrie Dykes*, Clinician/Scientist and Hulin Wu*, Statistician, University of Rochester

10:00 **Discussant:**
Susan Plaeger, National Institute of Allergy and Infectious Diseases

98. JOINT MODELING APPROACHES FOR LONGITUDINAL DATA UNDER COMPLEX STUDY DESIGNS *(Regency B)*

Sponsors: ASA Health Policy Statistics Section, ASA Biometrics Section and ASA Survey Research Methods Section
Organizer: Yulei He, Harvard University
Chair: Chiu-Hsieh Hsu, University of Arizona

8:30 **Case-control Studies with Longitudinal Covariates**
Honghong Zhou*, Schering-Plough Corporation, Bin Nan, University of Michigan, Xihong Lin, Harvard University

9:00 **Modeling Multivariate Latent Trajectories as Predictors of a Univariate Outcome**
Sujata Patil*, Memorial Sloan Kettering Cancer Center, Trivellore E. Raghunathan and Jean T. Shope, University of Michigan

9:30 **Statistical Methods for Adjusting Selection Bias in Longitudinal Survey Studies in Aging Research**
Wen Ye*, University of Michigan

10:00 **Floor Discussion**

99. MULTI-TASK LEARNING FOR BORROWING INFORMATION FROM DISPARATE DATA SOURCES *(Regency C)*

Sponsors: ENAR and ASA Section on Statistics in Defense and National Security
Organizer: Amy Herring, University of North Carolina at Chapel Hill
Chair: Amy Herring, University of North Carolina at Chapel Hill

8:30 **Semiparametric Bayes Borrowing of Strength in Related Analyses**
David B. Dunson*, National Institute of Environmental Health Sciences

9:00 **What Next in Gene Set Analysis?**
Giovanni Parmigiani*, Luigi Marchionni and Sierra Li, Johns Hopkins University, Dongmei Liu, London School of Hygiene and Tropical Medicine, Leslie Cope, Johns Hopkins University

9:30 **Estimating Variable Structure and Dependence in Multi-task Learning via Gradients**
Sayan Mukherjee*, Duke University

10:00 **Floor Discussion**

100. DYNAMIC TREATMENT REGIMES: PRACTICE AND THEORY *(Regency D)*

Sponsors: IMS, ASA Biometrics Section, ASA Biopharmaceutical Section
Organizer: Peter F. Thall, M.D. Anderson Cancer Center
Chair: Ken Cheung, Columbia University

8:30 **A Two-Stage Selection Trial with Intrapatient Sequential Randomization in Prostate Cancer**
Randall E. Millikan*, Peter F. Thall and Sijin Wen, M. D. Anderson Cancer Center

8:55 **Two-Stage Treatment Strategies Based on Sequential Failure Times**
Peter F. Thall*, Leiko H. Wooten, Nizar Tannir, Randall E. Millikan and Christopher Logothetis, M.D. Anderson Cancer Center

9:20 **Screening Experiments for Dynamic Treatment Regimes**
Susan Murphy*, University of Michigan, Derek Bingham, Simon Fraser University

9:45 **Optimal Treatment and Testing Strategies with Possibly Non-ignorable Observation Processes**
Andrea Rotnitzky*, Di Tella University-Argentina and Harvard University, James Robins, Harvard University, Liliana Orellana, Facultad de Ciencias Exactas, Universidad de Buenos Aires, Miguel Hernan, Harvard University

10:10 **Floor Discussion**

101. TEXT DATA MINING *(Washington A)*

Sponsors: ASA Section on Statistics in Defense and National Security and ENAR
Organizer: David Marchette, Naval Surface Warfare Center
Chair: David Marchette, Naval Surface Warfare Center

8:30 **GO-Driven Literature-based Discovery Using Semantic Analysis**
Anthony E. Zukas*, Science Applications International Corporation, Jeffrey L. Solka, George Mason University, Jennifer W. Weller, University of North Carolina-Charlotte

9:00 **Text Analysis with Iterative Denoising**
Kendall Giles*, Virginia Commonwealth University, David Marchette, Naval Surface Warfare Center, Carey Priebe, Johns Hopkins University

9:30 **Tracking Trends in Health Articles**
Elizabeth L. Hohman*, Naval Surface Warfare Center

10:00 **Discussant:** Wendy Martinez, Office of Naval Research

102. NONPARAMETRIC BAYES: THE PRACTICAL USE FOR GENOMIC DATA *(Washington B)*

Sponsors: ASA Biometrics Section and ENAR
Organizer: Sinae Kim, University of Michigan
Chair: Sinae Kim, University of Michigan

8:30 **Using Gene Annotation as Prior Information in Bayesian Nonparametric Models**
David B. Dahl*, Texas A&M University

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

8:55 **Nonparametric Graphical Models With Applications to Microarray Experiments**
Abel Rodriguez*, University of California, Santa Cruz

9:20 **A Gaussian Process Modeling Approach to Survival Data with Bayesian Variable Selection**
Naijun Sha*, The University of Texas at El Paso, Marina Vannucci, Rice University

9:45 **Integrative Dirichlet Process Mixtures for Constructing Transcriptional Modules**
Mario Medvedovic*, Xiangdong Liu and Siva Sivaganesan, University of Cincinnati

10:10 **Floor Discussion**

103. CONTRIBUTED PAPERS:METHODS IN CAUSAL INFERENCE *(Potomac 1)*

Sponsors: ENAR and ASA Section on Statistics in Epidemiology
Chair: Darryl Downing, GlaxoSmithKline

8:30 **Evaluating Longitudinal Treatments Using Regression Models on Propensity Scores**
Aristide CE Achy-Brou*, Constantine Frangakis, and Michael Griswold, Johns Hopkins University

8:45 **Structural Principal Effects Models for Randomized Trials with Continuous Measures of Compliance**
Yan Ma*, University of Rochester, Jason Roy, Geisinger Center for Health Research

9:00 **Causal Inference in Mediator Analysis: Old Problems and New Solutions**
Andreas G. Klein*, University of Western Ontario, Canada

9:15 **Compare Median Treatment Difference in Causal Inference**
Jing Qin*, National Institute of Allergy and Infectious Diseases

9:30 **Confounding in Observational Studies Comparing Propensity Score and Traditional Regression Analyses**
Megan E. Price*, Vicki Hertzberg and Michael Frankel, Emory University

9:45 **Optimal Propensity Score Stratification**
Jessica A. Myers* and Thomas A. Louis, Johns Hopkins University

10:00 **A Mixture Model Approach to Assessing the Performance of Instrument Variable Analysis in Measuring the Effect of Hormone Therapy on Prostate Cancer Survival**
Dirk F. Moore*, University of Medicine and Dentistry New Jersey, Grace Lu-Yao, Cancer Institute of New Jersey

104. CONTRIBUTED PAPERS:BIOASSAY AND CANCER APPLICATIONS *(Potomac 2)*

Sponsors: ENAR and ASA Biometrics Section
Chair: John Peterson, GlaxoSmithKline

8:30 **Stochastic and State Space Models of Human Eye Cancer**
W. Y. Tan, University of Memphis, H. Zhou*, Arkansas State University

8:45 **Improving Biological Plausibility of Classification Trees for Cancer Research**
Douglas Landsittel*, Duquesne University, Megan McLaughlin, Independent Consultant, Anna Lokshin, University of Pittsburgh

9:00 **On the Robustness of the Ploy-k Test for Life Time Animal Studies**
Mulugeta G. Gebregziabher* and David Hoel, Medical University of South Carolina

9:15 **Interval Approach to Assessing Antitumor Activity for Tumor Xenograft Studies**
Jianrong Wu*, St Jude Children's Research Hospital

9:30 **Comparison of Properties of Tests for Assessing Tumor Clonality**
Irina Ostrovnaya* Memorial Sloan-Kettering Cancer Center, Venkatraman Seshan, Columbia University, Colin Begg, Memorial Sloan-Kettering Cancer Center

9:45 **Detecting Group Differences With Right-Censored Counts From Serial Dilution Assays**
Tim Bancroft* and Dan Nettleton, Iowa State University

10:00 **An Optimal Dilution Experiment Design for Sample DNA Concentration Estimation**
Ming Li*, Robbert Slebos and Yu Shyr, Vanderbilt University Medical Center

105. CONTRIBUTED PAPERS:MISSING DATA AND IMPUTATION *(Potomac 3)*

Sponsors: ASA Survey Research Methods Section, ASA Section on Statistics in Epidemiology and ASA Biometrics Section
Chair: Kimberly Drews, George Washington University

8:30 **Estimating Spatial Intensity from Locations Coarsened by Incomplete Geocoding**
Dale L. Zimmerman*, University of Iowa

8:45 **Imputation Model Assessment Using Posterior Predictive Checking**
Yulei He* and Alan M. Zaslavsky, Harvard Medical School

9:00 **Sequential Semi and Nonparametric Regression Multiple Imputations**
Irina Bondarenko* and Trivellore Raghunathan, University of Michigan

9:15 **Imputation Adjusted for Covariate for Nonrespondents with Applications**
Juan Li* Amgen Inc., Shein-Chung Chow, Duke University, Amy Feng, Amgen Inc., Jr-Rung Lin, Duke University, Eric Chi, Amgen Inc.

9:30 **Fractional Imputation for Categorical Variables with Missing Values**
Michael D. Larsen*, Iowa State University

9:45 **Multiple Imputation Methods for Treatment Noncompliance and Missing Data in Clustered Encouragement Design Studies**
Leslie L. Taylor*, University of Washington and Axio Research, Xiao-Hua Zhou, University of Washington

10:00 **Multiple Imputation of Ordinal and Count Outcomes in a Multiple Sclerosis Clinical Trial Using Data at Dropout**
Peter B. Imrey* and John Barnard, Cleveland Clinic at Case Western Reserve University, Matthew Karafa, Cleveland Clinic Foundation

106. CONTRIBUTED PAPERS:LONGITUDINAL MODELS-CONTINUOUS *(Potomac 4)*

Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Renee Moore, University of Pennsylvnia

8:30 **The Mixture of Nonlinear Models for Gastric Emptying Studies**
Inyoung Kim*, Virginia Tech University, Noah D. Cohen, Allen Roussell, and Naisyin Wang, Texas A & M University

8:45 **Comparisons of Methods for Handling Missing Continuous Longitudinal Outcome with Mixed Effects Models**
Tulay Koru-Sengul*, McMaster University-Canada

9:00 **Modeling Covariance Structure in Unbalanced Longitudinal Data**
Min Chen*, Texas A&M University

9:15 **The GLM with a Generalized AR(1) Covariance Structure**
Sean L. Simpson* and Lloyd J. Edwards, University of North Carolina at Chapel Hill, Keith E. Muller, University of Florida, Pranab K. Sen, University of North Carolina at Chapel Hill

9:30 **Nonparametric Estimation for Conditional Distribution Functions and Time-Varying Transformation Models with Longitudinal Data**
Colin O. Wu, Xin Tian* and Jarvis Yu, National Heart, Lung and Blood Institute

9:45 **Inference for Censored Quantile Regression Models in Longitudinal Studies**
Huixia Judy Wang*, North Carolina State University, Mendel Fygenson, University of Southern California

10:00 **Characterization of Variability in Longitudinal Data Using the Spectrum**
Wei Yang*, Marshall Joffe and Steven Brunelli, University of Pennsylvania School of Medicine

107. CONTRIBUTED PAPERS:BAYESIAN AND MULTI-LEVEL SURVIVAL ANALYSIS *(Potomac 5)*

Sponsors: ENAR and ASA Section on Statistics in Epidemiology
Chair: Dennis O. Dixon, National Institute of Allergy and Infectious Diseases

8:30 **Hierarchical Dynamic Time-to-Event Models for Post-treatment Preventive Care Data on Breast Cancer Survivors**
Freda W. Cooner*, U.S. Food and Drug Administration, Xinhua Yu, Sudipto Banerjee, Patricia L. Grambsch and A. Marshall McBean, University of Minnesota

8:45 **Bayesian Semiparametric Modeling using Mixtures for Stratified Survival Data**
Bo Cai*, University of South Carolina, Renate Meyer, University of Auckland

9:00 **A Transformation Approach for the Analysis of Interval-censored Failure Time Data**
Liang Zhu*, University of Missouri, Xingwei Tong, Beijing Normal University, Jianguo Sun, University of Missouri

9:15 **Bayesian Threshold Regression with Random Effects**
Michael L. Pennell* and Mei-Ling Ting Lee, The Ohio State University

9:30 **The Application of the Bayesian Dynamic Survival Model**
Jianghua He*, University of Kansas Medical Center, Daniel L. McGee and Xufeng Niu, Florida State University

9:45 **Linear Regression Analysis of Survival Data from Stratified Case-cohort Studies**
Lan Kong*, University of Pittsburgh, Jianwen Cai, University of North Carolina at Chapel Hill

10:00 **On Risk Reversal Due to Heterogeneity**
David Oakes*, University of Rochester

108. CONTRIBUTED PAPERS:MULTIPLE TESTING IN GENOMICS *(Potomac 6)*

Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Xiaonan Xue, Albert Enstein College of Medicine

8:30 **A Sequential Testing Procedure for Detection of Association Between Disease and Haplotype Blocks: A Simulation Study**
Andres Azuero* and David T. Redden, University of Alabama at Birmingham

8:45 **A Parametric Permutation Test for Regression Coefficients in LASSO Regularized Regression for High Dimensional Data**
Michael C. Wu*, Tianxi Cai and Xihong Lin, Harvard School of Public Health

9:00 **Empirical Bayes Modeling with Optimal Discovery Procedure for Improved Significance Testing of Microarray Data**
Xiting Cao* and Baolin Wu, University of Minnesota

9:15 **A Flexible Bayesian Framework for Large Scale Simultaneous Testing**
Seungbong Han*, Adin Cristian Andrei and Kam Wah Tsui, University of Wisconsin - Madison

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

Wednesday, March 19 *continued*

9:30 **Two-Stage Group Sequential Robust Tests in Family-Based Association Studies: Controlling Type I Error**
Lihan K. Yan*, George Washington University; The EMMES Corporation, Gang Zheng, National Heart, Lung and Blood Institute, Zhaohai Li, George Washington University, National Institute of Child Health and Human Development

9:45 **Improved Estimation of False Discovery Rates Based on Subsampling with Applications to Microarray Data**
Long Qu*, Dan Nettleton and Jack CM Dekkers, Iowa State University

10:00 **Data Complexity Dependence of False Discovery Rate Estimates in Differential Gene Selection Procedures**
Michael G.. Schimek*, Medical University of Graz-Austria, Tomas Pavlik, Masaryk University Czech Republic

Wednesday, March 19

10:15—10:30 a.m. **Refreshment Break and Visit the Exhibitors**

10:30 a.m.—12:15 p.m.

109. ADVANCES IN THE DESIGN AND ANALYSIS OF RANDOMIZED CLINICAL TRIALS *(Regency A)*

Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Section on Statistics in Epidemiology
Organizer: Vance Berger, National Cancer Institute
Chair: Kirsten Doehler, University of North Carolina-Greensboro

10:30 **Re-formulating Non-inferiority Trials as Superiority Trials: The Case of Binary Outcomes**
Valerie Durkalski*, Medical University of South Carolina, Vance Berger, National Cancer Institute

10:55 **Placebo Effects or Medication Compliance?**
William Grant*, James Madison University

11:20 **Enrollment Projections: Estimating with Confidence**
Venita DePuy*, INC Research, Inc.

11:45 **Selection Bias and Covariate Imbalance in Randomized Clinical Trials**
Vance Berger*, National Cancer Institute

12:10 **Floor Discussion**

110. RECENT ADVANCES IN GRAPHS/GRAPHICAL MODELS FOR GENETIC NETWORK ANALYSIS *(Regency B)*

Sponsors: ENAR and ASA Biometrics Section
Organizer: Jie Peng, University of California-Davis
Chair: Fan Li, Harvard Medical School

10:30 **Network Structure Inference in Biology**
Wing H. Wong*, Stanford University

11:00 **The Modal Oriented Stochastic Search for Graphical Models**
Adrian Dobra*, University of Washington

11:30 **High Dimensional Graphs Inference by Sparse Regression**
Jie Peng*, University of California-Davis, Pei Wang, Fred Hutchinson Cancer Research Center, Ji Zhu, University of Michigan

12:00 **Floor Discussion**

111. NEW STATISTICAL METHODS FOR BIOMEDICAL IMAGING DATA *(Regency C)*

Sponsors: ENAR, ASA Biopharmaceutical Section and ASA Biometrics Section
Organizer: Ying Guo, Emory University
Chair: F. DuBois Bowman, Emory University

10:30 **Statistical Analysis of Neuroimaging Data using Adjusted Exponetially Tilted Likelihood**
Hongtu Zhu* and Haibo Zhou, University of North Carolina at Chapel Hill, Jiahua Chen, University of British Columbia- Canada, Yimei Li and Martin Styner, University of North Carolina at Chapel Hill

10:55 **Analyzing High Temporal Resolution fMRI data**
Martin A. Lindquist*, Columbia University

11:20 **Pharmacologic Imaging Using Principal Curves in Single Photon Emission Computed Tomography**
Brian S. Caffo*, Johns Hopkins University, Lijuan Deng, Boston Scientific Company, Ciprian Crainiceanu and Craig Hendrix, Johns Hopkins University

11:45 **A Unified Framework for Group Independent Component Analysis for Multi-subject fMRI Data**
Ying Guo*, Rollins School of Public Health, Emory-University

12:10 **Floor Discussion**

112. STATISTICAL CHALLENGES IN GENOMEWIDE ASSOCIATION STUDIES *(Regency D)*

Sponsors: IMS, ASA Biometrics Section and ASA Biopharmaceutical Section
Organizer: Danyu Lin, University of Michigan
Chair: Danyu Lin, University of Michigan

10:30 **Searching for Disease Susceptibility Variants in Structured Populations**
Kathryn Roeder*, Carnegie Mellon University

11:00 **Are We Ready to Look at Copy Number Variation in Whole-genome Association Studies?**
Eleanor Feingold*, University of Pittsburgh

11:30 **Using Genotype Imputation to Combine Data Across Studies**
Goncalo Abecasis*, Yun Li and Paul Scheet, University of Michigan

12:00 **Floor Discussion**

113. RECENT ADVANCES IN ANALYZING BIOMARKER DATA WITH LIMITS OF DETECTION *(Washington A)*

Sponsors: ENAR, ASA Biometrics Section and ASA Biopharmaceutical Section
Organizer: Paul Albert, National Cancer Institute
Chair: Zonghui Hu, National Institutes of Health

10:30 Relating a Reproductive Health Outcome to Subject-specific Features Based on Left- or Interval-censored Longitudinal Exposure Data
Kathleen A. Wannemuehler*, Centers for Disease Control and Prevention, Robert H. Lyles, Amita K. Manatunga, Renee H. Moore, Metricia L. Terrell and Michele Marcus, Emory University

11:00 A Bayesian Approach Estimating Treatment Effects on Biomarkers Containing Zeros with Detection Limits
Haitao Chu*, University of North Carolina at Chapel Hill, Lei Nie, Georgetown University, Thomas W. Kensler, Johns Hopkins University

11:30 A Combined Efficient Design Based on Data Subject to Detection Limit
Enrique F. Schisterman*, National Institute of Child Health and Development, Albert Vexler, The State University of New York at Buffalo

12:00 Discussant:
Jim Hughes, University of Washington

114. BIOLOGICAL APPLICATIONS OF MACHINE LEARNING *(Washington B)*

Sponsor: ENAR
Organizer: Sayan Mukherjee, Duke University
Chair: Sayan Mukherjee, Duke University

10:30 Modeling and Detection of Spatial Patterns of Brain Activation from fMRI Data
Polina Goland*, Massachusetts Institute of Technology

11:00 Regularization Approach to Screening Biomarkers
Yoonkyung Lee*, The Ohio State University

11:30 Learning Predictive Models of Gene Regulation
Christina Leslie*, Memorial Sloan-Kettering Cancer Center

12:00 Floor Discussion

115. CONTRIBUTED PAPERS: HEALTH SERVICES AND POLICY RESEARCH *(Potomac 1)*

Sponsors: ENAR, ASA Health Policy Statistics Section and ASA Section on Statistics in Epidemiology
Chair: A. James O'Malley, Harvard Medical School

10:30 Informative Screening
Joshua M. Tebbs*, University of South Carolina, Christopher R. Bilder, University of Nebraska

10:45 Small-area Estimation of Mental Illness Prevalence for Schools
Fan Li*, Alan M. Zaslavsky and Ronald Kessler, Harvard University

11:00 Bayesian Ordering of a Trinomial Sample Space for Classical Hypothesis Testing
A John Bailer, Robert B. Noble and Douglas A. Noe*, Miami University

11:15 Backward Estimation of Medical Cost in the Presence of a Failure Event
Kwun Chuen Gary Chan* and Mei-Cheng Wang, Bloomberg School of Public Health-Johns Hopkins University

11:30 Some Issues in Suicide Attempt Prediction
Steven P. Ellis*, Columbia University

11:45 Gamma Shape Mixtures for Heavy-tailed Distributions
Sergio Venturini*, Bocconi School of Management-Italy, Francesca Dominici and Giovanni Parmigiani, Johns Hopkins University

12:00 A Bayesian Location Shifted Mixture Model for HMO Pharmacovigilance Research Database
Fang Zhang* and Martin Kulldorff, Harvard University

116. CONTRIBUTED PAPERS:MEASUREMENT ERROR *(Potomac 2)*

Sponsors: ENAR, ASA Survey Research Methods Section and ASA Section on Statistics in Epidemiology
Chair: Yu Zhao, Harvard University

10:30 Misclassification Adjustment in Threshold Models for the Effects of Subject-Specific Exposure Means and Variances
Chengxing Lu* and Robert H. Lyles, Emory University

10:45 Average Causal Effect Estimation Allowing Covariate Measurement Error: A Finite Mixture Modeling Framework
Yi Huang*, University of Maryland, Karen Bandeen-Roche and Constantine Frangakis, Johns Hopkins University

11:00 Measurement Error Model with Unknown Error
Peter Hall, University of Neuchatel, Yanyuan Ma*, Australian National University

11:15 Modeling Heaping in Self-reported Cigarette Counts
Hao Wang* and Daniel F. Heitjan, University of Pennsylvania

11:30 Linear Model Covariate Measurement Error Correction via Multiple Imputation
Miguel A. Padilla*, University of Alabama at Birmingham, Jasmin Divers, Wake Forest University, Hemant K. Tiwari, University of Alabama at Birmingham

11:45 Recursive Estimation Method for Predicting Residual Bladder Urine Volumes to Improve Accuracy of Timed Urine Collections
David Afshartous* and Richard A. Preston, University of Miami, Miller School of Medicine

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

12:00 **Using Orthogonal Decomposition to Adjust Regression Calibration in Physical Activity Studies**
Sanguo Zhang*, Vanderbilt University School of Medicine

117. CONTRIBUTED PAPERS: MICROARRAY DATA ANALYSIS *(Potomac 3)*
Sponsors: ENAR, ASA Biometrics Section and ASA Biopharmaceutical Section
Chair: Boris Iglewicz, Temple University

10:30 **A Geometric Approach for Detecting Cell Cycle Gene Expression**
Omar De la Cruz* and Dan L. Nicolae, University of Chicago

10:45 **Two-Criterion: A New Bayesian Differentially-Expressed Gene Selection Algorithm**
Fang Yu*, University of Nebraska Medical Center, Ming-Hui Chen and Lynn Kuo, University of Connecticut

11:00 **A Non-parametric Meta-analysis Approach for Combining Independent Microarray Datasets: Application Using Two Microarray Datasets Pertaining to Chronic Allograft Nephropathy**
Xiangrong Kong*, Valeria Mas and Kellie J. Archer, Virginia Commonwealth University

11:15 **RXA: Analysis of Gene Expression Microarrays Based on Expression Orderings**
Xue Lin*, Daniel Naiman, Leslie Cope, Giovanni Parmigiani, and Donald Geman, Johns Hopkins University

11:30 **Profiling Time Course Expression of Virus Genes**
I-Shou Chang*, Li-Chu Chien and Chao Hsiung, National Health Research Institutes-Taiwan

11:45 **Identify Equivalently Expressed Genes from Microarray Data Using a Mix-scaled Equivalence Criterion**
Jing Qiu, University of Missouri-Columbia, Xiangqin Cui*, University of Alabama at Birmingham

12:00 **A Statistical Framework for Integrating Different Microarray Data Sets in Differential Expression Analysis**
Yinglei Lai*, George Washington University

118. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS METHODS AND APPLICATIONS *(Potomac 4)*
Sponsors: ENAR and ASA Biopharmaceutical Section
Chair: Xiaoxi Zhang, Pfizer

10:30 **Accelerated Quantile Residual Life Model**
Hanna Bandos* and Jong-Hyeon Jeong, University of Pittsburgh

10:45 **Bayesian Case Influence Diagnostics for Survival Models**
Hyunsoon Cho* and Joseph G. Ibrahim, University of North Carolina at Chapel Hill, Debajyoti Sinha, Medical University of South Carolina, Hongtu Zhu, University of North Carolina at Chapel Hill

11:00 **A Robust Weighted Kaplan-Meier Approach Using Linear Combinations of Prognostic Covariates**
Chiu-Hsieh Hsu*, University of Arizona, Jeremy M.G. Taylor, University of Michigan

11:15 **Modeling Discrete Survival Data with Application to Reproductive Study**
Huichao Chen*, Harvard University, Amita K. Manatunga, Limin Peng and Michele Marcus, Emory University

11:30 **Goodness-of-fit Testing for the Cox Proportional Hazards Model with Time-dependent Covariates**
Susanne May*, Jennifer Emond, Steve Edland and Loki Natarajan, University of California-San Diego

11:45 **Censored Median Regression and Profile Empirical Likelihood**
Sundar Subramanian*, New Jersey Institute of Technology

12:00 **It's All About the Proportional Hazards Assumption**
Kyung Y. Lee*, Yuan-Li Shen and Kallappa M. Koti, U.S. Food and Drug Administration

119. CONTRIBUTED PAPERS: VARIABLE SELECTION AND MODEL BUILDING - APPLICATIONS *(Potomac 5)*
Sponsor: ENAR
Chair: Kenneth Portier, American Cancer Society

10:30 **Recursive Partitioning Analysis and Proportional Hazard Models in Prognostic Factors Analysis: A Joint Cooperative Groups Study for High Grade Recurrent Glioma**
Wenting Wu*, Mayo Clinic, Kathleen Lamborn, University of California-San Francisco, Paul Novotny and Terry Therneau, Mayo Clinic

10:45 **Network-constrained Regularization and Variable Selection for Analysis of Genomic Data**
Caiyan Li* and Hongzhe Li, University of Pennsylvania School of Medicine

11:00 **Confounder-selection using the Change in Estimate approach; Is 10% All We Need to Know?**
Gheorghe Doros*, Boston University, Robert Lew, V.A. Boston, Alexander Ozonoff, Boston University

11:15 **Incorporating Prior Knowledge of Gene Functional Groups into Regularized Discriminant Analysis of Microarray Data**
Feng Tai* and Wei Pan, University of Minnesota

11:30 **Hierarchically Penalized Cox model for Survival Data with Grouped Variables and Its Oracle Property**
Sijian Wang*, Nengfeng Zhou, Bin Nan and Ji Zhu, University of Michigan

11:45 **Classification of Families of Locally Stationary Time Series**
Robert T. Krafty*, University of Pittsburgh, Wensheng Guo, University of Pennsylvania

120. CONTRIBUTED PAPERS:NONPARAMETRIC METHODS *(Potomac 6)*

Sponsor: ENAR
Chair: Hany Zayed, Nektar Therapeutics

10:30 **Nonparametric Additive Regression for Repeatedly Measured Data**
Raymond Carroll, and Arnab Maity*, Texas A&M University, Enno Mammen and Kyusang Yu, University of Mannheim-Germany

10:45 **Local Post-stratification and Diagnostics in Dual System Accuracy and Coverage Evaluation for US Census**
Chengyong Tang* and Song X. Chen, Iowa State University

11:00 **Analyzing Forced Unfolding of Protein Tandems via Order Statistics**
Efstathia Bura*, George Washington University, Dmitri Klimov, George Mason University, Valeri Barsegov, University of Massachusetts-Lowell

11:15 **On Distribution-free Runs Test for Symmetry Using Median Ranked Set Sampling**
Hani M. Samawi*, Georgia Southern University-Jiann-Ping Hsu College of Public Health

11:30 **Beta Regression Model for the Area under the ROC Curve**
Lin Zhang* and Jack Tubbs, Baylor University

11:45 **Density Estimation in Phase II Studies in Large Scale Community-based Research**
Wan Tang*, Hua He and Xin Tu, University of Rochester

Times may change slightly prior to the meetings. Please check the on-site program for final times.
Asterisks (*) indicate the presenter of the paper. Circles Denote Student Award Winners.

1. POSTERS: EPIDEMIOLOGIC METHODS AND SURVEY RESEARCH

1a. EXTREME VERIFICATION BIAS IN PAIRED CONTINUOUS TESTS MAY MASK THE MAGNITUDE OF THE DIFFERENCE BETWEEN THE DIAGNOSTIC ACCURACIES OF SCREENING MODALITIES

Deborah H. Glueck*, University of Colorado
Molly M. Lamb, University of Colorado
Colin O'Donnell, University of Colorado
Keith E. Muller, University of Florida
John M. Lewin, Diversified Radiology of Colorado

Extreme verification bias occurs when disease status is verified with further testing only when the test result is suspicious. A common design for the comparison of new and existing diagnostic tests involves screening all participants with both methods, and then using a paired comparison of the areas under the ROC curves. We show that such parallel study designs with continuous tests are subject to paired extreme verification bias. In exceptional cases, the magnitude of the observed difference between the diagnostic accuracy of the screening tests may be incorrect because of this bias. The extent of paired extreme verification bias is strongly affected both by the rate at which cases undetected by the screening test present with signs and symptoms of disease, and by the test score that triggers a confirmatory test. Paired extreme verification bias is also influenced by the distributions of the ROC scores for the cases and non-cases for each test, as well as the underlying disease prevalence. We use a comparison of screen film and digital mammography as a driving example of paired extreme verification bias.

email: Deborah.Glueck@uchsc.edu

1b. CASE-CONTROL STUDY OF LUNG-CANCER RISK FROM RESIDENTIAL RADON EXPOSURE IN WORCESTER COUNTY, MASSACHUSETTS

Richard E. Thompson*, Johns Hopkins Bloomberg School of Public Health
Donald F. Nelson, Worcester Polytechnic Institute
Joel H. Popkin, St. Vincent Hospital and Fallon Clinic, Worcester Medical Center
Zenaida Popkin, St. Vincent Hospital and Fallon Clinic, Worcester Medical Center

A study of lung cancer risk from residential radon exposure and its radioactive progeny was performed with 200 cases (58% male, 42% female) and 397 controls matched in age and sex, all from the same health maintenance organization. Emphasis was placed on accurate and extensive, year-long dosimetry with etch-track detectors in conjunction with careful questioning about historic patterns of in-home mobility. Conditional logistic regression was used to model the outcome of cancer on radon exposure, while controlling for years of residency, smoking, education, income, and years of job exposure to known or potential carcinogens. Radon exposure was divided into six categories with break points at 25, 50, 75, 150, and 250 Bq m-3, the lowest being the reference. The adjusted odds ratios (AOR) were, in order, 1.00, 0.53, 0.31, 0.47, 0.22, and 2.50 with the third category significantly below 1.0 ($p<0.05$), and the second, fourth, and fifth categories approaching statistical significance ($p<0.1$). An alternate regression analysis using natural cubic splines allowed calculating AORs as a continuous function of radon exposure. That analysis produced AORs that are substantially less than 1.0 with borderline statistical significance (0.048 d p d 0.05) between approximately 85 and 123 Bq m-3.

email: rthompso@jhsph.edu

1c. EFFECT OF BMI ON LIFETIME RISK OF DIABETES FOR U.S. ADULTS

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

We estimated lifetime risk of diagnosed diabetes among U.S. adults (18-84 years) without diabetes at baseline age by race, sex, and Body Mass Index (BMI) category. Heirarchical Bayesian models were applied to National Health Interview Survey data (1997-2004) to estimate age-, race-, sex-, and BMI category-specific prevalence and incidence of diabetes in year 2004. U.S. Census Bureau age-, race-, sex- specific population and mortality rate estimates for 2004 were combined with two previous studies of mortality by BMI and diabetes status to estimate mortality rates for the diabetic and non-diabetic populations by BMI. These estimates were then entered into a Markov model to estimate risk of diabetes at baseline age, by race, sex, and BMI. At any baseline age, and for every race-sex subgroup, risk of diabetes increases with increasing BMI. For example, among young adults without diabetes at age = 18 years in 2004, the risk of diabetes is projected to be 29.7% for males and 35.4% for females with BMI 25 to <30. And, for BMI > 35, 70.3% for males and 74.4% for females. Also, minority race has a strong influence on remaining risk of diabetes.

email: jboyle@cdc.gov

1d. VARIABLE SELECTION FOR MULTIPLY-IMPUTED LARGE DATA SET IN DIOXIN EXPOSURE STUDY

Qixuan Chen*, University of Michigan School of Public Health
James M. Lepkowski, Institute for Social Research, University of Michigan
Brenda Gillespie, University of Michigan School of Public Health
David H. Garabrant, University of Michigan School of Public Health

Multiple imputation is frequently used to handle the missing data in large public use survey data sets. Variable selection methods are applied in such data sets to identify important predictors in multiple predictor models. However, variable selection within the framework of multiple imputation has been an important issue since the standard variable selection methods such as stepwise selection commonly identify different important predictors across the multiply imputed data sets. We describe a combine, then select (C-S) strategy implemented by combining multiply imputed data first and then selecting variables on the combined data using multiple imputation inference procedures in each step of selection. The C-S selection stops when the combined inference for all the variables staying in the model satisfies the selection criteria. The strategy is illustrated using data from the University of Michigan Dioxin Exposure Study which generated a large survey data set. In addition, simulation studies are presented that compare the performance of the C-S approach with other variable selection strategies applied to multiply imputed data in epidemiological studies. A SAS macro for the C-S stepwise selection is also provided.

email: qixuan@umich.edu

1e. A LATENT VARIABLE MODEL FOR AN EPIDEMIOLOGICAL STUDY WITH MULTIPLE CORRELATED EXPOSURES

Abbie Stokes-Riner*, University of Rochester School of Medicine and Dentistry
Sally W. Thurston, University of Rochester School of Medicine and Dentistry
Jason Roy, Geisinger Center for Health Research

Epidemiologic researchers are often interested in studying the effect of environmental exposures on a specified health condition. Such studies may include a number of highly correlated exposures, making multiple regression models problematic due to multicollinearity. In addition the exposure biomarkers may be measured with error. If the health condition considered is complex, it may only be possible to assess it through measurement of multiple outcomes. If we assume a latent variable structure underlying both the exposure and outcome measurements, then we can examine the relationship between a few latent variables rather than the relationships between numerous exposure and outcome variables. Such a latent variable model will alleviate the problem of multicollinearity in the exposures and may also reduce measurement error bias. Using Bayesian methods allows not only estimation of the relationship between the latent exposure and latent outcome variables, but also the subject-specific latent variables, which quantify the complex health condition. We use a latent variable model to estimate the relationship between multiple correlated environmental exposures and male fertility using data from the Study for Future Families.

email: abbie_stokesriner@urmc.rochester.edu

1f. PARAMETER ESTIMATION ON A PARTIALLY GROUPED SAMPLE

Sergey Tarima*, Medical College of Wisconsin

In some questionnaires, grouped responses are occasionally provided instead of exact answers expected by an investigator. For example, if a respondent is asked about his/her priority regarding a treatment choice, he/she can write ‘1-2’ instead of providing an exact number 1, 2, 3, or 4. Then, statisticians face partially grouped data. Discarding grouped responses may lead to significant loss of efficiency and bias. In this work partially grouped sample is transformed into a dataset with missing data for further analysis. After such transformation missing data almost are exclusively MAR not MCAR. An example of a survey question with partially grouped responses is presented for illustrative purposes.

email: starima@mcw.edu

1g. BAYESIAN HIERARCHICAL METHODS FOR SMALL AREA ESTIMATION: DIABETES PREVALENCE BY U.S. COUNTIES

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

There is an ongoing need to assess health status at both the state and county levels. Estimates of health status help determine demand for health care and inform health policy decision making. The Behavioral Risk Factor Surveillance System (BRFSS) is a very large ongoing telephone survey designed to provide state-level estimates of health status including prevalence of diagnosed diabetes. However, the BRFSS cannot provide design-based direct estimates for diabetes prevalence by some smaller geographical areas or subpopulations such as counties. Using BRFSS and US census data we develop Bayesian multilevel models to estimate diabetes prevalence by county. The models are fit using WinBUGS. The estimates will be included in the Centers for Disease Control and Prevention’s ‘Diabetes Data and Trends’ web site.

email: tat5@cdc.gov

1h. SHIFT IN AGE DISTRIBUTION OF PERTUSSIS INFANT MORTALITY TO VERY YOUNG INFANTS, UNITED STATES, 193-2004

Andrew L. Baughman*, Centers for Disease Control and Prevention
Tracy Pondo, Centers for Disease Control and Prevention
Margaret M. Cortese, Centers for Disease Control and Prevention

To characterize the change in the pertussis infant mortality rate and in the shape of the age distribution of pertussis infant deaths over time, we analyzed data on pertussis deaths in infants aged <12 months from U.S. vital statistics for 1933-2004. U.S. natality data on number of live births were used for calculating death rates. Data were analyzed in four time periods: pre-vaccine era (1933-1944) and post-vaccine era (1945-1964, 1965-1984, 1985-2004). The average annual death rate per 100,000 live births decreased from 97 during 1933-1944 to 0.19 during 1985-2004. We fit a zero-deflated right truncated negative binomial model to the age distribution of pertussis deaths in single months of age at death (0 to 11). Results from fitting the model indicated that the average age at death decreased from 4.5 months for 1933-1944 to 1.9 months for 1985-2004. The estimated proportion of infant pertussis deaths among infants aged <3 months increased from 35% to 77% over this time span. The covariates gender, race, and ethnicity were also considered in the analyses. These results suggest that the age distribution of pertussis deaths in infants has shifted over the study period with deaths now concentrated in very young infants. Additional vaccination strategies are needed to protect young infants from death due to pertussis.

email: alb1@cdc.gov

2. POSTERS: LONGITUDINAL AND CATEGORICAL DATA

2a. PROCESS MODELING FOR ORDERED CATEGORICAL DATA

Simone Gray*, Duke University
Alan Gelfand, Duke University

We propose a method which models the expected cell counts of a contingency table using a spatial process. Suppose that all m classifying variables in a multi-way contingency table are ordinal. Then expected cell counts in the table can be viewed as being driven by an intensity surface over R^m. We propose to model this intensity surface as a realization of an m-dimensional spatial process. With this new approach our expected cell counts are driven by a continuous process, allowing flexible modeling for the joint probabilities in the table, interpolation to other ordinal classifications of interest not necessarily specified by the given dataset and an alternative method for modeling multi-dimensional

ordered categorical data. We detail this methodology with Bayesian hierarchical modeling and show the associated computation and dimension reduction techniques necessary to fit these models. Finally we will illustrate this new approach with both simulated data and real data from the North Carolina detailed birth records.

email: simone@stat.duke.edu

2b. TESTS FOR ZERO INFLATION IN A BIVARIATE ZERO-INFLATED POISSON MODEL

Byoung Cheol Jung*, University of Seoul
JungBok Lee, Korea University
Seo Hoon Jin, Korea University

The score test statistics for testing zero inflation and covariance parameter are proposed in the zero-inflated bivariate Poisson regression model. The Monte Carlo studies show that the score test and LR test for testing zero inflation underestimate the nominal significance level, while the score test for covariance parameter keeps the significance level close to the nominal one. To overcome this nominal level underestimation, we propose a bootstrap method of the score test for the testing problem of zero inflation. Two empirical examples with and without covariates are provided to illustrate the results.

email: bcjung@uos.ac.kr

2c. ESTIMATION OF CAUSAL EFFECTS IN LONGITUDINAL OBSERVATIONAL STUDIES

Elizabeth Johnson*, Johns Hopkins University
Constantine Frangakis, Johns Hopkins University
Scott L. Zeger, Johns Hopkins University

We investigate via a case-study methodologies for estimating the causal effect of interventions that change in time when the propensity of receiving a treatment depends upon the past outcome values. We contrast a recent method by Archy-Brou and Frangakis (2007) with marginal structural models (Robins, 1997, 1999). The Archy-Brou and Frangakis method is a likelihood method utilizing the g-computation formula of Robins (1986) and the longitudinal history of propensity scores to remove the effects of time-varying confounders. We develop graphical representations and generalize the modeling strategies of Archy-Brou and Frangakis allowing for smooth functions of the history of propensity scores. We study the existing methodology based on weights and the new methodology in the case of labor interventions on the first stage of labor (time from onset of labor to full dilation) among women experiencing their first delivery. We assess the sensitivity of our results to our testable modeling assumptions. In our application, the methodologies yield similar results; however, the Archy-Brou and Frangakis method allows for a deeper understanding of the contribution of observable data in the estimation of the causal effects.

email: ejohnson@jhsph.edu

2d. SELECTING THE WORKING CORRELATION STRUCTURE BY A NEW GENERALIZED AIC INDEX FOR LONGITUDINAL DATA

Jiawei Liu*, Georgia State University
Bruce G. Lindsay, Pennsylvania State University
Wei-Lun Lin, Georgia State University

The analysis of longitudinal data has been a popular subject for the recent years. In this paper, we are interested in the influence of different 'working' correlation structures for modeling the longitudinal data. We propose a new AIC-like method for the model assessment which generalized AIC from the point of view of data generating. By comparing the log-likelihood functions of different correlation models, we define the interval of sample size, on which a correlation structure will model the data well, for our model selection.

email: jliu16@gsu.edu

2e. DIAGNOSTICS FOR THE COVARIANCE STRUCTURE OF NONLINEAR MIXED EFFECTS MODELS FOR BALANCED LONGITUDINAL DATA

Karen Chiswell*, GlaxoSmithKline
John F. Monahan, North Carolina State University

Longitudinal data, where the response for each experimental unit is measured repeatedly on several occasions, are common in biomedical and agricultural studies. The nonlinear mixed effects model (NLMM) provides a useful statistical framework for characterizing longitudinal data when a nonlinear regression function describes the trend for each experimental unit or subject. Our methods explore the structure of variation in longitudinal data using traditional tools of multivariate analysis (principal component analysis, factor analysis, and multivariate regression). We propose methods for identifying the non-trivial random effects in the NLMM, for assessing the presence of within-subject heteroscedasticity (non-constant variance), and for testing evidence of within-subject autocorrelation. In simulation studies our methods demonstrated high power to correctly identify random effects, to detect heteroscedasticity, and to find strong AR(1) structure in within-subject residuals. We apply our methods to a physiologically-based pharmacokinetic (PBPK) model for carbon tetrachloride inhalation exposure in rats.

email: karen.e.chiswell@gsk.com

2f. ESTIMATING SAMPLE SIZE FOR MISCLASSIFIED BINARY RESPONSE WITH COVARIATE MEASUREMENT ERROR

Dunlei Cheng*, Institute for Health Care Research and Improvement, Baylor Health Care System

This work endeavors to estimate sample size after considering dual effects of misclassification and measurement error. Since the outcome is a binary variable, logistic regression is applied for the paper. Bayesian paradigm is employed giving each parameter a prior density. The sample size crieterion used is Bayesian power, based on posterior effect size. Simulations suggest that overall Bayesian power increases as sample size enlarges. This presetation looks at an example of sample size estimation for modeling the relationship of breast cancer occurrence (misclassification) and calorie intake amount (measurement error).

email: dunleic@baylorhealth.edu

2g. SPATIAL AND FUNCTIONAL MODELING OF VARIABLITY IN TOOTH-LEVEL PERIODONTAL OUTCOMES

Thomas M. Braun*, University of Michigan

It is common practice in the analysis of periodontal data to collapse the tooth-level data of each subject into a single patient-level outcome, such as the overall (crude) mean among all teeth. However, the use of crude means is inefficient because such an approach not incorporate the tooth-level variability found in the original data. We propose a random effects model to estimate the within-mouth variability of teeth that incorporates both the spatial relationship of the teeth, as well as the shared functionality of subsets of teeth. Through this model, can we can assess the variability of several continuous and categorical periodontal measures, as well as determine whether biologic function or general location is a stronger predictor of the similarity of periodontal measures among teeth in the same mouth. We present some results for our model when fit to data collected from a recent longitudinal study of 50 healthy and 50 periodontitis patients.

email: tombraun@umich.edu

3. POSTERS: SURVIVAL ANALYSIS

3a. A NONLINEAR MODELING FRAMEWORK FOR ANALYZING TIME-TO-EVENT OUTCOMES FROM COMPLEX MULTI-STAGE SAMPLES

Scott W. Keith*, University of Alabama at Birmingham
David B. Allison, University of Alabama at Birmingham

Splines are a useful and exceptionally flexible tool for modeling nonlinear relationships in study variables. Estimating both the number and locations of knots as free parameters (join points in a 'free-knot' spline) can provide a fit to data that optimizes model parsimony. We have designed a free-knot spline framework for estimating nonlinear relationships between a censored time-to-event outcome variable and a continuous prognostic variable adjusted for relevant covariates in large nationally representative samples. We use direct search methods to maximize partial likelihood equations with piecewise linear free-knot splines having functional bases including the truncated power basis and B-splines. The model selection and inference methods incorporate parametric and nonparametric bootstrap methodology. The parameter estimates are structured for interpretability so that the results from our framework may be understood by investigators familiar with Cox proportional hazards survival modeling. Unlike other nonlinear modeling frameworks, our methods are capable of handling data from complex multistage sampling designs common to nationally representative surveys. Our presentation will include a summary of methodological details as well as results from a simulation study evaluating the performance of the framework under a variety of underlying conditions.

email: swkeith@uab.edu

3b. A NEW DISTRIBUTION FOR CUMULATIVE INCIDENCE FUNCTIONS

Sarah R. Haile*, University of Pittsburgh
Jong-Hyeon Jeong, University of Pittsburgh

The cumulative incidence function is of great importance when analyzing data where competing risks are present. We present a new distribution for parametric inference on competing risks. As the cumulative incidence function is used to model a subset of events, it is logical to model them using a distribution which is improper. The 4-parameter Gompertz distribution proposed is very flexible and permits several different hazard shapes, including unimodal, and can be extended to include covariates. The model is applied to data from National Surgical Adjuvant Breast and Bowel Project breast cancer trial B-14.

email: sah34@pitt.edu

3c. OUTLIER DETECTION AND GOODNESS OF FIT FOR RECURRENT EVENT MODELS WITH FRAILTY

Jonathan T. Quiton*, Western Kentucky University
Edsel A. Pena, University of South Carolina

Consider a recurrent event data where frailty models are used to account for differences in failure rates between units or subjects. In this context, we considered the problem of detecting outlying inter-event times or units, and the problem of deciding whether the data fits a chosen frailty model. We present a general test for outlier detection and goodness of fit using the mathematical framework similar to that of Neyman's smooth embedding. Large sample and finite sample results will be presented, and the procedures will be illustrated by applying to biomedical and engineering data sets.

email: Jonathan.Quiton@wku.edu

3d. SURVIVAL ANALYSIS BASED ON QUANTILE REGRESSION MODELS

Limin Peng*, Emory University
Yijian Huang, Emory University

Quantile regression offers great flexibility in assessing covariate effects on event times, thereby attracting considerable interests in its applications in survival analysis. However, currently available methods often require stringent assumptions or complex algorithms. In this paper we develop a new quantile regression approach for survival data subject to conditionally independent censoring. The proposed martingale-based estimating equations naturally lead to a simple algorithm that only involves minimizations of convex functions. We establish uniform consistency and weak convergence of the resultant estimators. Inferences are developed accordingly, including hypothesis testing, second-stage submodels, and model diagnostics. We evaluate the finite-sample performance of the proposed methods via extensive simulation studies. An analysis of a recent dialysis study illustrates the practical utility of our proposals.

email: lpeng@sph.emory.edu

3e. MINIMIZATION-BASED INTERVAL ESTIMATION OF HAZARD RATIO PARAMETERS

Kenichi Yoshimura*, National Cancer Center-Japan

We propose exact methods of interval estimation of proportional hazard parameter under Cox models. These methods are consistent to the randomization schemes that we used in the RCT. In oncology, minimization is exclusively used for randomization. We also propose the exact methods that are consistent to minimization. Simulation study was used for evaluation of the performance of our proposed methods. Our methods showed more favorable tendency than the asymptotic methods.

email: keyoshim@gan2.res.ncc.go.jp

3f. SURVIVAL ANALYSIS FOR MULTIVARIATE FAILURE TIME DATA

Zugui Zhang*, Christiana Care Health System
Paul Kolm, Christiana Care Health System
Edward F. Ewen, Christiana Care Health System
Claudine Jurkovitz, Christiana Care Health System
James Bowen, Christiana Care Health System
Joseph Jackson, Bristol-Myers Squibb
Dogan Fidan, Sanofi-Aventis Research-France
William Weintraub, Christiana Care Health System

Multiple failure time data frequently arise in biomedical studies when subjects can potentially experience several events. For example, patients who develop cardiovascular disease (CVD) may suffer a number of CVD events such as non-fatal myocardial infarction (MI), stroke or peripheral vascular disease. However, many analyses in published medical research journals involved only first events, ignoring additional failures, which may result in misleading conclusions. The purpose of this study was to apply multiple-event survival methods (MESM) to examine the risk of subsequent CVD events associated with increasing blood pressure. A total of 19,167 patients were selected from an electronic medical record encompassing office and hospital care. Results from three types of analyses were compared: single event survival analysis, unconditional multiplicative hazards model assuming independence of multiple events, and the conditional multiplicative hazards model assuming correlated multiple events. It was found that, across methods, hazard ratios were similar for mild and moderate hypertension categories. For severe hypertension, the conditional hazard ratio was greater than the single event and unconditional multiple event hazard ratios. Reasons for similarities and differences in results are discussed. The survival analysis for multivariate failure time data provides a more complete understanding of the burden of CVD.

email: ZZhang@ChristianaCare.org

3g. TESTING FOR SHAPE RESTRICTED HAZARD FUNCTION USING RESAMPLING TECHNIQUES

Desale Habtzghi*, Georgia College and State University
Somnath Datta, University of Louisville
Mary Meyer, Colorado State University

We consider testing the hypothesis that the lifetimes come from a population with a parametric hazard rate, such as Weibull against a shape restricted alternative, which comprises a broad range of hazard rate shapes, such as convex, concave or increasing. The alternative may be appropriate when the shape of parametric hazard is not constant and monotone. Most papers in the literature on shape restricted hazard rate function estimation, however, focus on testing the null hypothesis of a constant hazard rate versus the alternative of a nondecreasing. We compare the hazard function estimates obtained under shape restriction with its parametric counterpart using log rank and Kolmogorov's goodness-of-fit. We use appropriate resampling based computation to conduct our tests since the asymptotic distributions of the test statistics in these problems are largely intractable. We evaluate the performance of our approach via simulation studies and illustrate it on some real data sets.

email: desaleh@math.gcsu.edu

3h. SEMIPARAMETRIC METHODS FOR THE ANALYSIS OF CLUSTERED SURVIVAL DATA FROM CASE-COHORT STUDIES

Hui Zhang*, University of Michigan
Douglas E. Schaubel, University of Michigan
Jack D. Kalbfleisch, University of Michigan

The case-cohort study is a common and efficient design for large cohort studies. Most existing methods of analysis have been exclusively concerned with the analysis of univariate failure-time data. However, clustered failure time data are commonly encountered in public health studies. In this article, we propose methods based on estimating equations for three case-cohort designs for clustered failure time data. A marginal model is assumed, with a common baseline hazard and common regression coefficient across clusters. The proposed estimators are shown to be consistent and asymptotically normal. The estimators are easily computed using any standard Cox regression software that allows for offset terms. Consistent estimators of the asymptotic covariance matrices are presented. We show that these estimators have increased efficiency relative to some existing methods. The performance of the proposed estimators is investigated through simulation studies and the methods are applied to the analysis of liver transplant data.

email: huizh@umich.edu

4. POSTERS: CLINICAL TRIALS: GENERAL

4a. CASE EVALUATION OF OUTCOMES FOR STROKE: AN ILLUSTRATION OF THE SLIDING DICHOTOMY

Sharon D. Yeatts*, Medical University of South Carolina
Yuko Y. Palesch, Medical University of South Carolina
Scott W. Miller, Medical University of South Carolina

Common outcomes in phase III clinical trials in stroke, such as the modified Rankin Scale, are ordinal scales measuring functional outcomes. The statistical analysis of these outcomes, however, is often based on a fixed dichotomization approach; each subject is classified, based on the appropriate outcome scale, into one of two categories: good or bad outcome. Such an analysis, while easy to interpret, does not take into account the natural ordering of the response scale. It also ignores the notion that subjects with more severe strokes may have very little chance of achieving a good outcome. The sliding dichotomy approach tailors the definition of good or bad outcome based on each subject's

baseline prognosis. Two versions of the sliding dichotomy concept were applied to the data from a clinical trial of stroke (NINDS tPA study). The sliding dichotomy model as proposed in the literature determines the dichotomy based on a prognostic model using baseline factors (ie, age, severity of stroke, and time from symptom onset to treatment) as predictors. The second approach, a simplification, involves stratifying subjects using baseline data, and tailoring the definition of good outcome to each stratum. The performance of these approaches is compared with more traditional analysis methods, such as the fixed dichotomization described above.

email: yeatts@musc.edu

4b. EFFECTS OF HETEROGENEITY ON SIMON'S TWO-STAGE OPTIMAL DESIGN

Rebecca B. McNeil*, Mayo Clinic
Beth K. Rush, Mayo Clinic
Thomas G. Brott, Mayo Clinic
James F. Meschia, Mayo Clinic

The optimal Simon design has been recommended for phase II studies in stroke recovery. This design minimizes the expected sample size with respect to Type I and II errors under the null hypothesis. However, baseline participant heterogeneity may increase variation in the probability of successful recovery beyond nominal levels. We hypothesized that this overdispersion may alter the parameters of the study. Methods: We conducted simulations of this design as follows: for each simulated participant, the probability of success was randomly selected from either two or three values (representing study participant subgrouping), or was drawn from a beta prior distribution. Study termination rates and average sample sizes were estimated under the null and alternative hypotheses. Ten-thousand simulations were performed for each parameter set. Results: Negligible variation from nominal study parameters was observed when participants were equally likely to have been assigned each probability of success, or when the probability of success was drawn from a representative beta distribution. When the probability of a favorable outcome was skewed, the Type I and II error rates, as well as average sample sizes, may vary from nominal levels to a worrisome extent. In this situation, adjustment of the design parameters may be justified.

email: mcneil.rebecca@mayo.edu

4c. SENSITIVITY AND SPECIFICITY OF IMMUNOLOGICAL CORRELATES OF PROTECTION

Yunjie Chen*, Virginia Polytechnic Institute and State University
Andrew J. Dunning, Sanofi Pasteur

Immunological assays measure characteristics of the immune system, such as antibody concentrations, that are believed to be associated with protection from disease. In vaccines research, there is considerable interest in determining the threshold assay value that confers protection from disease. Existing approaches estimate the threshold assay value by modeling the relationship between the disease status and assay values. In this research, links are established between the threshold assay value and diagnostic screening accuracy measures, sensitivity and specificity. The losses associated with the two types of misclassification measured by 1-sensitivity and 1-specificity are accounted for. The problem of exposure is examined and a lower bound established. Results are illustrated using data from the Swedish Pertussis household trial (1992-1995) and the robustness of the estimated thresholds is assessed.

email: yjchen3@vt.edu

4d. A COMPARISON OF THE POWER OF SEVERAL TESTS FOR DETECTING QUALITATIVE TREATMENT BY COVARIATE INTERACTIONS IN CLINICAL TRIALS

Scott W. Miller*, Medical University of South Carolina
Yuko Y. Palesch, Medical University of South Carolina
Renee H. Martin, Medical University of South Carolina
Peng Huang, Medical University of South Carolina

Phase III randomized, controlled clinical trials are the definitive method to determine the safety and efficacy of an agent or strategy to prevent or ameliorate a disease. It is possible that the effect of the intervention under investigation may vary across different levels of some clinically important covariate; for example, the treatment may have a significant effect in males but only a marginal effect in females: a quantitative interaction. More clinically important are qualitative interactions; for example, the treatment may have a significantly beneficial effect in males but a significantly harmful effect in females. Several methods to detect such differences exist (the likelihood ratio test, the pushback method, the range test and the extended range test). While some have been studied, the operating characteristics of the pushback method in particular have not been previously explored. We conducted a simulation study to compare the statistical operating characteristics of these methods. The effect of the number of subgroups, the types of qualitative differences and the magnitude of those differences were examined. We find that the pushback method using the normal distribution is highly anticonservative, while the pushback method using the t distribution compares more favorably with the other tests..
email: millersw@musc.edu

4e. ON TESTS FOR QUALITATIVE INTERACTION

Robert A. Parker*, Amgen

I reconsider the underlying structure for a test of qualitative interaction of a treatment across various subsets. Such a test should require that the overall probability of the positive and negative results combined should equal the Type I error selected for the test. The standard test for qualitative interactions (Gail and Simon, Biometrics, 1985;41:361) and subsequent developments of this approach do not meet this criterion. For example, for the case of two subsets, with a nominal Type I error of 0.05, the chance of a significant finding when all treatment groups are zero (e.g., for example when the effect is centered) would be 0.005 (2 x 0.05 x 0.05), which is only one-tenth of the nominal value. Simulations of the null case confirm this finding, and suggest that the criterion tabled in the Gail-Simon paper are very conservative. These results help explain why the test is known to have very low power to detect interaction. In contrast, the lesser known test originally proposed by Azzalini and Cox (J. R. Statist Soc B, 1984:46:335) does seem to be consistent with the proposed interpretation.

email: raparker@amgen.com

4f. STATISTICAL ASSESSMENT OF MEDICATION ADHERENCE DATA: A TECHNIQUE TO ANALYZE THE J-SHAPED CURVE

Jeffrey M. Rohay*, Pinney Associates
Gary M. Marsh, University of Pittsburgh
Stewart Anderson, University of Pittsburgh
Vincent C. Arena, University of Pittsburgh

Ada Youk, University of Pittsburgh
Jacqueline Dunbar-Jacob, University of Pittsburgh

Medication non-adherence is a major reason for a lack of effectiveness of efficacious drug regimens. Adherence distributions are J-shaped with many taking their medication completely, a large number taking none, and a small number taking it intermittently. Comparisons are typically made using parametric and non-parametric techniques, or by dichotomization, which all have limitations. Parametric techniques may be inappropriate as the assumptions (e.g., normality) are violated and transformations fail; Central tendency measures and non-parametric techniques do not adequately depict the distribution; and, dichotomization results in information loss, making small improvements indiscernible. We propose an analytic technique using a beta mixture characterized by pb(1,b)+(1-p)b(a,1), producing a J-shaped distribution. The proportion of values in each tail will be determined by p. The EM algorithm will be used to produce parameter and standard error estimates. This technique will allow for the description of the distribution's shape and comparisons of parameter estimates for two distributions to determine if two interventions differ. We will assess, via simulation studies, alpha levels and power for this new method as compared to other standard methods.

email: jrohay@pinneyassociates.com

4g. BIOEQUIVALENCE AND PHARMACOKINETICS WITH THE USE OF WINNONLIN

Anna M. Maeser*, University of North Carolina-Wilmington and Biostudy Solutions, LLC

The Food and Drug Administration has developed strict requirements with regards to the pharmaceutical regulatory process and the bioequivalence of generic drugs to drugs already existing in the marketplace. This session will explore the analysis of generic drugs based on pharmacokinetic parameters such as the rate and extent of absorption, the terminal rate of elimination, and the terminal half-life of elimination in order to meet these FDA requirements. The usage of the computer program WinNonlin will be discussed in depth and a simulation of a noncompartmental analysis will be presented. The notions of oral drugs and topical corticosteroids will also be addressed.

email: AMM1928@uncw.edu

4h. CROSSOVER DESIGNS UNDER SUBJECT DROPOUT

Shi Zhao* and Dibyen Majumdar, University of Illinois-Chicago

Crossover experiments are used for comparing the responses to various different stimuli or treatments in areas ranging from psychology and human factor engineering to medical and agricultural applications. They are widely used in the pharmaceutical industry. There is an extensive literature that assures us that a carefully designed crossover study will produce a wealth of information that will enable inference with high precision. This is based on the implicit, but critical, assumption that the experiment will yield all the planned observations. Yet in many situations, such as clinical trials, there is a substantial probability that some subjects will drop out of the study prior to the completion of their treatment sequence. We will study UBRMDs under the subject dropouts. (1) We will derive the best UBRMDs for the situation where all subjects may drop out in the final period and provide methods for constructing these designs. (2) We will study UBRMDs under subject dropout for the model where subject effects are random and show that the Low, Lewis and Prescott (1999) result on lack of connectedness of the Williams Latin Square of order 4 is no longer valid. Compound symmetry and AR (1) covariance structures will be considered. (3) Expected loss under various dropout probabilities will be studied.

email: szhao5@uic.edu

5. POSTERS: CLINICAL TRIALS: ADAPTIVE DESIGN AND ADAPTIVE RANDOMIZATION

5a. GLMIP 1.0: SAS/IML SOFTWARE FOR PLANNING REPEATED MEASURES INTERNAL PILOTS

Keith E. Muller*, University of Florida
Christopher S. Coffey, University of Alabama at Birmingham
John A. Kairalla, University of Florida
Jacqueline L. Johnson, Novartis

Internal pilot designs involve conducting interim power analysis (without interim data analysis) to modify the final sample size. Recently developed techniques avoid the type I error rate inflation inherent to unadjusted tests, while still providing the advantages of an internal pilot design. The free software GLMIP allows planning internal pilot studies in the multivariate general linear model with Gaussian errors, using the univariate approach to repeated measures. Such models are equivalent to a useful range of mixed models. The multivariate approach is planned in future versions. The GLMIP software makes it easy to perform approximate power analysis and sample size selection for repeated measures internal pilots with excellent accuracy even in very small samples (N=10).

email: Keith.Muller@ufl.edu

5b. A TWO-STAGE PROCEDURE FOR FACTORIAL DESIGNS UTILIZING A FORMAL FUTILITY RULE

Nitin K. Nair*, University of Alabama at Birmingham
Christopher S. Coffey, University of Alabama at Birmingham

Determination of sample size for factorial clinical trials is complicated by the fact that one is forced to make an a priori decision as to whether an interaction exists. To address this issue, Russek-Cohen and Simon (1997) introduced a two-stage procedure using a formal hypothesis test for the interaction to determine whether to stop at the first stage (if there is no interaction and effects can be combined across groups) or to proceed to a second stage (if there is an interaction and effects must be examined separately). If the test suggests that a second sample is required, it is taken as long as the estimated stratum-specific treatment effect is positive. An attractive alternative might be to assess the futility of second stage using conditional power. We present software to compute the exact probabilities and assess the performance of a modified two-stage procedure using conditional power calculated under three assumptions for future trend: 1) design alternative, 2) current observed trend, and

3) current observed trend plus one standard error. The performance of each of the three modified procedures will be summarized and compared to the original procedure.

email: nnair@uab.edu

5c. ESTIMATION AND HYPOTHESIS TESTING PROPERTIES FOR OUTCOME-BASED ADAPTIVE RANDOMIZATION IN CLINICAL TRIALS WITH BINARY OUTCOMES --- A SIMULATION STUDY COMPARING URN MODELS AND SEQUENTIAL ESTIMATION METHODS

Xuemin Gu*, M. D. Anderson Cancer Center
J. Jack Lee, M. D. Anderson Cancer Center

Equal randomization (ER) is the default standard in randomized controlled trials for comparing treatment efficacy. However, ER focuses on collective ethics and neglects intermediate results as the trial progresses. On the other hand, outcome-based adaptive randomization (AR) methods can improve the ethical concern by allocating more patients in the more effective arms based on the interim results, and yet still be able to make valid statistical inferences. Extensive simulations have been conducted on trials with moderate to small number of patients to investigate the statistical properties of two urn models: randomized play-the-winner and drop-the-loser, both of which have a limiting allocation ratio of relative risk and three sequential estimation methods: sequential maximum likelihood estimation, doubly adaptive biased coin design, and sequential estimation adjusted urn, all of which can target different allocation ratios. For binary outcome data, we compare the performance of these five AR methods with ER on parameter estimation and hypothesis testing. The estimation of response rate, allocation ratio, and overall response rate are compared among different randomization procedures. The power of hypothesis testing statistics are also compared at the same type I error rate. Recommendations on the choice of adaptive randomization procedures are discussed.

email: xuegu@mdanderson.org

5d. CAN RESPONSE-ADAPTIVE RANDOMIZATION BENEFIT YOUR TRIAL?

Amy S. Nowacki*, Medical University of South Carolina
Wenle Zhao, Medical University of South Carolina

A method is proposed to aid in the decision of whether or not response-adaptive randomization will benefit a trial. This is achieved by exploring the relationship between test power and the expected number of additional successes. If deemed beneficial, this method identifies a target allocation resulting in no loss of power over equal allocation, but a higher success rate. This target allocation can be implemented in any sequential response-adaptive randomization scheme.

email: bardeen@musc.edu

5e. OPTIMAL ADAPTIVE GROUP SEQUENTIAL DESIGN FOR PHASE II CLINICAL TRIALS: A BAYESIAN DECISION THEORETIC APPROACH

Yiyi Chen*, University of Iowa
Brian J. Smith, University of Iowa

Bayesian decision theoretic approaches have been widely studied in the literature as tools for designing and conducting phase II clinical trials. However, full Bayesian approaches that consider multiple endpoints are lacking. Since the monitoring of toxicity is a major goal of phase II trials, we propose an adaptive group sequential design using a Bayesian decision theoretic approach, which characterizes efficacy and toxicity as correlated bivariate binary endpoints. We allow trade-off between the two endpoints to a certain extent. Interim evaluations are conducted group sequentially, but the number of interim looks and the size of each group are chosen adaptively based on current observations. We utilize a loss function consisting of two components: the cost associated with accruing, treating, and monitoring patients, and the loss associated with making incorrect decisions. The operating characteristics of the design are evaluated over a range of parameter values assigned to the loss function. We also examine the robustness of prior specifications to the decision rule, and evaluate the effect of a misspecified correlation between efficacy and toxicity. Our method is illustrated in the context of a single-arm phase II trial of bevacizumab, gemcitabine, and oxaliplatin in patients with metastatic pancreatic adenocarcinoma.

email: yiyi-chen@uiowa.edu

6. POSTERS: APPLIED DATA ANALYSIS

6a. BIMODALITY AND VARIANCES: BUMP HUNTING IN MOTOR CONTROL OF PARKINSON'S DISEASE

Sue Leurgans*, Rush University Medical Center
Julie Robichaud, University of Illinois-Chicago
David Vaillancournt, University of Illinois-Chicago
Daniel Corcos, University of Illinois-Chicago

Familiar measures of central tendency and measures of spread do not give accurate summaries of empirical distributions that contain multiple modes. Kernel density estimates can be adapted to provide more faithful descriptions. Silverman (1981) showed that the smallest bandwidth resulting in a unimodal kernel density estimate can be used to search for multimodality, or bumps. We applied this methodology to the electromyographic signals measured during a study which examined rapid voluntary elbow flexion movements in patients with Parkinson's disease and in healthy control subjects. When many repeated movement trials are available for an individuals, grid searches on bandwidth parameters provides a bandwidth measure from the electromyographic signal that shows differences in patterns that distinguish individual Parkinson's Disease patients from control subjects. The bandwidth measure is sensitive to bimodality and reveals insight that is hidden if averages are used to summarize repeated trials and that is difficult to discern from measures of spread

email: Sue_E_Leurgans@rsh.net

6b. ORDER-PRESERVING DIMENSION REDUCTION PROCEDURE FOR THE DOMINANCE OF THE TWO MEAN CURVES WITH APPLICATION OF TIDAL VOLUME CURVES

Sang Han Lee*, Nathan Kline Institute
Johan Lim, Yonsei University
Marina Vannucci, Rice University
Eva Petkova, New York University
Maurice Preter, Columbia University
Donald K. Klein, Columbia University

The work we present in this paper was motivated by a case study involving high-dimensional and high-frequency tidal volume traces measured during induced panic attacks. Our focus was to develop a procedure to determine the significance of whether a mean curve dominates another one. The key idea of the suggested method relies on preserving the order in mean while reducing the dimension of the data. Dimension reduction is achieved by projecting the observed data matrix onto a set of lower rank matrices with a positive constraint. A multivariate testing procedures is then applied to the coefficient vectors that represent the data matrix in a lower dimension. We provide an iterative algorithm to solve the projection problem. Our procedure is simple and powerful. We use simulated data to illustrate its statistical properties. Our results on the case-study data confirm the preliminary hypothesis of the investigators and provide critical support to their overall goal of creating an experimental model of the clinical panic attack in normal subjects.

email: hanul31@stat.tamu.edu

6c. ADJUSTMENT OF SYSTEMATIC MEASUREMENT ERROR IN HUMAN BODY SURFACE AREA MEASURED BY THREE DIMENSIONAL SCAN

JungBok Lee*, Korea University
Young Ju Kim, Kangwon University
Dongdeuk Jang, Korea National Institute of Toxicological Research
Bong Hyun Nam, Korea National Institute of Toxicological Research
Eunhee Kim, Korea National Institute of Toxicological Research

For the purpose of nationwide measurements of human body surface (BSA) which are fundamental data for risk assessments of drug, cosmetcis etc, Korea national institute of toxicological research prepared nationwide measurements plan of BSA. Several measurement methods of BSA are proposed such as a coating method using paper, inelastic tape, plaster bandage or aluminum foil, surface integration and an alginate method. These methods were too laborious and time consuming to measure for large populations. As an alternative, three dimensional whole body scan provided BSA estimates with easy and fast. However, there were systematic measurement error between 3D scan and the other methods stated above. In this study, we evaluate accuracy of 3D scanned BSA data and propose adjustment modeling approach of systematic measurement variations between 3D scan and alginate method.

email: jungboklee@korea.ac.kr

6d. STATIONARY STATE OF THE POPULATION AFTER A SCREENING INTERVENTION: AN M(T)/G/INF NETWORK MODEL

Shih-Yuan Lee*, University of Michigan School of Public Health
Alexander Tsodikov, University of Michigan School of Public Health

A population under screening for prostate cancer is considered as an M(t)/G/Inf queuing network. Birth process in the population and cancer incidence, the system's input and output, are modeled as non-homogeneous Poisson processes with infinite number of servers. The service time distribution is general but finite. Improper distributions (cure models) as well as competing risks are used to model over-diagnosis of the disease. The disease free stage and the pre-clinical disease stage represent two series-connected servers. Stationary states of the system before and after the intervention are derived and linked by a regression model with intervention process characteristics as covariates. Estimation and inference procedures for the model are developed. Effects of secular trends in disease characteristics on the system are assessed. Surveillance, Epidemiology and End Results data as well as other data sources are used to fit the model.

email: shihylee@umich.edu

6e. A GENERALIZED GROWTH MIXTURE MODELING APPROACH TO PREDICT INSTITUTIONALIZATION OF PEOPLE WITH ALZHEIMER'S DISEASE (AD) FROM ANXIETY AND DEPRESSION TRAJECTORIES OF CAREGIVERS

Song Zhang*, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh
Steven H. Belle, University of Pittsburgh
Richard Schulz, University of Pittsburgh

Medical and social-behavioral studies suggest that study populations are not usually homogenous and can be characterized by qualitatively different patterns of selected characteristics over time. It is also often the case that an outcome event is related to the configuration of a targeted set of correlated measurements observed over time. A generalized mixture modeling (GMM) approach can be used to identify subpopulations with distinct trajectory patterns that predict subsequent outcomes. This modeling approach is illustrated using data from the ResourceS for Enhancing Alzheimer's Caregiver Health (REACH) study. We applied a GMM to identify the subpopulations characterized by the trajectory patterns of the interrelated anxiety and depression among caregivers who provided care to family members with Alzheimer's disease. A piecewise linear model was applied to take into account the nonlinear growth trend at different times following study entry. The membership determined by the growth trajectory patterns was used to predict the probability of the care recipients being institutionalized. Three subpopulations were identified and the features of their trajectories are found to be substantively informative in predicting care recipient institutionalization.

email: zhangs@edc.pitt.edu

6f. ECOLOGICAL INFERENCE IN META-ANALYSIS

Mireya Diaz*, Case Western Reserve University

Ecological inference manifests in meta-analysis when the distribution of an influential characteristic at the individual level is provided at an aggregated (study) level and then linked to an outcome that is, there is already ecological associations at the study level. The resulting association will be valid as long as the outcome proportions are equal, information unfortunately unknown, unless the overall estimate remains constant for any allocation ratio. In general the validity of ecological inferences in this missing strata case will depend on two sources of information, (1) whether the overall estimate is homogeneous or heterogeneous, and (2) whether there is a set of studies assessing outcomes throughout a wide range of allocation ratios. Homogeneity could be verified if a portion of the studies have unique groups (i.e. allocation ratios of 0 or 1). Conditions in the two settings of homogeneity and heterogeneity under which ecological inferences are valid will be examined.

email: mcd8@cwru.edu

6g. HIERARCHICAL MODELING OF REMOVAL EXPERIMENTS TO ESTIMATE CATCHABILITY OF BLUE CRAB IN VIRGINIA

Mary Christman, University of Florida
Xiaobo Li*, University of Florida
Thomas Bohrmann, University of Florida

To estimate blue crab (Callinectes sapidus) population dynamics in the Chesapeake Bay, a dredge survey has been conducted annually in winter between 1992 and 2007. As part of this survey removal experiments have been conducted in several years at randomly selected locations throughout the Bay. The removals take place during the winter when the crabs are extremely lethargic and thus it is reasonable to assume a closed population during the tows which occur in a short time span. These removal experiments generally consist of six rounds of a dredge dragged over an area of approximately 550 square meters. In addition, depth, vessel and age of crabs (younger than one year or older than one year) were recorded as potential covariates. A hierarchical Bayesian model is used to estimate the abundance and catchability of the crab of each removal experiments at each year. Catchability is assumed constant at a location over the six rounds but varying among locations depending on the covariates. The abundance at each location is assumed to have a Poisson distribution with its parameter drawn from a flat uniform prior. The number of crabs caught in the kth round follows a binomial distribution of catchability from a size of sum of the catches in the k-1 previous rounds. The catchability is assumed to be Beta distributed with the mean a function of the covariates depth and vessel. In order to further test if the crab age has an effect on catchability, we compare estimates from the same model using the total crab data and only the mature crab data.

email: xbli@ufl.edu

7. POSTERS: LATENT VARIABLES AND MISSING DATA

7a. BAYESIAN LATENT VARIABLE MODELING OF OUTCOMES FOR BIPOLAR DISORDER

Brian Neelon*, Harvard Medical School
A. James, O'Malley, Harvard Medical School
Sharon-Lise Normand, Harvard Medical School

Researchers often monitor symptoms and functioning longitudinally to determine how bipolar patients respond to treatment. Patients may be clustered into underlying classes of response trajectories that depend on patient characteristics. The outputs of interest include predicting class membership and estimating a mean response curve for each class. A number of interesting practical issues arise from such endeavors, including choosing the number of classes; jointly modeling multiple, possibly mixed, outcomes; and capturing between-subject heterogeneity within each class. Using a Bayesian framework, we consider several approaches to analyzing such data, including latent class models, growth mixture models and shape-constrained mixture models. These models are applied to a multi-site, longitudinal study of a bipolar cohort. This work is funded by a grant from the National Institute of Mental Health (R01-MH61434)

email: neelon@hcp.med.harvard.edu

7b. USING CONCORDANCE CORRELATION AND MULTIPLE IMPUTATION TO COMBINE TWO DIFFERENT ASSESSMENTS OF THE SAME CONSTRUCT

Christopher J. Swearingen*, Medical University of South Carolina

Two nearly identical Parkinson's disease clinical trials ascertained participant caffeine consumption using different, previously non-validated assessments, complicating an ancillary analysis of caffeine's possible effect on treatment outcomes. A concordance analysis was used to determine the agreement between the assessments. The concordance analysis was also used to validate an extension of multiple imputation methods, where known responses of one assessment estimate the missing responses of the other assessment and vice versa, by comparing post-imputation concordances. Post-imputation models are presented for each caffeine measure, assessing caffeine's possible effect on a selected outcome, as well as a model utilizing observed data recorded from each assessment, demonstrating the utility of this imputation.

email: swearinc@musc.edu

7c. LONGITUDINAL ASPECTS OF NONRESPONSE IN THE SURVEY OF INDUSTRIAL RESEARCH AND DEVELOPMENT

Adriana Pérez*, University of Lousiville

The Survey of Industrial Research and Development (SIRD) is important because provides statistics on research and development (R&D) by companies in the United States, which is used to evaluate the status of science and technology diffusion in the United States relative to other nations. Missing data is present in many of the SIRD key variables. Several imputation techniques (auxiliary ratio and auxiliary trend) are implemented to overcome missingness. Currently, year by year estimates are reported for SIRD as if imputation values were observed values. A maximum likelihood estimate (MLE) of the data accounting for the uncertainty due to missing values is proposed from a longitudinal perspective of companies in years 2002 & 2003. Empirical and simulations studies were developed to compare the results of current and proposed estimators through precision and accuracy measures (bias and mean square error).

email: adriana.perez@louisville.edu

7d. THE USE OF SAMPLE WEIGHTS IN HOT DECK IMPUTATION

Rebecca R. Andridge*, University of Michigan
Roderick J. Little, University of Michigan

A common strategy for handling item nonresponse in survey sampling is hot deck imputation, where each missing value is replaced with an observed response from a 'similar' unit. We discuss here the use of sampling weights in the random hot deck. The naive approach is to ignore sample weights in creation of adjustment cells, which effectively imputes the unweighted sample distribution of respondents in an adjustment cell, potentially causing bias. Alternative approaches have been proposed that use weights in the imputation by incorporating them into the probabilities of selection for each donor. We show by simulation that these weighted hot decks do not correct for bias when the outcome is related to the sampling weight and the response propensity. The correct approach is to use the sampling weight as a stratifying variable alongside additional adjustment variables when forming adjustment cells.

email: fedarko@umich.edu

7e. USING A MIXTURE MODEL FOR MULTIPLE IMPUTATION IN THE PRESENCE OF OUTLIERS

Michael R. Elliott*, University of Michigan

To obtain population-based inference in the presence of missing data and outliers, we develop a latent class model that assumes each observation belongs to one of K unobserved latent classes, with each latent class having a distinct covariance matrix. The latent class covariance matrix with the largest determinant is assumed to form an outlier class, and we conduct inference after removing these outliers. As in Ghosh-Dastidar and Schafer (2003), we use multiple imputation to promulgate uncertainty in the outlier status. We extend their work by embedding the outlier class in a larger mixture model, consider penalized likelihood and posterior predictive distributions to assess model choice and fit, and construct the model to account for complex sample designs. We apply our methods to estimate obesity prevalence and body-mass index (BMI) measures in the Healthy For Life Study.

email: mrelliot@umich.edu

7f. ESTIMATION IN HIERARCHICAL MODELS WITH INCOMPLETE DATA

Yong Zhang*, University of Michigan
Trivellore E. Raghunathan, University of Michigan

Hierarchical models are often used when data are observed at different levels and the interaction effects of variables measured at different levels on the outcome are of interest. Missing data can occur at all levels and in both outcomes and covariates. Ignoring the subjects with missing data generally leads to biased estimates, yet methods for analyzing incomplete data using hierarchical models are less developed. We develop an approach that combines the features of the EM algorithm and multiple imputations assuming missing at random (MAR) and ignorable missing mechanism (Rubin, 1976; Little and Rubin, 2002). Simulation study is used to demonstrate that our proposed method has desirable repeated sampling properties. The method is also applied to a national high school survey data.

email: yonzhang@umich.edu

7g. MODELING SENSITIVITIES AND SPECIFICITIES FOR LONGITUDINAL DATA: APPLICATION TO PSYCHOSOCIAL RESEARCH

Qin Yu*, University of Rochester
Wan Tang, University of Rochester
Yan Ma, University of Rochester
Xin Tu, University of Rochester

Analysis of accuracy of diagnostic test is used in a wide range of behavioral, biomedical, psychosocial, and health-care related research. Test sensitivity and specificity are the most popular measures of accuracy for diagnostic tests. In this talk, we consider inference for sensitivity (specificity) when both the reference standard and test outcome are subjected to missing within a longitudinal data setting. We develop an inverse probability weight (IPT) approach to address the inference issue and discuss improvement of efficiency using augmented IPT-based or double robust estimates. The approach is illustrated with real data in sexual health research.

email: qin_yu@urmc.rochester.edu

8. POSTERS: STATISTICAL MODELS AND METHODS

8a. GROWTH RATE ESTIMATES ASSUMING A WEIBULL GROWTH CURVE

Alaina M. Houmard*, University of North Carolina-Wilmington

We have witnessed an increase interest in estimating an ‘average’ growth rate under a sigmoidal growth curve. This paper illustrates a method to estimate an ‘average’ growth rate when assuming a Weibull growth curve. In addition, we show how to calculate a confidence interval for this growth rate estimate. Simulation studies will show the overall effectiveness of this model.

email: amh6957@uncw.edu

8b. A COMPARISON STUDY OF GLOBAL MODELS AND CONFIDENCE INTERVALS UNDER THE EXPONENTIAL GROWTH MODEL

Lifang Du*, University of North Carolina-Wilmington

A new and creative way of testing equivalence between two models is known as Global Model testing. In this session, we compare the effectiveness of global models to the more conventional method of constructing confidence intervals around the parameters, A simulation study demonstrate the performance of these two methodologies.

email: dulifang@gmail.com

8c. USING A STOCHASTIC SEARCH IN A BAYESIAN HIERARCHICAL REGRESSION MODEL

Qijun Fang*, University of North Carolina-Wilmington

Many datasets inherently have a complex multi-level structure that needs to be accounted for. Numerous literature cite examples where hierarchical models fit these complex structures well. In this session, we illustrate how to use a stochastic search to identify important predictors in a Bayesian hierarchical regression model where p>n.

email: fangqijun33@hotmail.com

8d. LINEAR REGRESSION MODELS FOR SYMBOLIC INTERVALVALUED DATA AND THE CONFIDENCE INTERVAL

Wei Xu*, University of Georgia
Lynne Billard, University of Georgia

Symbolic data is represented by p-dimensional hypercubes in Rp, or a Cartesian product of p distributions. It includes single, multi-valued,

interval-valued, histogram, and even distribution data. Unlike classical data, it usually has internal variations. Recently, several linear regression models have been introduced to analyze interval-valued data. However, depending on the values of the estimated regression coefficients, they either can give predictions with the lower bounds larger than the upper bounds, or force the parameter estimators to be non-negative which do not necessarily reflect the true relationship between dependent and independent variables. We propose an alternative linear model to solve these problems and to obtain better predictions. We also introduce a confidence interval calculation method based on interval-valued data covariance matrix cross product structure. The proposed method is evaluated and compared with previous methods in the frame work of a Monte Carlo experiment and a real data-set application.

email: xuwei@uga.edu

8e. BAYESIAN ROC CURVE ESTIMATION UNDER BINORMALITY USING A PARTIAL LIKELIHOOD BASED ON RANKS

Jiezhun Gu*, Duke Clinical Research Institute
Subhashis Ghosal, North Carolina State University

There are various methods to estimate the parameters in the binormal model for the ROC curve. In this paper, we propose a conceptually simple and computationally accessible Bayesian estimation method using a partial likelihood based on ranks. Posterior consistency is also established. We compare the new method with other estimation methods and conclude that our estimator generally performs better than its competitors.

email: sherrygu2001@yahoo.com

8f. THE TEST OF STOCHASTIC ORDERING TO DETECT TREATMENT RELATED TREND WITH CLUSTERED DISCRETE DATA

Kyeongmi Cheon*, University of Memphis
Aniko Szabo, Medical College of Wisconsin
Ebenezer O. George, University of Memphis

The definition of treatment related trend in studies with clustered data has always been ambiguous. And treatment effects are defined in terms of either per cluster member response rate or in terms of per whole cluster response rate. We introduce the concept of stochastic ordering to define treatment related trend in clustered exchangeable multinomial data. Most existing definitions are special cases of this definition, and its robustness incorporates various forms of monotone responses. We use a saturated model of the multinomial exchangeable distribution by George, et al. (2007) to model the joint distribution of the endpoints including death/resorption and malformation. The saturated model representation ensures that our procedure is essentially nonparametric. We address the problem of sparseness and random cluster sizes by using the marginal compatibility assumption of Pang and Kuk (Biometrics, 2006). We device an EM algorithm by Lindsay (Ann. of Stat, 1983) and Hoff (J. of Comp. and Graphical Stat., 2000) for MLE computation under stochastic ordering. We compare our work to the existing procedures, including the exact method of Corcoran et al. (Biometrics, 2001).

email: katie.cheon@gmail.com

8g. AN ALTERNATIVE METHOD FOR ANALYZING 2 X C TABLES

Julia L. Sharp*, Clemson University
John J. Borkowski, Montana State University

Fisher’s Exact Test is predominate in the analysis of 2 x c tables. When the number of replicates is small, Fisher’s Exact Test can be performed by hand. The definition of ‘as or more extreme’ and the calculation of the p-value in Fisher’s Exact Test relies on tables with hypergeometric probabilities smaller than or equal to the hypergeometric probability observed. In some instances, Fisher’s Exact Test may consider some tables that are not of interest to researchers in the computation of the p-value. Combinatorial generating functions computed for a conditional frequency distribution are considered as an alternative to Fisher’s Exact Test when equal and small replicates are used in each analyzed category. An example using affinity isolation experimental data from Pacific Northwest National Laboratory will be presented.

email: jsharp@clemson.edu

8h. MODEL-BASED CANONICAL CORRELATION ANALYSIS FOR FUNCTIONAL DATA

Hyejin Shin*, Auburn University
Seokho Lee, Texas A&M University

Analysis of accuracy of diagnostic test is used in a wide range of behavioral, biomedical, psychosocial, and health-care related research. Test sensitivity and specificity are the most popular measures of accuracy for diagnostic tests. In this talk, we consider inference for sensitivity (specificity) when both the reference standard and test outcome are subjected to missing within a longitudinal data setting. We develop an inverse probability weight (IPT) approach to address the inference issue and discuss improvement of efficiency using augmented IPT-based or double robust estimates. The approach is illustrated with real data in sexual health research.

email: hjshin@auburn.edu

9. POSTERS: STATISTICAL GENETICS

9a. MULTIPLE IMPUTATION FOR FAMILY ASSOCIATION STUDIES

Miguel A. Padilla, University of Alabama at Birmingham
Howard Wiener, University of Alabama at Birmingham
Donald Rubin, University of Alabama at Birmingham
Hemant K. Tiwari*, University of Alabama at Birmingham

Little has been done to handle missing data problems in genetic association analysis using pedigree data. Recently, multiple imputation has become a powerful alternative to traditional missing data methods. Here we present multiple imputation as a tool that can be used to impute probable values within each of the families using genotypic and phenotypic information in the linear model. In this context, the imputation is more general in that phenotypic and/or covariate data can be imputed jointly. The proposed method takes into account both the within and between familial information. A Gibbs sampler imputation scheme is presented within the context of a polygenic model. We performed a simulation study to investigate bias and MSE of the estimates, and the frequentist validity (type I error rate) and power of the association test before and after adjusting for missing data.

email: htiwari@uab.edu

9b. IMPUTATION OF MISSING GENOTYPES: AN EMPIRICAL EVALUATION

Zhenming Zhao*, Boston University
Paola Sebastiani, Boston University

We tested and evaluated the quality and accuracy of the program Impute, which imputes unobserved genotypes in genome-wide case-control studies based on a set of known haplotypes. The evaluation set consisted of real genotype data of 5,872 SNPs on Chromosome 21 of 271 subjects. We randomly removed genotype data of 0.1%, 1% or 10% of the original SNPs and the average accuracies of the imputed genotypes were 97.42%, 97.41% or 97.05% in 1,000 simulations. By using 0.95 as minimum posterior probability of a match, the accuracies increased to 99.24%, 99.24% or 99.22%. If still selecting 0.1%, 1% or 10% of SNPs, but only 80% genotypes for each selected SNP were missing, the imputation accuracies were slightly higher than the imputations above. To emulate combining genotype data from Illumina 370K chip and Affymetrix 500K chip, we randomly removed 40% or 60% of SNPs, and we observed good imputation accuracies as 95.20% or 91.88%, and 99.06% or 98.86% if using the 0.95 threshold. We also found there was a correlation between imputation accuracy and linkage disequilibrium. Our results confirmed the high accuracy of Impute program, and we are optimistic about the future applications of imputations on combining data from different studies and different platforms.

email: zmzhao@bu.edu

9c. POPULATION GENETICS OF BLUEFISH

Christina M. Stuart*, University of North Carolina-Wilmington

Bluefish, Pomatomus saltatrix, are common along the east coast of the United States and are an important species to the commercial and recreational fishing industry. Bluefish have a unique spawning behavior in that they produce two cohorts of juveniles, or young of the year (YOY), each year. This study is being done to determine if the two cohorts, termed spring spawned and summer spawned, are genetically different. The D-loop region of the mitochondrial DNA will be used to determine if there is a significant difference between the two cohorts. The YOY bluefish were collected from estuaries and surf zones around Wrightsville Beach, NC using haul seines. Tissue samples were taken from each bluefish for DNA extraction, amplification and sequencing of the D-Loop region. The results of this ongoing project will be presented at the meeting.

email: cms7288@uncw.edu

9d. A COMPUTATIONAL PROCEDURE FOR IDENTIFYING ESTROGEN RECEPTOR BINDING SITES IN HUMAN OSTEOSARCOMA CELLS USING HIGH RESOLUTION TILING ARRAYS

Hui Tang*, Mayo Clinic

The mechanisms by which estrogen interacts with the estrogen receptor to regulate gene expression are poorly understood and only a small number of regulatory elements activated by estrogen have been identified before. Our collaborators conducted a set of expression microarrays and whole genome tiling arrays on estrogen-treated human osteosarcoma cells to study the regulatory mechanism. We have developed a computational procedure to streamline the processes of quality assessment, binding interval prediction, sequence retrieval, and regulatory element discovery using these microarrays, and have found several important elements previously identified by other studies, including the classical ERE, AP-1 and SP1, Oct and C/EBP. We have also found some elements which have not been characterized as estrogen activated motifs and some unknown motifs which can be further examined by PCR and ChIP assay.

email: h.tang@mayo.edu

9e. GENETIC ASSOCIATION STUDIES USING SAMPLES ASCERTAINED ON THE BASIS OF A CORRELATED TRAIT

Genevieve M. Monsees*, Harvard School of Public Health
Peter Kraft, Harvard School of Public Health

Large cohorts with prospectively-measured biomarkers are very expensive. Utilizing existing data from nested case-control studies may reduce ascertainment expenses when investigating gene-biomarker relationships; however, improperly accounting for selection based on case/control status could bias the estimated gene-biomarker association. The magnitude of this bias may be sufficient to contribute to the lack of replication across genetic association case-control studies. We use simulation to compare the relative bias and efficiency of six standard methods for testing association between a biallelic genetic factor (G) and a continuous biomarker (X), including regressing X on G restricted to controls and inverse probability-of-sampling weighted (IPW) linear regression. All methods have appropriate Type I error rate when either X is independent of case-control status D or G is independent of D. The magnitude and direction of bias depend on the pattern of association among X, G and D, disease prevalence, and the analysis method. IPW is the only method that is unbiased in all disease scenarios. Method selection should rely on knowledge of the underlying disease model and the study purpose: gene characterization studies may favor IPW to avoid bias, whereas large genomic screens may necessitate the power afforded by an unadjusted approach.

email: gmonsees@hsph.harvard.edu

9f. ALLELIC BASED GENE-GENE INTERACTION ASSOCIATED WITH QUANTITATIVE TRAITS

Jeesun Jung*, Indiana University
Bin Sun, Indiana University
Deukwoo Kwon, National Cancer Institute
Daniel L. Koller, Indiana University
Tatiana M. Foroud, Indiana University

Recently, it has been widely recognized that gene-gene interaction is likely to have an important contribution to the etiology of disease genes. Most of the statistical methods to detect SNP by SNP interaction contributing to quantitative traits are ways to divide the multi-locus genotypes into subgroups tending to ignore specific allele information that may be associated. In this study, we propose a new statistical approach based on the nonrandom association of these alleles at multiple unlinked loci and designed to detect SNP by SNP interaction for those cases where the effect of each SNP can not be detected when analyzed as a main effect. Our proposed method assigns a score to allelic combination inferred by genotypes of subjects and tests association with quantitative traits at the allelic level. Based on a simulation study of type I error rates and power, we show that the allelic approach achieves greater power than the genotypic approach through data dimension reduction by a smaller number of allelic combinations rather than a larger number of genotypic combinations. We applied our method to a candidate genetic study of kidney sodium retention in a sample of African Americans. We found that this method identifies the ability to detect interaction effects between SNPs in the Calcium-sensing Receptor (CaSR) gene and the chloride channel (CLCNKB) gene.

email: jeejung@iupui.edu

9g. ASSOCIATION STUDIES OF CASE-CONTROL DATA WITH GENOTYPING UNCERTAINTY

Youfang Liu*, North Carolina State University
Jung-Ying Tzeng, North Carolina State University

Current genotyping technology produces two dimensional fluorescent intensity (FI) data. Genotypes are inferred from FI data by a scoring algorithm and association analysis is conducted between genotypes and phenotypes. Genotyping scoring errors remain a challenge for automated scoring programs and it renders a negative impact on association analysis. Here, we propose a new score test that incorporates the genotyping uncertainty to assess the association between traits and SNPs. In this method, we directly use the original FI data and regard genotypes as unobserved variables. Therefore genotyping scoring error would be no longer a problem. Extensive simulation studies for both binary and continuous traits demonstrate that our method outperforms other approaches using inferred genotypes.

email: yliu7@ncsu.edu

9h. RAPID INFERENCE OF HAPLOTYPES IN PEDIGREES

Xiaohua Gong*, North Carolina State University
Zhao-bang Zeng, North Carolina State University

In dairy cattle half-sib populations one sire is mated with many dams. Usually sires and their sons are genotyped but dams are not. Thus one must reconstruct sire haplotypes in order to map quantitative trait loci. Here we propose a likelihood-based algorithm to reconstruct sire haplotypes given the genotypes observed in the sire and sons. We demonstrate that our method outperforms other methods in terms of accuracy and computing time. Simulations indicate our methods performance with respect to the number of offspring, the number and density of markers, and the extent of missing values in marker genotype data. Our algorithm also applies to other types of outbred populations. We have implemented the algorithm into a computer program freely available.

email: xgong@ncsu.edu

9i. MUTATION DETECTION FOR VIRUS DATA

Hongyuan Cao*, University of North Carolina-Chapel Hill
Xingye Qiao, University of North Carolina-Chapel Hill

In the past few years there has been increased interest in using data-mining techniques to extract interesting patterns from time series data generated by sensors monitoring temporally varying phenomenon. In some cases the rule for determining when a sensor reading should generate an event is well known. However, it is difficult to state such a rule when phenomenon is ill-understood. Detection of events in such an environment is the focus of this poster. The problem we are interested in is to identify the time points at which the behavior change occurs, the so called change-point detection problem. Generally there are two approaches to address this problem: (a) fix the number of change-points that are to be discovered; (b) decide the function that will be used for curve fitting as well as the location of change points. In this paper, we propose an iterative algorithm that fits the model to a time segment, and use empirical RSS, F test, t test and BIC as our model selection criteria. Evaluation of these criteria includes function recognition, empirical 95% bounds, randomized mean performance and case sensitivity analysis. Finally, we apply this method to virus data to detect the mutation points at which plague sizes of virus change.

email: hycao@email.unc.edu

9j. AN EMPIRICAL BAYESIAN METHOD TO CORRECT FOR WINNER'S CURSE IN GENETIC ASSOCIATION STUDIES

Rui Xiao*, University of Michigan
Michael Boehnke, University of Michigan

Genetic association mapping is a powerful method to detect genetic variants that predispose to human disease. Investigators are also interested in estimating the genetic effect on disease risk of each identified variant. Initial positive findings of the genetic effect estimate tend to be upwardly biased particularly if they were the first to reach the statistical significance level, a phenomenon known as the winners curse. Overestimation of genetic effect size in initial studies may cause follow-up studies to be underpowered and so to fail. In this paper, we propose an empirical Bayesian method to correct for the overestimation. Our method incorporates information from genomewide association studies as a prior distribution for the genetic effect size throughout the genome, and then combines the locus specific data to reduce the bias. We compare our method with the existing resampling-based method and the likelihood-based method.

email: xiaor@umich.edu

9k. LINKAGE DISEQUILIBRIUM FILTER IN GENOME WIDE ASSOCIATION STUDIES

Nadia Timofeev*, Boston University
Paola Sebastiani, Boston University

Genome wide association studies examine hundreds of thousands of SNPs across the genome for associations with genetic factors affecting disease. These studies generate a large amount of data and have introduced many interesting challenges to statisticians. We will address the issue of multiple comparisons and show through simulations that a filter which accounts for the magnitude of linkage disequilibrium (LD) surrounding a significant SNP can reduce the false positive rate by almost half. SNPs in LD with each other are expected to show similar strengths of association if the association is in fact real. If a single significant SNP is not supported by the significance of other SNPs within its LD block, then the association is likely to be spurious. Our procedure checks for patterns of LD and significance and excludes statistically significant SNPs not supported by the SNPs in its LD block. We will also discuss implications of the power necessary to detect significance in neighboring SNPs.

email: ntimofee@bu.edu

10. POSTERS: GENOMICS AND PROTEOMICS

10a. BOOTSTRAP AGGREGATION FOR ORDINAL RESPONSE PREDICTION IN HIGH-THROUGHPUT GENOMIC DATASETS

Kellie J. Archer*, Virginia Commonwealth University

Ensemble methods have been demonstrated to be competitive with other machine learning approaches for classification, particularly in predicting phenotypic class when the number of covariates (p) greatly exceed the sample size (n), such as is the case in high-throughput genomic experiments. To date, ensemble methods have been described for nominal, continuous, and survival responses. However, in a large number of biomedical applications, the class to be predicted may be inherently ordinal. Examples of ordinal responses include TNM stage (I, II, III, IV) and drug toxicity (none, mild, moderate, severe). While nominal response methods may be applied to ordinal response data, in so doing some information is lost that may improve the predictive performance of the classifier. Herein we compare the effectiveness of combining an ordinal impurity function with bootstrap aggregation to traditional bagging algorithms which use the Gini impurity function. Results are presented for both simulated and genomic datasets.

email: kjarcher@vcu.edu

10b. MACHINE LEARNING ALGORITHMS FOR GENETIC ASSOCIATION STUDIES

Bareng Aletta S. Nonyane*, University of Massachusetts
Andrea S. Foulkes, University of Massachusetts

Population-based studies aimed at uncovering genotype-trait associations often involve high-dimensional genetic polymorphism data as well as information on multiple environmental and clinical parameters. Machine learning (ML) algorithms offer a straightforward analytic approach for selecting small subsets of these inputs that are most predictive of a pre-defined trait. The performance of these algorithms, however, in the presence of covariates is not well characterized. In this study, we investigate two approaches: Random Forests (RFs) and Multivariate Adaptive Regression Splines (MARS). Through multiple simulation studies, the performance under several underlying models is evaluated. An application to a cohort of HIV-1 infected individuals receiving anti-retroviral therapies is also provided. Consistent with more traditional regression modeling theory, our findings indicate that ignoring population-level confounding covariates in the application of ML algorithms to genomic data may result in high false positive rates.

email: aletta@schoolph.umass.edu

10c. A GLOBAL TEST PROCEDURE WITH APPLICATIONS TO GENE EXPRESSION PROFILING STUDIES

Chien-Ju Lin*, U.S. Food and Drug Administration
Ching-Wei Chang, U.S. Food and Drug Administration
James J. Chen, U.S. Food and Drug Administration

In order to understand biological functions of individual genes, gene-class testing (GCT) has been proposed for gene expression profile analysis. A gene class refers to a group of genes with related functions or a set of genes grouped together based on biologically relevant information. In GCT, the hypothesis is: if there are treatment effects in the gene class. Most statistical methods for GCT are the one-sided test, that is, the changes of individual gene expressions in a gene class are all in one direction: either up or down. The one-side test might not be realistic, since a significant gene class often consists of mixture of up- and down regulated genes. In this study, we apply the generalized linear model by assuming random effects of gene expressions to the two-sided GCT. The generalized linear model approach can be expended to take other clinical covariates into consideration. We combine the random effect linear model and LASSO to estimate the large effect while shrinking the smaller effects toward zero. The proposed procedure is applied to two data sets, a diabetes data set and a breast cancer data set, for illustration.

email: Chien-Ju.Lin@fda.hhs.gov

10d. AN APPROACH TO DETECT STATISTICAL INTERACTORS IN GENE NETWORKS

Alina Andrei*, University of Wisconsin-Madison
Christina M. Kendziorski, University of Wisconsin-Madison

Complex traits ranging from mRNA expression to disease phenotype are affected by multiple genes, and by interactions among these genes. Gene association networks (GANs) provide a simple ,yet effective,summary of relationships among genes and are often used as a foundation upon which more complex inference is carried out. In GANs, nodes represent genes and edges represent some measure of association between genes, such as correlation, or partial correlation for graphical Gaussian models (GGMs). Although useful, GANs are limited in that the association measures usually quantify linear relationships among nodes. We propose an efficient approach to accommodate interactions, extending the class of GGM's. This approach utilizes information from the partial correlation matrix, following several steps for which theoretical support is derived. Extensive simulations show that the approach is also applicable when the sample size is smaller than the number of genes considered, a situation common in high dimensional studies. A comparison between the resulting gene interaction network and traditional GGMs illustrates the advantages of the proposed approach in gene expression studies of breast cancer and diabetes.

email: aandrei@biostat.wisc.edu

10e. MULTIVARIATE MODELS TO DETECT GENOMIC SIGNATURES FOR A CLASS OF DRUGS

Brooke L. Fridley*, Mayo Clinic
Greg Jenkins, Mayo Clinic
Daniel Schaid, Mayo Clinic
Liewei Wang, Mayo Clinic
Fang Li, Mayo Clinic

Recently, there has been an increased interest in individualized medicine and thus pharmacogenetics and pharmacogenomics in cancer research. The NIH-funded Mayo PGRN (Pharmacogenomic Research Network) has pharmacogenomic studies involving many classes of cancer drugs on cell lines in which cytotoxicity measurements at multiple dose concentrations. Along with the cytotoxicity data, both expression and genotype data are measured. When studying multiple drugs from the same class of drugs that have similar genetic mechanisms, one may wish to determine genomic variation that explains the difference in cytotoxicity between individuals for the class of drugs as opposed to each individual drug. Thus, we have developed a multivariate model to assess if genomic variation impacts a class of drugs. We will illustrate the utility of this multivariate model for cytotoxicity and genomic data collected on the Coriell Human Variation Panel for the class of anti-purine metabolite that involved 6TG and 6MP that are used to treat childhood acute lymphoblastic leukemia (ALL), Inflammatory bowl disease (IBS) and transplant patients. The multivariate model will be compared to the univariate models for each individual drug along with comparison to the data reduction method of principal components.

email: fridley.brooke@mayo.edu

10f. SYSTEMATIC QUANTIFICATION OF THE EFFECTS OF POPULATION STRUCTURE ON GENOME-WIDE ASSOCIATION STUDIES

Hongyan Xu*, Medical College of Georgia
Varghese George, Medical College of Georgia

Large-scale genome-wide association studies are promising for unraveling the genetic basis of complex diseases. Population structure is a potential problem, the effects of which on genetic association studies are controversial. Systematic quantification of the effects of population structure on large scale genetic association studies is needed for valid analysis of the data and correct interpretation of the results. In this study, we performed extensive coalescent-based simulations of samples with varying levels of population structure to investigate the effects of population structure on large-scale genetic association studies. The effects of population structure are measured by the multiplicative changes of the probability of type I error rate, which is then correlated with the levels of population structure. It is found that at each nominal level of association tests, there is a positive relationship between the level of population structure and its effects, which could be summarized well with a regression function. It is also found that at a specific level of population structure, its effect on association study increases drastically as the significance level of the test decreases. Therefore in genome-wide association studies, the effects of population structure cannot be safely ignored and must be accounted for with proper methods.

email: hxu@mcg.edu

10g. ESTIMATION OF GENETIC VARIANCE EXPLAINED FOR INDIVIDUAL QUANTITATIVE TRAIT LOCI

William Bridges*, Clemson University
Steve Knapp, University of Georgia

In most studies involving the detection and location of individual Quantitative Trait Locus (QTL), the genetic variance associated with the QTL is estimated. The total genetic variance is also estimated and the ratio of these two variance estimates is often used as an estimate of the contribution of an individual QTL to total genetic variance (i.e., a heritability-like estimate for the QTL). Using some linear model and variance component estimation results we show that variance estimates associated with individual QTL are not a simple additive function of the total genetic variance estimate, and that the contribution of individual QTL to total genetic variance is often an over-estimate. An example of the variance estimates for an actual data set is shown and a correction for the over-estimation is derived.

email: wbrdgs@clemson.edu

11. POSTERS: MICROARRAY ANALYSIS

11a. MODIFIED LINEAR DISCRIMINANT ANALYSIS APPROACHES FOR CLASSIFICATION OF HIGH-DIMENSIONAL MICROARRAY DATA

Ping Xu*, University of Louisville
Guy N. Brock, University of Louisville
Rudolph S. Parrish, University of Louisville

Motivation: Linear discriminant analysis (LDA) is one of the most popular methods of classification. In high-dimensional microarray data classification problems, due to the small number of samples and large number of features (genes), a problem with classical LDA is the singularity and instability of the within-group covariance matrix. In general, the performance of classical LDA is sub-optimal in high-dimensional situations. Methods: In this study, we applied two modified LDA approaches (MLDA and NLDA) to microarray classification and compared the performance of the two modified LDA approaches with other popular classification algorithms across a range of feature set sizes (number of genes) using both simulated and real benchmark cancer datasets. Results: We found that when the feature set size is close to or greater than the training set size, classical LDA often has poor performance. However, the overall performance of the two modified LDA approaches is as competitive as support vector machines and better than diagonal linear discriminant analysis (DLDA), k-nearest neighbor, and classical LDA. We recommend the use of the modified LDA approaches in limited sample size and high-dimensional microarray classification problems in consideration of computational simplicity and overall performance.

email: xupingky@gmail.com

11b. EXPLORING THE BREAST CANCER GENE EXPRESSION DATA: IDENTIFICATION AND ANALYSIS OF FUNCTIONAL ERALPHA REGULATORY NETWORK IN SILICO

Li-yu D. Liu*, National Taiwan University
Mei-Ju M. Chen, National Taiwan University
Ming-Shian Tsai, National Taiwan University
Cho-Han S. Lee, National Taiwan University
Chien-yu Chen, National Taiwan University
Tzu L. Phang, University of Colorado Health Sciences Center
Li-Yun Chang, National Taiwan University
Wen-hung Kuo, National Taiwan University
Hsiao-Lin Hwa, National Taiwan University
Huang-Chun Lien, National Taiwan University

A study of lung cancer risk from residential radon exposure and its radioactive progeny was performed with 200 cases (58% male, 42% female) and 397 controls matched in age and sex, all from the same health maintenance organization. Emphasis was placed on accurate and extensive, year-long dosimetry with etch-track detectors in conjunction with careful questioning about historic patterns of in-home mobility. Conditional logistic regression was used to model the outcome of cancer on radon exposure, while controlling for years of residency, smoking, education, income, and years of job exposure to known or potential carcinogens. Radon exposure was divided into six categories with break points at 25, 50, 75, 150, and 250 Bq m-3, the lowest being the reference. The adjusted odds ratios (AOR) were, in order, 1.00, 0.53, 0.31, 0.47, 0.22, and 2.50 with the third category significantly below 1.0 ($p<0.05$), and the second, fourth, and fifth categories approaching statistical significance ($p<0.1$). An alternate regression analysis using natural cubic splines allowed calculating AORs as a continuous function of radon exposure. That analysis produced AORs that are substantially less than 1.0 with borderline statistical significance (0.048 d p d 0.05) between approximately 85 and 123 Bq m-3.

email: lyliu@ntu.edu.tw

11c. ADJUSTING FOR COVARIATES IN THE ANALYSIS OF MICROARRAY AND OTHER HIGH DIMENSIONAL BIOLOGICAL DATA

Lang Chen*, University of Alabama at Birmingham
T. Mark Beasley, University of Alabama at Birmingham
Christopher S. Coffey, University of Alabama at Birmingham
Grier P. Page, University of Alabama at Birmingham

Microarrays enable investigators to simultaneously measure the expression of thousands of genes. Covariates are expected to have an effect on the expression of only some of the genes on a microarray. Thus, adjusting all genes for a covariate will increase the accuracy of variance estimation for some genes, but will introduce extra variability to other genes. The purpose of this study is to asses when to adjust genes in a microarray study for covariates and when not to do so. We used a simulation approach to asses the impact of adjusting for covariates on the results of microarray and other high dimensional biological experiments. We found that never adjusting for existing covariates (ANOVA) reduced power. Similarly, always adjusting for a covariate (ANCOVA) also reduced power. We tested four approaches (AIC,BIC, MSE, and p-value) to determine when to adjust a gene for a covariate. We found these methods to be similar in power and type I error rate. Use of these modeling approaches is superior to the use of ANOVA or ANCOVA alone. When covariates are measured in the context of a microarray study, they should be used to adjust gene expression values in order to generate more accurate inferences. But Adjustments should be done to each individual gene expression value, and we have identified several methods for determining whether or not to adjust a gene expression profile for a covariate that improve power while controlling the type I error rate.

email: lchen@ms.soph.uab.edu

11d. A CENSORED BETA MODEL FOR ESTIMATING THE PROPORTION OF TRUE NULL HYPOTHESES IN A MICROARRAY EXPERIMENT

Anastasios Markitsis*, The George Washington University
Yinglei Lai, The George Washington University

Microarray technology uses the hybridization property of DNA to measure the expression of tens of thousands of genes simultaneously. In a two-sample microarray experiment, gene expression data from two types of cells are collected. To detect differentially expressed genes, a certain test is used to screen all the genes in the study. Since tens of thousands of genes are tested simultaneously, it is necessary to adjust the p-values. A widely used method is controlling the False Discovery Rate (FDR), the proportion of false positives among the claimed positives. To obtain an accurate estimate of the FDR, it is necessary to estimate the proportion of true null hypotheses, \pi_0 (i.e., the proportion of nondifferentially expressed genes). Although many methods have been proposed, there is still a need for efficient estimation methods. In this study, we propose a censored beta model as the underlying distribution of p-values for a conservative estimation of \pi_0. We compare the proposed method with six other methods (convest, qvalue, HIST, BUM, SPLOSH, and LBE); in two simulation studies, the proposed method performs competitively, and shows satisfactory performance. In an application to an experimental data set, the censored beta model provides a comparatively stable estimate for \pi_0.

email: amarkits@gwu.edu

11e. HIERARCHICAL MIXTURE METHOD WITH DATA-ADAPTIVE DIMENSION REDUCTION FOR ASSESSING DIFFERENTIAL EXPRESSION IN TWO-SAMPLE MICROARRAY EXPERIMENTS WITH SMALL SAMPLE SIZES

Sang-Hoon Cho*, University of Wisconsin - Madison

In two-sample microarray experiments, it is natural to expect that there exists a strong positive linear association between intensity measurements, especially under the null hypothesis of no differentially expressed genes. In this talk, we will illustrate with two examples how much improvement we can achieve once the overall dependent structure is considered in statistical approaches. In addition, we will introduce a hierarchical mixture model with data-adaptive dimension reduction when the number of sample sizes is small. By a simulation study and data analysis, the proposed method will be compared to several methodologies in terms of sensitivity and specificity.

email: cho@stat.wisc.edu

11f. REALISTIC SIMULATION OF AFFYMETRIX GENE EXPRESSION ARRAYS

Andrew C. Hardin*, Southern Methodist University

AffyMetrix Gene Expression Microarrays are used to detect difference in gene expression between different treatment conditions. Analysis methods perform preprocessing and significance tests to produce lists of differentially expressed genes. Simulated microarray experiments can be used to compare the effectiveness of different analysis methods. Though models are available for comparing cDNA microarrays these models are not easily extended to AffyMetrix arrays. A statistical model is proposed to address issues specific to AffyMetrix gene expression arrays.

email: ahardin@smu.edu

11g. BAYESIAN ANALYSIS OF MICROARRAY EXPERIMENTS WITH MULTIPLE SOURCES OF VARIATION

Cumhur Y. Demirkale*, Iowa State University
Dan Nettleton, Iowa State University
Tapabrata Maiti, Iowa State University

Some microarray experiments have complex experimental designs that call for modeling of multiple sources of variation through the inclusion of multiple random factors. While data on thousands of genes are collected in these experiments, the sample size for each gene is usually small. Therefore, in a classical gene-by-gene mixed linear model analysis, there will be very few degrees of freedom to estimate the variance components of all random factors considered in the model and low statistical power for testing fixed effect(s) of interest. To address these challenges, we propose a hierarchical Bayesian modeling strategy to account for important experimental factors and complex correlation structure among the expression measurements for each gene. We use half-Cauchy priors for the standard deviation parameters of the random factors with few effects. We rank genes with respect to evidence of differential expression across the levels of a factor of interest by calculating a single summary statistic per gene from the posterior distribution of the fixed treatment effects considered in the model. Simulation shows that our hierarchical Bayesian approach is much better than a traditional gene-by-gene mixed linear model analysis at distinguishing differentially expressed genes from non-differentially expressed genes.

email: cyusuf@iastate.edu

11h. A FLEXIBLE MODEL SELECTION ALGORITHM FOR THE COX MODEL WITH HIGH-DIMENSIONAL DATA

Alexander T. Pearson*, University of Rochester School of Medicine and Dentistry
Derick R. Peterson, University of Rochester School of Medicine and Dentistry

Gene array information can be collected and related to the time to a specific disease-related event. A prognostic gene signature, or risk score, should be a well-defined function of the expression of a group of genes that are jointly predictive of survival. Such signatures could lead to new avenues of cancer treatment or prevention. We consider the problem of selecting which among a large number of genes should be included in a prognostic gene signature, using the Cox proportional hazards model framework. Standard model selection methods perform poorly in this high-dimensional context. We propose a search algorithm that lies between forward selection and searching over all subsets. This method uses an evolving subgroup paradigm to heuristically search over a large model space. We show that our method can yield significant improvements in the number of predictive variables selected and the prediction error, compared with forward selection and univariate screening, as demonstrated via simulation studies. We apply our new method to publicly available lymphoma gene array data (Rosenwald et. al., 2002) with 7399 genes and 240 patients. Training our method on a subsample of the data leads to predictive results in the remaining validation data.

email: alexander_pearson@urmc.rochester.edu

12. STATISTICAL MONITORING OF CLINICAL TRIALS: WHERE THE RUBBER MEETS THE ROAD

DISOBEYING RULES: UNPLANNED EARLY STOPPING OF A NON-INFERIORITY TRIAL

Janet T. Wittes*, Statistics Collaborative

The ACTIVE-W study was a noninferiority trial compaing two approaches to anticoagulation in patients with atrial fibrillation. The standard treatment, oral anticoagulation with warfarin, can cause major bleeding. The experimental method, aspirin plus clopidogrel, was medically appealing because it was easier to administer and was not expected to cause bleeding. ACTIVE-W was designed to continue with 1450 participants experienced at least one major cardiovascular outcome. The DSMB carefully established a monitoring plan to ensure safety of the participants; the plan called for a first look at the data when 25 percent of the outcomes had occurred. The DSMB was forced to make decisions on the fly. This paper will deal with methods for monitoring noninferiority trials and how a DSMB should deal with composite outcomes when the components of that endpoint behave differently from each other.

email: janet@statcollab.com

A CONSTRAINED BOUNDARIES APPROACH FOR FLEXIBLY MONITORING TIME TO EVENT OUTCOMES

Sean Brummel, University of California-Irvine
Daniel L. Gillen*, University of California-Irvine

It is well known that the use of group sequential stopping rules materially affects the frequentist operating characteristics of hypothesis tests. As such it is necessary to choose an appropriate stopping rule during the planning of a study. However, due to logistical constraints it is often the case that the number and timing of analyses are not precisely known at the time of trial design, and the implementation of a particular stopping rule must allow for flexible determination of the schedule of interim analyses. In this talk we consider the use of a constrained boundaries approach for implementing stopping rules in the setting of a survival endpoint. This approach requires estimation of the proportion of maximal statistical information at the time of an interim analysis. Here we consider the estimation of statistical information in the setting of potentially time-varying treatment effect on survival where a weighted logrank statistic might be considered. A simulation study will be presented to demonstrate the operating characteristics of the proposed method together with a case study to illustrate the procedure.

email: dgillen@uci.edu

SPENDING FUNCTIONS AND THE UPDATING OF INFORMATION FRACTIONS

Michael A. Proschan*, National Institute of Allergy and Infectious Diseases

Spending functions offer great flexibility by allowing arbitrary timing and number of interim looks at the data of a clinical trial. Still, to determine the current boundary in a trial with a survival outcome, one must estimate the current information fraction, defined as the ratio e/E of the current number of patients with events, to the number expected by trials end, E. This talk explores different ways to deal with an inaccurate initial estimate of E.

email: ProschaM@mail.nih.gov

A CONDITIONAL POWER APPROACH TO THE EVALUATION OF PREDICTIVE POWER

Kuang-Kuo G. Lan*, Johnson & Johnson
Peter Hu, Johnson & Johnson
Michael A. Proschan, National Institute of Allergy and Infectious Diseases, National Institutes of Health

We consider the use of a standard Brownian motion process on the unit interval [0,1] to predict the final outcome of a medical study during interim data analyses. The drift parameter theta is zero under the null hypothesis, and is positive under alternative hypotheses. Many of the test statistics in two-sample comparisons can be converted to the Brownian motion with a linear drift. For a fixed amount of information, the drift theta is determined by the standardized treatment difference D. For a given drift theta, the conditional power (CP) of a positive study during interim analysis is easy to visualize and simple to evaluate. A more sophisticated way to predict the future outcome is the use of the predictive power (PP) which considers the drift parameter as random and takes a weighted average of the CPs. Unfortunately, the evaluation of PP involves the integration of CPs, and many clinicians found that hard to follow. We propose the use of CP under a modified current trend as an alternative method to evaluate PP for the prediction of clinical trial outcomes. We will also discuss CP and PP for sequential designs and the choice of prior and posterior distributions under Bayesian settings.

email: glan@prdus.jnj.com

13. STATISTICAL ISSUES IN GENOMIC STUDIES IN POPULATION SCIENCES

GENE-ENVIRONMENT INTERACTION STUDIES

Raymond J. Carroll*, Texas A&M University
Nilanjan Chatterjee, National Cancer Institute
Yi-Hau Chen, Academia Sinica
Bhramar Mukherjee, University of Michigan

In gene-environment case-control studies, it is well known that making assumptions about the distribution of covariates in the population can lead to major gain in efficiency for estimating the effects of genes and gene-environment interactions. I will review this work, and discuss methods for building robustness against model assumptions into this framework.

email: carroll@stat.tamu.edu

POWERFUL MULTILOCUS ASSOCIATION TESTING OF COMPLEX TRAITS

Michael P. Epstein*, Emory University

Association mapping of complex traits typically employs tagSNP genotype data to identify functional variation within a region or genome of interest. However, considerable debate exists regarding the most powerful strategy for utilizing such tagSNP data for inference. A popular approach tests each tagSNP within the region individually, but such tests could lose power due to incomplete linkage disequilibrium between the genotyped tagSNP and the functional variant. Alternatively, one can jointly test groups of tagSNPs simultaneously (using multivariate genotypes or haplotypes), but such multivariate tests have large degrees of freedom that can also compromise power. Here, we consider a semiparametric model for complex-trait mapping that uses genetic information from multiple tagSNPs simultaneously in analysis but produces test statistics with reduced degrees of freedom compared to existing multivariate approaches. We fit this model using a dimension-reducing technique called least-squares kernel machines, which we show is identical to analysis using a specific linear mixed model (which we can fit using standard software packages like SAS and R). Using simulated SNP data based on real data from the International HapMap Project, we demonstrate our approach often has superior performance for association mapping of complex traits compared to existing approaches. Our approach is also flexible, as it allows easy modeling of covariates and high-dimensional interactions among tagSNPs and environmental predictors.

email: mepstein@genetics.emory.edu

INCORPORATING PRIOR BIOLOGICAL KNOWLEDGE IN GENOME-WIDE ASSOCIATION STUDIES

Hongyu Zhao*, Yale University School of Medicine

The last three years have seen great successes in many Genome-Wide Association Studies (GWAS) which have identified numerous genetic variants underlying complex traits. The analysis and interpretation of data from GWAS presents great statistical and computational challenges, especially after the initial discoveries of variants carrying relatively large effects. Although various statistical approaches have been or are being developed to better analyze GWAS data, it has become apparent that the incorporation of information from prior studies and other sources is indispensable. In this presentation, we discuss our recently developed statistical methods and bioinformatics tools that are designed to more effectively integrate diverse types of prior biological information in analyzing GWAS data. The usefulness of these methods will be illustrated through their applications to some recent large scale GWAS data. This is joint work with Iryna Lobach, David Ballard, Ji Young Lee, and Judy Cho.

email: hongyu.zhao@yale.edu

14. HIERARCHICAL MODELING IN ENVIRONMENTAL EXPOSURE AND TOXICOLOGICAL RISK ASSESSMENT

BAYESIAN HIERARCHICAL REGRESSION FOR MODEL SELECTION AND CLUSTERING

Amy H. Herring*, The University of North Carolina at Chapel Hill
David B. Dunson, National Institute of Environmental Health Sciences
Stephanie M. Engel, Mount Sinai School of Medicine

In epidemiologic studies, there is often interest in assessing the relationship between highly-correlated predictors and a health outcome. Examples include studies of polymorphisms in functionally-related genes or exposures to mixtures of chemicals. Because instabilities can result in analyses that include all predictors, dimensionality is typically reduced by conducting single predictor analyses. This article proposes an alternative Bayesian approach for reducing dimensionality. A multi-level Dirichlet process prior is used for the distribution of the predictor-specific regression coefficients, incorporating a variable selection-type mixture structure in the base measure to allow predictors with no effect. This structure allows simultaneous selection of important predictors and clustering of predictors having similar impact on the health outcome.

email: amy_herring@unc.edu

REGIONAL SPATIAL MODELING OF ARSENIC IN ENVIRONMENTAL MEDIA: IMPLICATIONS FOR HUMAN EXPOSURE ASSESSMENT

Catherine A. Calder*, The Ohio State University
Peter F. Craigmile, The Ohio State University
Hongfei Li, IBM Corporation
Rajib Paul, The Ohio State University
Jian Zhang, Freddie Mac

Characterizing variation in human exposure to toxic substances over large populations often requires an understanding of the geographic variation in environmental levels of toxicants. This knowledge is essential when the primary routes of exposure are through interaction with environmental media, as opposed to more individual-specific exposure routes (e.g., occupational exposure). In this study, we focus on modeling the spatial variation in the concentration of arsenic, a toxic heavy metal, in air, soil, and water across the state of Arizona. We then synthesize this background information with individual-specific arsenic exposure measurements from EPAs National Human Exposure Assessment Survey (NHEXAS) in a Bayesian hierarchical pathways model. We discuss the implications of including this background exposure information in assessing the relative importance of various exposure pathways and on the general issue of assessing model fit in large multilevel statistical models.

email: calder@stat.osu.edu

USING HIERARCHICAL PK/PD MODELS FOR PROBABILISTIC DOSE-RESPONSE ASSESSMENT

Ralph L. Kodell*, University of Arkansas for Medical Sciences

Probabilistic risk assessment is being used increasingly to characterize and communicate uncertainties in estimates of benchmark doses (BMDs) as reference levels of exposure to toxic substances. This presentation describes the use of hierarchical statistical models as tools for implementing probabilistic dose-response assessments, in that such models provide a natural connection between the pharmacokinetic (PK) and pharmacodynamic (PD) components of dose-response models. For a given benchmark response (BMR), a complete distribution of BMD100BMR can be simulated by Monte Carlo bootstrap resampling of the observed response data. Results show that incorporating internal dose information into dose-response assessments via coupling of PK and PD models in a hierarchical structure can reduce the uncertainty in the dose-response assessment. However, information on the mean of the internal dose distribution is sufficient; having information on the variance of a PK model of internal dose does not appreciably affect the uncertainty in the resulting estimates of benchmark doses. It is also shown that the hierarchical structure is a natural vehicle for incorporating information on upstream precursor effects (e.g., biomarkers of effect) into the dose-response assessment for downstream frank toxic endpoints.

email: rlkodell@uams.edu

15. MODEL VALIDATION AND MODEL SELECTION IN LONGITUDINAL DATA ANALYSIS

QUADRATIC INFERENCE FUNCTIONS IN MARGINAL MODELS FOR LONGITUDINAL DATA

Peter X K Song*, University of Waterloo

Quadratic Inference Function (QIF) is getting increasingly popular in the analysis of longitudinal data, because this method has been proved at many occasions, analytically and numerically, to perform better than Liang and Zeger's GEE. This presentation will give a brief introduction to the QIF, and then discuss some pros and cons of this method in comparison to the GEE, including efficiency, goodness-of-fit and robustness against data contamination. Also, some recent developments on model selections based on the QIF approach will be presented. All numerical illustrations given in the talk are produced via a SAS MACRO QIF, a free software package available for users to download from webpage: www.stats.uwaterloo.ca/~song.

email: song@uwaterloo.ca

REDUCING THE BIAS OF BETWEEN - / WITHIN- CLUSTER COVARIATE METHODS WHEN DATA ARE MISSING AT RANDOM

John M. Neuhaus*, University of California-San Francisco
Charles E. McCulloch, University of Chicago

Generalized linear mixed models that partition covariates into between- and within-cluster components can provide effective analysis of longitudinal data in settings where covariates or responses are missing completely at random. However, like conditional likelihood methods, such between-/within-cluster approaches can yield inconsistent covariate effect estimates when data are missing at random. This talk describes and evaluates several strategies, including weighted methods, to reduce bias when data are missing at random. We illustrate these methods with simulation studies and fits to example data.

email: john@biostat.ucsf.edu

REGULARIZATION PARAMETER SELECTION FOR PENALIZED LIKELIHOOD VARIABLE SELECTION PROCEDURES

Runze Li*, Penn State University

Penalized likelihood approach with continuous penalty has been proposed to select significant variables for various statistical models in the recent literature. Its regularization parameter controls the model complexity, and therefore plays a critical role in its implementation. In this paper, we systematically study the issue of regularization parameter selection. We proposed a new regularization parameter selector, and further investigate the asymptotic behaviors of the newly proposed regularization parameter selector for generalized linear models and Cox's models. The ideas are applicable for other statistical models. We demonstrate that under certain conditions, the penalized likelihood procedures with some regularization parameter selectors may yield an overfit model, while with carefully chosen regularization parameter, the resulting penalized likelihood estimate possesses the oracle property in the terminology of Fan and Li (2001). Monte Carlo simulation study is conducted to examine the finite sample performance of the proposed regularization parameter selectors. The simulation results confirms the theoretic findings. A real data example is used to illustrate the proposed procedures.

email: ril4@psu.edu

A BIC CRITERION FOR LONGITUDINAL DATA MODEL SELECTION

Lan Wang*, University of Minnesota
Annie Qu, Oregon State University

Model selection for marginal regression analysis of longitudinal data is challenging due to the presence of correlation and the difficulty of specifying the full likelihood, particularly for correlated categorical data. This paper introduces a novel BIC-type model selection criterion based on the quadratic inference function (Qu, Lindsay and Li, 2000), which does not require the full likelihood or quasilikelihood. With probability approaching one, the criterion selects the most parsimonious correct model. Although a working correlation matrix is assumed, there is no need to estimate the nuisance parameters in the working correlation matrix; moreover, the model selection procedure is robust against the misspecification of the working correlation matrix. The BIC-type criterion can also be used to construct a data-driven Neyman smooth test for checking the goodness-of-fit of a postulated model. This test is especially useful and often yields much higher power in situations where the classical directional test behaves poorly. The finite sample performance of the model selection and model checking procedures is demonstrated through Monte Carlo studies and analysis of a clinical trial data set.

email: lan@stat.umn.edu

16. NEW STATISTICAL METHODS IN DIAGNOSTIC MEDICINE

BOOTSTRAP AND EMPIRICAL LIKELIHOOD-BASED NONPARAMETRIC INFERENCE FOR THE DIFFERENCE BETWEEN TWO PARTIAL AUCS

Gengsheng Qin*, Georgia State University
Yan Yuan, Georgia State University
Xiao-Hua Zhou, University of Washington

Comparing the accuracy of two continuous-scale diagnostic tests is increasingly important when a new test is developed. Traditional way of comparing entire areas under two Receiver Operating Characteristic (ROC) curves is not sensitive when two ROC curves cross each other. Comparing the areas under two ROC curves over a specific interval of false positive rates is a more appropriate way to compare the accuracy of two diagnostic tests. In this paper, we have proposed bootstrap and empirical likelihood (EL) approach for inference of the difference between two partial areas under the ROC curves (partial AUCs). The empirical likelihood ratio for the difference between two partial AUCs is defined and its limiting distribution is a scaled chi-square distribution. Using this scaled chi-square distribution, the EL-based confidence intervals for the difference between two partial AUCs are obtained. Additionally we conduct a simulation study to compare four proposed EL and bootstrap based intervals.

email: gqin@gsu.edu

ESTIMATING DIAGNOSTIC ACCURACY FROM DESIGNS WITH NO GOLD STANDARD, PARTIAL GOLD STANDARD, OR IMPERFECT GOLD STANDARD EVALUATION

Paul S. Albert*, National Cancer Institute

Interest often focuses on estimating sensitivity and specificity of a group of raters or a set of new diagnostic tests in a situation where gold standard evaluation is invasive or expensive. For situations in which no gold standard evaluation is available, various authors have proposed latent class models for estimating diagnostic accuracy. We show that these approaches lack robustness to key modeling assumptions and, in most practical situations, it is difficult to check these assumptions. We discuss two alternatives to using latent class models. First, we propose methodology (semilatent class models and imputation approaches) for estimating diagnostic accuracy when the gold standard test is available on a subset of individuals. Second, we propose an approach for estimating diagnostic accuracy using an imperfect reference standard when the sensititivity and specificity of the imperfect test relative to the gold standard test is available from other studies. Through data analysis, analytic results, and simulations, we demonstrate that these two approaches have improved statistical properties relative to latent class models without a gold standard.

email: albertp@mail.nih.gov

SEMIPARAMETRIC LEAST SQUARES ROC ANALYSIS OF CORRELATED BIOMARKER DATA

Liansheng Tang*, George Mason University
Xiao-Hua Zhou, University of Washington

The receiver operating characteristic (ROC) curve is a popular tool to evaluate the accuracy of continuous-scale biomarkers. In this article we propose a semiparametric least squares method to estimate ROC curves from correlated biomarker data. Our new method has several advantages over existing ROC methods. First, unlike most existing methods our method does not require iterations and is simple to implement. Second, our method allows unknown baseline functions. Third, our method includes interaction terms between discrete covariates and false positive rates. Such interaction terms are important in the ROC modeling because without them the ROC curves of biomarkers are not allowed to intersect. In addition, based on our semiparametric method we propose a separation curve method to identify the range of false positive rates for which two ROC curves differ or one ROC curve is superior than the other. We compared the finite sample performance of the newly proposed semiparametric method with a parametric least squares method and a semiparametric method in large-scale simulation studies. Finally, our method is illustrated through two real life examples.

email: ltang1@gmu.edu

17. REGULARIZATION AND HIERARCHICAL MODELING: TWO PHILOSOPHIES FOR VARIABLE SELECTION

MODEL SELECTION AND DIAGNOSTICS FOR GENETIC MARKERS

Dipak K. Dey*, University of Connecticut
Feng Guo, Virginia Tech University
Kent E. Holsinger, University of Connecticut

The distribution of genetic variation among populations is conveniently measured by Wright's F_{ST}. For single-nucleotide polymorphisms (SNPs), locus-specific estimates of F_{ST} will depart from a common mean only for loci with unusually high or low rates of mutation and for loci that are closely associated with genomic regions having a substantial effect on fitness. Thus, loci showing significantly more variation than background are likely to mark genomic regions subject to diversifying selection among the sample populations, while those showing significantly less variation than background are likely to mark genomic regions subject to stabilizing selection across the sample populations. We propose several Bayesian hierarchical models to estimate locus-specific effects on F_{ST}, and we apply these models to single nucleotide polymorphism data from the HapMap project. Because loci that are physically associated with one another are likely to show similar patterns of variation, we introduce conditional autoregressive models to incorporate the local correlation among loci. We estimate the posterior distributions of the model parameters using Markov chain Monte Carlo (MCMC) simulations.

email: dipak.dey@uconn.edu

DETECTING DIFFERENTIALLY EXPRESSED GENES VIA MULTILEVEL NONLINEAR MIXTURE DIRICHLET MODELS FOR HIGH-DIMENSIONAL EST DATA

Fang Yu, University of Nebraska Medical Center
Ming-Hui Chen*, University of Connecticut
Lynn Kuo, University of Connecticut
Peng Huang, Medical University of South Carolina
Wanling Yang, The University of Hong Kong

ESTs (Expressed Sequence Tags) are usually a one-pass sequencing reading of cloned cDNAs derived from a certain tissue. The frequency of unique tags among different unbiased cDNA libraries is used to infer the relative expression level of each tag. We develop a multinomial model with novel priors of nonlinear Dirichlet distributions for high-dimensional EST data with multiple libraries and/or multiple types of tissues. The properties of the priors and the implied posteriors are examined in detail. Gene selection algorithms are developed to detect differentially expressed genes between different types of tissues. The performance of the proposed gene selection algorithm is examined via several simulations. A real EST dataset is used to further illustrate the proposed methodology.

email: mhchen@stat.uconn.edu

VARIABLE SELECTION VIA ADAPTIVE ELASTIC NET AND ITS CONSISTENCY PROPERTY

Samiran Ghosh*, Indiana University, Purdue University-Indianapolis

Lasso proved to be an extremely successful technique for simultaneous estimation and variable selection. However lasso has two major drawbacks. First, it does not capture any grouping effect and secondly in some situations lasso solutions are inconsistent. To overcome inconsistency recently adaptive lasso was proposed where adaptive weights are used for penalizing different coefficients. Adaptive lasso enjoys oracle properties. Also recently a doubly regularized technique namely elastic net was proposed which encourages grouping effect i.e. either selection or omission of the correlated variable together and is particularly useful when the number of covariates (p) is much larger than the number of observations (n). However even for usual p < n case it does not deemed to be an oracle procedure. In this paper we propose a new version of the elastic net called adaptive elastic net which inherits some of the desirable properties of the adaptive lasso and elastic net. We explicitly prove its oracle properties for p < n case. An efficient algorithm was proposed in the line of LARS-EN which is then illustrated with simulated as well as real life data examples.

email: samiran@math.iupui.edu

18. METHODS FOR VARIABLE SELECTION AND MODEL BUILDING

BAYESIAN VARIABLE SELECTION UNDER HEREDITY CONSTRAINTS

Woncheol Jang*, University of Georgia
Johan Lim, Yonsei University

We propose a variable selection method for statistical models with high order interactions.The main challenge in this variable selection is to incorporate the effect heredity (Chipman, Hamda and Wu, 1997); higher order interaction can exist only if at least one of its parent effects exists. Using modern variable selection procedures such as LASSO may result in the violation of the effect heredity principle. We present a relatively simple Bayesian hierarchical variable selection procedure while still achieving the effect heredity principle. Examples in biomedical and engineering experiments are presented.

email: jang@uga.edu

A CP STATISTIC FOR FIXED EFFECTS VARIABLE SELECTION IN THE LINEAR MIXED MODEL FOR LONGITUDINAL DATA

Anita A. Abraham*, University of North Carolina at Chapel Hill
Lloyd J. Edwards, University of North Carolina at Chapel Hill

The Cp statistic has been of great use in the univariate linear model when looking at a pool of models which are each separately nested within a single full model. However, as of yet, no analog has been developed for the linear mixed model. In this paper, a Cp statistic is proposed for fixed effects variable selection in the linear mixed model for longitudinal data. Two example studies are used to demonstrate the performance of the proposed Cp statistic: a well-known, small, complete and balanced orthodontic study, and a larger study of blood pressure measurements containing missing data from the North Carolina Established Populations for the Epidemiologic Studies of the Elderly (EPESE).

email: aabraham@bios.unc.edu

A BAYESIAN APPROACH TO EFFECT ESTIMATION ACCOUNTING FOR ADJUSTMENT UNCERTAINTY

Chi Wang*, Johns Hopkins University
Giovanni Parmigiani, Johns Hopkins University
Ciprian Crainiceanu, Johns Hopkins University
Francesca Dominici, Johns Hopkins University

In this paper, we propose a novel Bayesian formulation, called `Bayesian Confounding Adjustment' (BCA) to account for adjustment uncertainty in effect estimation from a Bayesian perspective. BCA uses a different weighting mechanism than BMA (Raftery et al., 1997; Hoeting et al., 1999), wherein effect estimation is obtained by weighting effect estimates from models, all of which are fully adjusted for confounding.

In simulation studies we show that BCA provides estimates of the exposure effect that have lower mean squared error than BMA ad and correct coverage. We then compare BCA, the approach of Crainiceanu et al. (2008), and traditional BMA in a time series data set of hospital admissions, air pollution levels and weather variables in Nassau, NY for the period 1999-2005. Using each approach, we estimated the short-term effects of PM2.5 on emergency admissions for cardiovascular diseases, accounting for confounding. This application illustrates the potentially significant pitfalls of misusing variable selection methods in the context of adjustment uncertainty.

email: chwang@jhsph.edu

NONPARAMETRIC BAYES CONDITIONAL DISTRIBUTION MODELING WITH VARIABLE SELECTION

Yeonseung Chung*, University of North Carolina
David B. Dunson, National Institute of Environmental Health Sciences

This research considers methodology for flexibly characterizing the relationship between a response and multiple predictors. Goals are (1) to estimate the conditional response distribution addressing the distributional changes across the predictor space, and (2) to identify important predictors for the response distribution change both with local regions and globally. We propose a nonparametric Bayesian method that achieves the two goals simultaneously. This is accomplished through a product of two generalized product partition models (GPPMs) with stochastic search variable selection (SSVS) structure incorporated. The first GPPM characterizes the conditional response distribution as an infinite mixture model with the mixing weight varying across the important predictors, while the second GPPM flexibly models the distribution for non-important predictors independently from the first GPPM. The SSVS structure allows for the global variable selection and the predictor-dependent mixture structure for the conditional distribution obtained through first GPPM facilitates the local variable selection. An efficient stochastic search Gibbs sampling algorithm is proposed for posterior computation. The methods are illustrated through simulation and applied to an epidemiologic study.

email: chungy@email.unc.edu

THE ADAPTIVE DANTZIG SELECTOR FOR VARIABLE SELECTION

Lee Dicker*, Harvard University
Xihong Lin, Harvard University

When working with genomic or proteomic data, one frequently desires to select a subset of important predictors from a potentially very large number of candidates. We propose the adaptive Dantzig selector, a procedure which simultaneously selects variables and estimates parameters by approximately solving the normal score equations. Ridge regression estimates determine how close we must come to solving the score equations; a larger ridge regression estimate requires that we come closer solving the corresponding score equation. Parameter estimates are set equal to zero via L1-minimization. Our method combines features of the Dantzig selector (Candes and Tao, 2007) and the adaptive LASSO (Zou, 2006) and is easily implemented using linear programming. In addition to analytic results, we study the performance of our method via extensive simulations.

email: ldicker@hsph.harvard.edu

BIVARIATE DIMENSION REDUCTION AND VARIABLE SELECTION WHEN N<<P

Lexin Li, North Carolina State University
Xuerong Wen*, University of Missouri-Rolla

Under the general framework of sufficient dimension reduction, our method aims at achieving model-free dimension reduction and variable selections in regressions with bivariate responses under `large-p-small-n' settings. One immediate application is to microarray survival data. Our method can be applied directly to gene microarray survival data and pick out those significant (influential) genes without any modeling assumptions which can greatly facilitate further analysis.

email: wenx@umr.edu

A LINK-FREE METHOD FOR TESTING THE SIGNIFICANCE OF PREDICTORS

Peng Zeng*, Auburn University

One important step in regression analysis is to identify significant predictors from a pool of candidates so that a parsimonious model can be obtained using these significant predictors only. However, most of the available methods assume linear relationships between response and predictors, which may be inappropriate in some applications. In this talk, we discuss a link-free method that avoids specifying how the response depends on the predictors. Therefore, this method has no problem of model misspecification, and it is suitable for selecting significant predictors at the preliminary stage of data analysis. A test statistic is suggested and its asymptotic distribution is derived. Examples are used to demonstrate the proposed method.

email: zengpen@auburn.edu

19. RESEARCH METHODS

OUTCOME-DEPENDENT SAMPLING IN THE SURVIVAL ANALYSIS SETTING

Denise Esserman*, University of North Carolina at Chapel Hill
Yanyan Liu Wuhan, University of Peoples Republic of China
Haibo Zhou, University of North Carolina at Chapel Hill

Often times in a retrospective study, we are able to obtain complete data on our outcome variable of interest, but due most often to cost, we are unable to obtain full data on our exposure of interest. Work has been done using outcome-dependent sampling (ODS) schemes in the linear regression setting as a way of enhancing study efficiency and cost-effectiveness when only a subset of the full data set can be analyzed. We extend the idea of ODS to the survival setting. Through simulations, we compare the efficiency of our approach to the case-cohort method proposed by Prentice and weighted log-rank type estimator of Cai and Zeng, as well as to a simple random sample (SRS).

email: esserman@med.unc.edu

MARGINAL HAZARDS MODEL FOR CASE-COHORT STUDIES WITH MULTIPLE DISEASE OUTCOMES

Sangwook Kang*, University of Georgia
Jianwen Cai, University of North Carolina at Chapel Hill

A case-cohort study design was developed to reduce the cost of a large cohort study while achieving the same goals. A key advantage of the case-cohort design is its ability to use the same subcohort for several diseases or for subtypes of disease (e.g. Prentice, 1986; Langholz and Thomas, 1990). Since times to multiple disease outcomes from the same subject could be correlated, methods for a single disease outcome cannot be directly applied. A valid statistical method which takes the correlation into consideration needs to be developed. To this end, we consider marginal proportional hazards regression model and propose an estimating equation approach for parameter estimation with two different types of weights. We derive the asymptotic properties of the proposed estimators. Simulation studies are conducted to assess the finite sample properties of the proposed estimators and show that they perform well under the sample sizes considered. The proposed methods are illustrated to a data set from the Busselton Health Study.

email: skang@uga.edu

ESTIMATING FAMILY RELATIONSHIPS USING DNA FINGERPRINTS WITHIN THE NHANES-III HOUSEHOLD SURVEY

Hormuzd A. Katki*, National Cancer Institute
Christopher Sanders, National Center for Health Statistics
Barry I. Graubard, National Cancer Institute
Andrew Bergen, SRI International

The NHANES-III survey did not accurately assess familial relationships within household members. We estimated family relationships using each participant's Identifiler(tm) DNA fingerprint, a set of 15 DNA markers that overlap with the markers used by the FBI for forensic identification. We adapted two methods of estimating familial relatedness (an exact method requiring allele frequencies and a less efficient method that does not require allele frequencies) to survey data, accounting for genotyping error rates and potential correlation of alleles within ethnic groups. We both tested and estimated family relationships for all pairs of potential relatives within a household. This work will facilitate future family-based association and linkage studies within NHANES-III.

email: katkih@mail.nih.gov

FLEXIBLE MODELLING OF AGE-DEPENDENT INFECTIOUS DISEASE PARAMETERS USING SEROPREVALENCE DATA IN COMBINATION WITH SOCIAL CONTACT DATA

Marc Aerts*, Hasselt University-Belgium
Niel Hens, Hasselt University-Belgium
Nele Goeyvaerts, Hasselt University-Belgium
Kaatje Bollaerts, Hasselt University-Belgium
Ziv Shkedy, Hasselt University-Belgium
Christel Faes, Hasselt University-Belgium
Benson Ogunjimi, University of Antwerp-Belgium
Olivier Lejeune, University of Antwerp-Belgium
Pierre Van Damme, University of Antwerp-Belgium
Philippe Beutels, University of Antwerp-Belgium

Epidemiological parameters such as the prevalence, the force of infection, the transmission rate, and the reproduction number are playing a crucial role in models for infectious diseases. Whereas serological data allow estimation of the prevalence and the force of infection, social contact data contain the essential information about the way people mix. To get estimates for the transmission rates, both sources of data have to be combined appropriately into the model. Typically age is categorized, mainly because mathematical models are applied in this way. In this paper the objective is threefold: use of continuous type of data avoiding issues related to categorization; use of splines to model both serological and social contact data, to allow maximal flexibility; and use of bootstrap simulations to account for all sources of variability. The approach will be illustrated on Belgian data on Varicella Zoster Virus (VZV): serological data from the period November 2001 until March 2003; and social contact data from a diary survey conducted in a period from March until May 2006. This work has been funded by POLYMOD (European Sixth Framework Program, nr SSP22-CT-2004-502084), and by SIMID (Promotion of Innovation by Science and Technology in Flanders, nr 060081).

email: marc.aerts@uhasselt.be

BENCHMARK ANALYSIS FOR TWO PREDICTOR VARIABLES

Roland C. Deutsch*, University of North Carolina-Greensboro
John M. Grego, University of South Carolina
Brian T. Habing, University of South Carolina
Walter W. Piegorsch, University of Arizona

Benchmark analysis is a widely used tool in risk assessment, and the theory of finding estimators for minimum exposure levels (BMD) to induce a pre-specified benchmark risk (BMR) is well developed for the case of an adverse response to a single stimulus. In this presentation the authors extend this concept and introduce a statistical framework for conducting benchmark analysis with quantal response variables and two predictor variables. When considering two predictors the BMD is extended to a Benchmark Profile (BMP), a collection of exposure levels inducing the pre-specified BMR values. The estimation of the BMP is done by assuming a binomial response model based on a link function and finding the maximum likelihood estimators (MLE) for this model. Many inferential procedures in benchmark analysis require the asymptotic normality of the MLE. Easily verifiable regularity conditions establishing the asymptotic normality in binomial response models are presented and the quality of the normality approximation is studied for practical sample sizes. In addition, three methods of finding lower confidence bands on the BMP are presented and their coverages are studied via simulation.

email: rcdeutsc@uncg.edu

QUANTIFYING TREATMENT EFFECT WHEN FLEXIBLY MODELING INDIVIDUAL CHANGE IN A NONLINEAR MIXED EFFECTS MODEL

Robert J. Gallop*, West Chester University

The analysis of patterns of change during treatment may help identify unique mechanisms of change between treatments. Despite the importance of such work, progress in this area has been hampered by the fact that traditional data analytic strategies have not provided clear methods for examining the complexity of change processes over the course of treatment. Hierarchical linear models provided an important statistical advance in clinical trial methodology, the typical use of HLM assumes a linear trend across time for each person, which may not be valid for modeling clinical change trajectories. We will review the implementation and estimation of a flexible piecewise HLM, consisting of two treatment phases with differing rates of change for each treatment group: an early, rapid phase and a phase of reduced change. The breakpoint between these two phases is allowed to vary between

treatment groups as well as between individuals, such that the model fit is maximized. In addition to the implementation, this presentation compares various methods of quantifying the overall treatment efficacy. This is illustrated through data from a placebo-controlled trial comparing behavioral activation, cognitive therapy, and antidepressant medication in the treatment of depression.

email: rgallop@wcupa.edu

COVARIATE-ADJUSTED PUTATIVE PLACEBO ANALYSIS IN ACTIVE-CONTROLLED CLINICAL TRIALS

Zhiwei Zhang*, Center for Devices and Radiological Health-U.S. Food and Drug Administration

Even though an active-controlled trial provides no information about placebo, investigators and regulators often wonder how the experimental treatment would compare to placebo should a placebo arm be included in the study. A putative placebo analysis attempts to address this question by combining information from previous studies comparing the active control with placebo. Such an analysis often requires a constancy assumption, namely that the control effect relative to placebo is constant across studies. When the constancy assumption is in doubt, there are ad hoc methods that 'discount' the historical data in a conservative fashion. This paper presents a different approach that does not require constancy or involve discounting, but rather attempts to adjust for any imbalances in covariates between the current and historical studies. This covariate-adjusted approach is valid under a conditional constancy assumption which requires only that the control effect be constant within each subpopulation characterized by the observed covariates. Simulation results show that the proposed method performs reasonably well in moderate-sized samples. The method is illustrated with an example concerning benign prostate hyperplasia.

email: zhiwei.zhang@fda.hhs.gov

20. GENOMICS AND MICROARRAY ANALYSES

INFERRING THE TRUE CORRELATION IN CROSS-SPECIES MICROARRAY DATA

Sunghee Oh, University of Pittsburgh
George C. Tseng*, University of Pittsburgh

In the meta-analysis of microarray studies, gene effects are often poorly correlated across species. In this talk, we hypothesize that the poor correlation is due to the mixture of correlated genes and non-correlated genes in the genome. We propose to infer the underlying true correlation via an extreme-value correlation measure where the correlation is calculated only on genes of large absolute effects. We propose a Gaussian parametric approach and a non-parametric approach for the inference. The hypothesis and methods are validated by simulated data and an aging microarray data on human versus mouse comparison.

email: ctseng@pitt.edu

A LINEAR MIXED EFFECTS CLUSTERING MODEL FOR MULTI-SPECIES TIME COURSE GENE EXPRESSION DATA

Kevin H. Eng*, University of Wisconsin-Madison
Sunduz Keles, University of Wisconsin-Madison
Grace Wahba, University of Wisconsin-Madison

Environmental and evolutionary biology have recently benefited from the advances in experimental design and statistical analysis for complex gene expression microarray experiments. For example, the availability of high-throughput time course experiments allows the discovery of new and the investigation of old gene functions. We argue that their extension to sophisticated multi-factor designs incorporating closely related species allows the consideration of important biological questions. Motivated by time course gene expression experiments conducted in multiple strains of S. cerevisiae, we propose a regression model based clustering method preserving the factor information contained in time and in species and allowing inference on their parameters. We demonstrate via simulation that a mixed effects type model has good clustering properties and is robust to noise. On a multi-strain yeast data set, we give examples of detectable patterns and give suggestions for further analysis.

email: eng@stat.wisc.edu

RECONSTRUCTION OF GENETIC ASSOCIATION NETWORKS FROM MICROARRAY DATA: A PARTIAL LEAST SQUARES APPROACH

Vasyl Pihur*, University of Louisville
Somnath Datta, University of Louisville
Susmita Datta, University of Louisville

Gene association networks provide vast amounts of information about essential processes inside the cell. A complete picture of gene-gene interactions would open new horizons for biologists, ranging from pure appreciation to successful manipulation of biological pathways for therapeutic purposes. Therefore, identification of important biological complexes whose members (genes and their products proteins) interact with each other is of prime importance. Numerous experimental methods exist but, for the most part, they are costly and labor-intensive. Computational techniques, such as the one proposed in this work, provide a quick 'budget' solution that can be used as a screening tool before more expensive techniques are attempted. Here, we introduce a novel computational method based on the partial least squares (PLS) regression technique for reconstruction of genetic networks from microarray data. The proposed PLS method is shown to be an effective screening procedure for the detection of gene-gene interactions from microarray data. Both simulated and real microarray experiments show that the PLS based approach is superior to its competitors both in terms of performance and applicability.

email: v0pihu01@louisville.edu

MODELING SPATIAL CORRELATION IN GENE REGULATION

Guanghua Xiao*, University of Texas, Southwestern Medical Center
Cavan Reilly, University of Minnesota
Arkady B. 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 levels 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. By modeling these spatial correlations, we can obtain improved estimates of transcript levels. Here, we demonstrate the existence of spatial correlations in transcriptional activity in the Escherichia coli (E. coli) chromosome across more than 50 experimental conditions. Based on this finding, we propose a hierarchical Bayesian models that model the spatial correlation in gene expression. Furthermore, we extend the model to account for the structure of E. coli chromosome. The simulation studies and analysis of a real data example shows that the proposed method outperforms the commonly used SAM t statistic in detecting differentially expressed genes.

email: guanghua.xiao@utsouthwestern.edu

A VISUALIZATION OF NORMALITY TRANSFORMATIONS

Glen Satten, Centers for Disease Control and Prevention
Somnath Datta, University of Louisville
Bart Brown, University of Louisville
Guy Brock*, University of Louisville

We describe a simple way of visualizing the monotonic function that transforms data to a normal distribution. This graph can help us determine the right candidates for a parametric data transformation. It can also tell us when a simple parametric transformation may not be appropriate and we should consider nonparametric transformations of the data. A novel application to the calculation of a reference interval is provided.

email: guy.brock@louisville.edu

STATISTICAL TEST FOR THE ChIP-chip TILING ARRAYS WITHOUT REPLICATION

Youngchul Kim*, Seoul National University
Taesung Park, Seoul National University
Seungyeoun Lee, Sejong University
Jae K. Lee, University of Virginia

Recently, high-resolution tiling chromatin-immunoprecipitation chips (ChIP-chip) have being increasingly used to find the protein-binding sites, replication origins of chromosomes, and DNase hypersensitive sites. However, due to the non-ignorable noises and high-resolution of tiling arrays, it is very difficult to obtain a sufficient number of biological replicates of ChIP-chip tiling arrays with a high reproducibility. Further, not many solid statistical methods are currently available to analyze ChIP-chip tiling arrays. We propose a new statistical test to identify the transcription factor IID (TFIID) binding sites using the ChIP-chip tiling arrays without any replicate. The proposed method adopts a local error pooling method to control the high noise levels of tiling arrays caused by the correlations between the adjacent probes. The proposed method was applied to the real data application of 38 NimbleGene ChIP-chip tiling arrays containing a total of 14,535,659 50-mer oligonucleotides.

email: tspark@stats.snu.ac.kr

A META-ANALYSIS APPROACH FOR GENE ASSOCIATION NETWORK RECONSTRUCTION

YounJeong Choi*, University of Wisconsin-Madison
Christina M. Kendziorski, University of Wisconsin-Madison

Complex traits ranging from mRNA expression to disease phenotype are affected by multiple genes, and importantly, by interactions among those genes. Gene association networks (GANs) provide a simple, yet effective, summary of relationships among genes; and they are often used as a foundation upon which more complex inferences are made. In GANs, nodes represent genes and edges represent some measure of association between genes. The correlation coefficient is the most commonly used association measure, and GANs are often constructed by thresholding correlations so that a reasonable number of edges are identified. These study specific thresholds make interpretability and inter-study comparisons difficult. In addition, the edges identified correspond to relatively high correlations, a condition that is neither necessary nor sufficient among genes in known pathways. We here present a principled approach for GAN identification. The approach utilizes data from multiple studies, provides false discovery rate control, and, importantly, does not require that all genes in the network be highly correlated. An analysis of gene groups defined from Gene Ontology (GO) reveals the power of the approach to identify GANs important in lung cancer.

email: ychoi@stat.wisc.edu

21. EPIDEMIOLOGIC METHODS

ESTIMATING DISEASE PREVALENCE INDIRECTLY USING INCIDENCE AND SURVIVAL DATA

Kent R. Bailey*, Mayo Clinic

Disease prevalence is the proportion of people with the disease. It has large public health and economic significance, but may be difficult to directly assess. However, disease prevalence can be modeled as a function of disease incidence and the impact of the disease on survival. If one has: 1) a cohort of incidence cases from a population, 2) population census data, 3) survival follow-up for the incidence cases, and 4) standard mortality data reflective of the population, then one can model disease incidence and relative survival, by using the negative log of the expected survival probability through follow-up as the analysis variable. Given models for age and sex-specific incidence, relative survival, and overall population survival, we can then develop an age-recursive model for age-specific prevalence. The resulting age-sex-specific results can be recombined as if they belong to a single population. Many difficult issues arise: how to deal with trends over calendar year, how finely to subdivide age, how to model the components, how to deal with time-dependent effects of disease on relative survival, and how to generate standard errors. An example using atrial fibrillation data will be provided.

email: baileyk@mayo.edu

INFERENCE FOR MULTINOMIAL PARAMETERS AND ODDS RATIO UNDER ORDER RESTRICTIONS

Broderick O. Oluyede*, Georgia Southern University
Mavis Pararai, Indiana University of Pennsylvania

This paper presents strategies for the estimation and testing of binomial and multinomial parameters as well as odds ratios under order restrictions. Estimates of binomial and multinomial parameters under certain order restrictions are obtained and subsequently used in the estimation of the odds ratios. Also, estimates that make use of the marginal frequencies are obtained and studied. The estimated bias

and mean square error of the estimates are given. Test procedures are developed based on these estimates and their distributions given. Appropriate comparisons of the estimates as well as the developed test procedures are given for sparse to moderate samples. Furthermore, the results presented are extended to cover alternative reparametrization using appropriate terminology for failure time data.

email: boluyede@georgiasouthern.edu

ESTIMATION OF INCIDENCE AND REMISSION RATES FROM CROSS-SECTIONAL SURVEY DATA

Jason Roy*, Geisinger Center for Health Research
Walter Stewart, Geisinger Center for Health Research

Cross-sectional epidomioglic studies have often been criticized for containing limited information about the population. However, with a carefully constructed survey, we show how longitudinal information can be extracted from these studies. In particular, we develop statistical methods for estimating age-specific incidence and remission rates. We use a Bayesian semi-parameteric model to estimate smooth curves. The methodology is applied to a national migraine study. The cross-sectional design and analysis approach has several advantages over longitudinal studies.

email: jaroy@geisinger.edu

BAYESIAN INFERENCE FOR NON-ISOTROPIC CHANGES IN RISK ABOUT A FIXED POINT

Ronald E. Gangnon*, University of Wisconsin
Jane A. McElroy, University of Missouri

In environmental epidemiology, there is often interest in assessing the pattern of disease risk about a prespecified fixed point, e.g. a point source of pollution. In the absence of detailed exposure information, distance from the fixed point is frequently used as a surrogate measure of exposure, and patterns of risks assessed using parametric or isotonic regression estimators. In many settings, the assumption of isotropic decay in risk about the fixed point is overly restrictive. In this paper, non-isotropic decays in risk are modeled as order constraints on parameters; constraints are determined using a combination of distance and direction from the fixed point. Inferences for the order constrained model are obtained by projecting samples from an unconstrained posterior onto the constrained space. Comparisons of different potential constraints are conducted using DIC. Results from a simulation study are presented and the approach is illustrated using data on diffusion of early breast cancer diagnosis in Dane county, Wisconsin.

email: ronald@biostat.wisc.edu

USING MATCHING ESTIMATORS TO EXPLORE THE POTENTIAL INFLUENCE OF UNMEASURED SMOOTHLY TIME-VARYING CONFOUNDERS

Yun Lu*, Johns Hopkins University
Scott L. Zeger, Johns Hopkins University

Confounding bias is a very important problem in environmental epidemiologic studies where there often exist unmeasured time-varying confounders. We present a method to diagnose the possible influence of unmeasured confounders U on the estimated effect of exposure X on health outcome Y. We start with the time series case where X and Y are continuous variables at equally-spaced times and assume a linear model. We define matching estimator b(u)s that correspond to pairs of observations with specific lag u. When an unmeasured confounder U exists, and the model is otherwise correctly specified, the excess variation in b(u)s is evidence of confounding by U. We use the plot of b (u)s versus lag u, Lagged-Estimator-Plot (LEP), to diagnose the influence of U on the effect of X on Y. We use appropriate linear combination of b(u)s or extrapolate to b(0) to obtain more robust estimators. The methods can be extended to time series log-linear models. The LEP plot gives us a direct view of the magnitude of the estimators for each lag u and provides evidence when models did not adequately describe the data.

email: ylu@jhsph.edu

IDENTIFYING EFFECT MODIFIERS IN AIR POLLUTION TIMESERIES STUDIES USING A TWO-STAGE ANALYSIS

Sandrah P. Eckel*, Johns Hopkins Bloomberg School of Public Health
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health

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, city-specific estimates are combined into regional and national estimates using two-stage models. Our two-stage analysis identifies effect modifiers of the relation between single-day lagged PM10 and daily mortality in people age 65 and older from the 50 largest NMMAPS cities. We build on the standard approach by "fractionating" city-specific analyses to produce month-year-city specific estimated air pollution effects (slopes) in Stage I. In Stage II, we identify potential effect modifiers via weighted regression and weighted regression trees with the estimated slopes as dependent variables and predictors such as temperature, relative humidity, CO, NO2, O3, SO2, season, year, and other city-specific characteristics. We validate a sufficient condition on constructing Stage II predictors from daily measurements so that confounders omitted from the Stage I model are not spurious Stage II effect modifiers.

email: seckel@jhsph.edu

AN ALGORITHM FOR OPTIMAL TAPERED MATCHING, WITH APPLICATION TO DISPARITIES IN SURVIVAL

Shoshana R. Daniel*, University of Pennsylvania
Katrina Armstrong, University of Pennsylvania
Jeffrey H. Silber, University of Pennsylvania
Paul R. Rosenbaum, University of Pennsylvania

In a tapered matched comparison, one group of individuals, called the focal group, is compared to two or more nonoverlapping matched comparison groups constructed from one population in such a way that successive comparison groups increasingly resemble the focal group. An optimally tapered matching solves two problems simultaneously: it optimally divides the single comparison population into nonoverlapping

comparison groups and optimally pairs members of the focal group with members of each comparison group. We show how to use the optimal assignment algorithm in a new way to solve the optimally tapered matching problem, with implementation in R. This issue often arises in studies of groups defined by race, gender, or other categorizations such that equitable public policy might require an understanding of the mechanisms that produce disparate outcomes, where certain specific mechanisms would be judged illegitimate, necessitating reform. In particular, we use data from Medicare and the SEER Program of the National Cancer Institute as part of an ongoing study of black-white disparities in survival among women with endometrial cancer.

email: skrieger@mail.med.upenn.edu

22. BAYESIAN METHODS

HIERARCHICAL BAYES ESTIMATION FOR BIVARIATE BINARY DATA WITH APPLICATIONS TO SMALL AREA ESTIMATION

Malay Ghosh, University of Florida
Ananya Roy*, University of Nebraska-Lincoln
Ming-Hui Chen, University of Connecticut
Myron Katzoff, National Center for Heath Statistics
Van L. Parsons, National Center for Heath Statistics

The paper addresses small area estimation problems when the response is bivariate binary, and some of the covariates may be partially missing. Hierarchical Bayesian models which accommodate this missingness are considered. Estimators of small area means along with the associated posterior standard errors, and posterior correlations are provided. The method is applied to the analysis of a real dataset, and the superiority of the hierarchical Bayes estimators over the direct estimators is established.

email: aroy2@unlnotes.unl.edu

BAYESIAN ASSESSMENT OF HIERARCHICAL MODELS

Ying Yuan*, University of Texas, M.D.-Anderson Cancer Center
Valen Johnson, University of Texas, M.D.-Anderson Cancer Center

We propose a convenient and flexible diagnostic tool to assess the adequacy of hierarchical models based on pivotal quantities. Our method is based on the fact that the distribution of pivotal quantities evaluated at posterior draws of parameters is identical to the (known) distribution of pivotal quantities evaluated at true values of the parameters. By determining if posterior draws of a pivotal quantity follow the known reference distribution, we can assess the adequacy of hierarchical models. In contrast to many available methods, which often require intensive simulations, our method is computationally simple by directly utilizing outputs from Markov chain Monte Carlo algorithm without any additional simulations. Our approach allows using noninformative or vague priors, and is easily tailored to detect a particular model deviation of interest and perform level-wise model assessment. These features make our method especially suitable for early model building phase, where a large number of complex hierarchical models are assessed and screened. We illustrate our method with real datasets in the context of a linear normal hierarchical model and a logistic hierarchical model.

email: yyuan@mdanderson.org

JOINT EFFECT FROM ENVIRONMENTAL AND BEHAVIOR RISK FACTORS TO CANCER DEATH RATES USING BAYESIAN MULTI-LEVEL MODELING

Hongmei Zhang*, University of South Carolina

This project focuses on a secondary data analysis on Florida cancer death rates. Traditional researches usually either focus on environmental effects or behavior effects. In this paper, we consider joint effects from these two sources incorporating spatial structure. The effects are evaluated using a hierarchical spatial structure including regions, counties, and races. Bayesian multi-level Poisson models are constructed for this purpose. The advantage of multi-level modeling is that researchers are able to separately estimate the predictive effects of an individual predictor and its group-level mean. Simulations are used to demonstrate and evaluate the approach. We apply the approach to Florida cancer death rates. Further, based on the posterior distribution, we draw inferences on future death rates for some selected counties.

email: hzhang@gwm.sc.edu

BAYESIAN GENERALIZED PRODUCT PARTITION MODEL

Ju-Hyun Park*, University of North Carolina at Chapel Hill
David B. Dunson, National Institute of Environmental Health Sciences

Starting with a carefully formulated Dirichlet process (DP) mixture model, we derive a generalized product partition model (GPPM) in which the partition process is predictor dependent. The GPPM generalizes DP clustering to relax the exchangeability assumption through the incorporation of predictors, resulting in a generalized Polya urn scheme. In addition, the GPPM can be used for formulating flexible semiparametric Bayes models for conditional distribution estimation, bypassing the need for expensive computation of large numbers of unknowns characterizing priors for dependent collections of random probability measures. Properties are discussed, a variety of special cases are considered, and an efficient Gibbs sampling algorithm is developed for posterior computation. The methods are illustrated using simulation examples and an epidemiologic application.

email: juhyunp@email.unc.edu

BAYESIAN MIXTURE LABELLING BY HIGHEST POSTERIOR DENSITY

Weixin Yao*, Kansas State University
Bruce G. Lindsay, The Pennsylvania State University

One of the most fundamental problems for Bayesian mixtures is label switching, due to the non-identifiability of the mixture components under symmetric priors. We propose two labelling methods to solve this problem. The first labelling method, denoted by PM(ECM), is based on the posterior modes and the ECM algorithm. The PM(ECM) labelling method automatically matches the “ideal” labels in the highest posterior density (HPD) credible region. We also consider a computationally easier labelling method. This method is to do labelling by maximizing the normal likelihood (NORMLH) of the labelled Gibbs samples. Using a Monte Carlo simulation study we demonstrate that our new methods are superior at reproducing the ideal labelling to some existing methods.

email: wxyao@ksu.edu

A PRACTICAL PROCEDURE TO FIND MATCHING PRIORS FOR FREQUENTIST INFERENCE

Juan Zhang*, Rutgers University
John E. Kolassa, Rutgers University

We give a practical way to find the matching priors proposed by Welch and Peers (1963) and Peers (1965). Then we investigate the use of saddlepoint approximations combined with matching priors and obtain p-values of the test of interest. The advantage of our procedure is the flexibility of choosing different initial conditions so that one can adjust the performance of the test. Two examples have been studied via Monte Carlo simulation. One relates to the ratio of two exponential means, and the other is about the logistic regression model. One of the numerical studies is under small sample size settings.

email: janezh@stat.rutgers.edu

HOW TO CHOOSE HYPERPARAMETER VALUES IN NORMAL MODELS

Susan Alber*, University of Texas, M.D.-Anderson Cancer Center
J. Jack Lee, University of Texas, M.D.-Anderson Cancer Center

The primary objection to using Bayesian models for medical research is the dependence on the choice of the prior, and the resulting potential for results to be influenced by subjective preconceptions that may be incorrect. However, a carefully chosen prior can improve the accuracy of estimated parameters. We investigate the sensitivity of the posterior to the choice of hyperparameters for a normal sampling distribution with a conjugate prior. Under an assumed model for the true data generating mechanism, we present frequentist properties of posterior point and interval estimates as a function of hyperparameter values. We develop two approaches for selecting hyperparameters that yield posterior estimation that is in some way optimal. First, we show which hyperparameter values minimizes the mean square error of the posterior point estimate. We also define the credible interval optimal hyperparameter values as those that minimize the length of the posterior credible interval while still maintaining the nominal coverage. We use our results as a basis to provide practical advice for choosing hyperparameter values under the usual situation where the data generating model is not known. This approach can be extended to more general models.

email: salber@mdanderson.org

23. LATENT VARIABLES AND STRUCTURAL EQUATIONS MODELING

EXPLORATORY STRUCTURAL EQUATION MODELING

Bengt Muthen*, University of California at Los Angeles
Tihomir Asparouhov, Mplus

Exploratory factor analysis (EFA) has been said to be the most frequently used multivariate analysis technique in statistics. In 1966 Jennrich solved a significant EFA rotation problem by deriving the direct quartimin rotation. He was also the first to develop standard errors for rotated solutions although these have still not made their way into most statistical software programs. This is perhaps because Jennrich's achievements were overshadowed by the 1967 development of confirmatory factor analysis (CFA) by Joreskog. Joreskog developed CFA further into structural equation modeling (SEM) where CFA was used for the measurement part of the model. The strict requirement of zero cross-loadings in CFA does, however, often not fit the data and has led to a tendency to rely on extensive model modification to find a well-fitting model. Furthermore, misspecification of zero loadings tends to give distorted factors with over-estimated factor correlations and subsequent distorted structural relations. This paper describes an EFA-SEM approach, where instead of a CFA measurement model, an EFA measurement model with rotations is used in a structural equation model. The resulting transformation of structural coefficients is described. Standard errors and overall tests of model fit are derived. Simulated and real data are used to illustrate the points.

email: asparouhov@hotmail.com

INCORPORATING STRUCTURAL EQUATION MODELING TO FECUNDABILITY MODELS

Sung Duk Kim*, National Institute of Child Health and Human Development, NIH
Rajeshwari Sundaram, National Institute of Child Health and Human Development, NIH

In epidemiologic studies, human fecundability is measured by probability of conception in a menstrual cycle by a couple. There has been considerable statistical literature in finding better models to fit data for probabilities of conception by incorporating various biological and behavioral aspects starting with the work of Barrett and Marshall (1969) and recently by Dunson. Here, we propose a general model for joint modeling of intercourse behavior and human fecundability through a classic conception probability model and structural equation model (SEM) which involves a set of latent variables to capture dependence between intercourse behavior on consecutive days in a menstrual cycle, so that the proposed model can accommodate not only a broad variety of intercourse patterns and dependency structure, but also general covariate effects, heterogeneity among fecund couples in menstrual cycle viability and in frequency of intercourse. Further, we also consider the dependence between menstrual cycles for a couple using random effects. Markov Chain Monte Carlo method is used to carry out Bayesian posterior computation. Several variations of the proposed model are considered and compared via the deviance information criterion. A detailed analysis of the New York State Angler pregnancy data is presented to illustrate the proposed methodology.

email: kims2@mail.nih.gov

STRUCTURAL EQUATION MODELS WITH FUNCTIONAL LATENT VARIABLES

Mingan Yang*, National Institute of Environmental Health Sciences-NIH
David B. Dunson, National Institute of Environmental Health Sciences-NIH
Pablo Nepomnaschy, National Institute of Environmental Health Sciences-NIH

Structural equation models with latent variables (SEMs) have proven increasingly useful in a wide variety of application areas, providing a general framework for modeling of multivariate data with mixed measurement scales and for investigating relationships among latent variables. This article focuses on generalizing typical SEMs to accommodate functional latent variables, as may be naturally considered in longitudinal and spatial applications. We take a semiparametric Bayes approach, allowing not only the mean but also the distribution of latent functions to be unknown. This is accomplished by expressing the unknown functions as linear combinations of basis functions, with the joint distribution of the subject-specific basis coefficients characterized using a linear structural relations model (LISREL). By using centered Dirichlet process mixtures to characterize the random components of the LISREL model, we allow subject-specific functions to vary flexibly while borrowing information. Posterior computation proceeds via an efficient parameter-expanded Gibbs sampling algorithm. The methods are motivated by an application to a longitudinal andropology study of multiple sources of stress and reproductive health.

email: yangm2@niehs.nih.gov

MULTILEVEL LATENT CLASS MODELS WITH DIRICHLET MIXING DISTRIBUTION

Chongzhi Di*, Johns Hopkins University
Karen Bandeen-Roche, Johns Hopkins University

Latent class analysis (LCA) and latent class regression (LCR) are widely used for modeling multivariate categorical outcomes in social sciences and biomedical studies. Standard analyses assume data of different respondents to be mutually independent, limiting application of the methods to familial and other designs in which study participants are clustered. In this paper, we develop multilevel latent class models with Dirichlet mixing distribution. We assume that the subpopulation mixing probabilities are random effects which vary from cluster to cluster. We apply the EM algorithm for model fitting by maximum likelihood (ML). This approach is computationally reasonable when the number of classes and cluster sizes are small. When either of them is large, computational burden may be substantial. We propose a maximum pairwise likelihood (MPL) approach for this case. We also show that a simple LCA that ignores clustering, combined with robust standard errors, provides another consistent, robust, but less efficient inferential procedure. Simulation studies suggest that the three methods work well in finite samples, and the maximum pairwise likelihood estimates (MPLE) often enjoy high efficiency compared to the MLE. Finally, we apply our method to the analysis of comorbid symptoms in the Hopkins Obsessive Compulsive Disorder study.

email: cdi@jhsph.edu

HIERARCHICAL MIXED MEMBERSHIP MODELS FOR HONEY BEE GENOMES

Tanzy M. Love*, Carnegie Mellon University

In the New World, all existing Apis mellifera (honey bees) are introduced from existing (different) populations in the Old World. One effect of this is the prevalence of European honey bees in the north and African 'killer' honey bees in south (these frighten some by moving). Using 1136 SNPs genotyped in 341 individuals, Whitfield et al. (2006) characterized both native and introduced populations of honey bees over time. In their paper, they clustered native genomes into four genetic classes, and then compared introduced genomes to these clusters to measure the change in the honey bee gene pools in the Americas over time. We propose a mixed membership model for honey bee genomes where any particular bee's genome is a mixture of several pure types. With this genetic example, mixed membership is a more reasonable structure than hard clustering. Using a Dirichlet process prior, we can estimate the true number of pure types along with the mixture in each of the observed individuals. These pure types are interpretable as ancestral genotypes of which current bees have mixed heritage.

email: tanzy@cmu.edu

CONFOUNDING EFFECTS IN GENETIC ANALYSIS

Mariza de Andrade*, Mayo Clinic
Brooke L. Fridley, Mayo Clinic
Stephen T. Turner, Mayo Clinic

In quantitative trait linkage analyses, investigators may sometimes observe two linkage peaks in the same genomic region. It is usually thought to be the consequence of two causative loci in the same region. However, this may due to a hidden effect of a covariate. Using two distinct data sets, one using real data from GENOA Phase II sibships and the other using GAW 13 simulated data, we observed this hidden effect that we called "the bunny ears" syndrome. In the simulated data, the trait of interest was systolic blood pressure (SBP) and the covariate was height. Height was simulated to have an impact in the level of SBP, and also had a major gene effect in the same region as the major gene for SBP. When the linkage analysis was performed for SBP without adjusting for height, two linkage peaks appeared. After adjusting for height, one of the peaks disappeared and only the peak due the SBP major gene remained. In the real data, the same was observed when performing a linkage analysis with brain atrophy (BA) as the trait of interest and total intracranial volume (TIV) as the covariate. We observed this effect in the chromosome 17p region when the linkage analysis was performed without adjusting for TIV. After adjusting for TIV the bunny ears disappeared and only one peak remained for BA. These results emphasize the importance of adjusting for covariates that may otherwise exert confounding genetic effects.

email: mandrade@mayo.edu

A GENOMIC IMPRINTING TEST FOR ORDINAL TRAITS IN PEDIGREE DATA

Rui Feng*, University of Alabama at Birmingham
Heping Zhang, Yale University

Genomic imprinting can lead maternally and paternally derived alleles with identical nucleotide sequences to function differently and has been found to affect the complex inheritance of a variety of human disorders. Statistical methods that differentiate the parent-of-origin effects on human diseases are available for binary traits and continuous traits. However, numerous common diseases are measured on discrete ordinal scales. Imprinting may also contribute to the complex genetic basis of these traits. In a previous study, we proposed a latent variable model and developed computationally efficient score statistic to test linkage of ordinal traits for any size pedigree while adjusting for non-genetic covariates. In this study, we extend the latent variable model to incorporate parent-of-origin information and further develop a score statistic for testing the imprinting effect in linkage analysis. We evaluated the properties of our test statistic using simulations. We then applied our method to the Collaborative Study on the Genetics of Alcoholism (COGA) and found a novel locus on chromosome 18 that shows a strong signal for imprinting. In addition, we identified two loci on chromosomes 3 and 4 significantly (p-value < 0.0001) linked with alcoholism.

email: rfeng@ms.soph.uab.edu

24. NON-INFERIORITY CLINICAL TRIALS - FIXED MARGIN OR NO FIXED MARGIN

CONTROLLING THE TYPE 1 ERROR RATE IN THE PRESENCE OF DEPARTURES FROM THE ASSUMPTIONS OF ASSAY SENSITIVITY AND CONSTANCY

Steven Snapinn*, Amgen Inc.
Qi Jiang, Amgen Inc.

Two different approaches have been proposed for establishing the efficacy of an experimental therapy through a non-inferiority trial. The fixed-margin approach involves first defining a non-inferiority margin and then demonstrating that the experimental therapy is not worse than the control by more than this amount, and the synthesis approach involves combining the data from the non-inferiority trial with the data from historical trials evaluating the effect of the control. This presentation will describe a unified approach that includes both of these approaches, and show how the parameters of this approach can be selected to control the type 1 error rate in the presence of departures from the assumptions of assay sensitivity and constancy. Using this approach it is shown that an appropriately chosen synthesis method is always more efficient than the fixed-margin method that achieves the same control of the type 1 error rate.

email: ssnapinn@amgen.com

ASSESSING THE EVIDENCE FOR NON-INFERIORITY: A LIKELIHOOD APPROACH

Sue-Jane Wang, U.S. Food and Drug Administration
Jeffrey Blume*, Brown University

One of the most difficult tasks in non-inferiority trials is the selection of a margin that rules out the inferiority of the new treatment relative to its active comparator. Likelihood methods characterize and measure the strength of statistical evidence across the entire parameter space in a way that does not depend on the specification of the non-inferiority margin, but, with the understanding that the pre-specified non-inferiority margin is of primary interest. This is an important advantage, because investigators who have different clinical assessments of what the non-inferiority margin should be would still agree on the strength of evidence the empirical data provide. In this paper presentation, we discuss how to use likelihood methods in the setting of non-inferiority trials. We pay particular attention to the role of selecting the non-inferiority margin. We also assess the smallest margin supported by the data, incorporating the discounting of the estimated control effect to further refine the fixed margin, using historical trials to provide information about the control effect, and performing sensitivity analyses for a range of margins. Finally, we discuss the probability that these likelihood methods will be misleading and show that this probability is both low and controllable under quite general conditions.

email: suejane.wang@fda.hhs.gov

EMPIRICAL DATA ON THE CHOICE OF IRRELEVANCE MARGINS RESULTS FROM A SYSTEMATIC REVIEW

Stefan Lange*, Institute for Quality and Efficiency in Health Care-Germany
Guido Skipka, Institute for Quality and Efficiency in Health Care-Germany

In the past, regulatory guidance and recommendations on biostatistical methodology contained only general statements on the selection of an appropriate non-inferiority (irrelevance) margin (delta). This situation inspired a systematic review on the choice of delta in published non-inferiority or equivalence trials. Overall, 332 relevant trials were identified by means of an extensive literature search. In about 50% of these trials, a difference of 0.5 standard deviations (corresponding to an odds ratio of about 2) or more was chosen as the non-inferiority margin and consequently regarded as "irrelevant". Looking at trials with mortality as (part of) the primary endpoint, the median of the non-inferiority margins amounts to an odds ratio of 1.4. While nearly 60% of the trials did not make any statement on the chosen non-inferiority margin, less than 30% provided a more substantial rationale for the choice than simply describing the margin as, for example, "clinically irrelevant". Only about 10% of trials based their choice on estimates of ("historical") standard-placebo differences. Reasons for this very low proportion might be (1) insufficient "historical" information, and (2) the very small margins possibly resulting. Hence, a more global definition of "irrelevance" might be warranted.

email: stefan.lange@iqwig.de

25. MODEL SELECTION FOR HIGH DIMENSIONAL DATA WITH PRACTICAL SOLUTIONS

SPLINE ESTIMATION OF SINGLE INDEX MODELS

Lijian Yang*, Michigan State University
Li Wang, University of Georgia

For the past two decades, single-index model, a special case of projection pursuit regression, has proven to be an efficient way of coping with the high dimensional problem in nonparametric regression. In this paper, based on weakly dependent sample, we investigate a robust single-index model, where the single-index is identified by the best approximation to the multivariate prediction function of the response variable, regardless of whether the prediction function is a genuine single-index function. A polynomial spline estimator is proposed for the single-index coefficients, and is shown to be root-n consistent and asymptotically normal. An iterative optimization routine is used which is sufficiently fast for the user to analyze large data of high dimension within seconds. Simulation experiments have provided strong evidence that corroborates with the asymptotic theory. Application of the proposed procedure to the river flow data of Iceland has yielded superior out-of-sample rolling forecasts.

email: yang@stt.msu.edu

PARSIMONIOUS MODELS FOR CORRELATION MATRICES USING ORDERED PARTIAL CORRELATIONS

Michael J. Daniels*, University of Florida
Mohsen Pourahmadi, Northern Illinois University

We propose an approach to parsimoniously model a correlation matrix using a particular set of ordered partial correlations. These correlations are free to vary independently in the interval (-1,1) and lend themselves to constructing parsimonious models for ordered (e.g., longitudinal) data. We propose various parametric models, explore the behavior of several prior specifications, and discuss Bayesian variable selection techniques to reduce the dimension of the correlation matrix.

email: mdaniels@stat.ufl.edu

SEMIPARAMETRIC MODELING OF CLUSTERED BIOMEDICAL DATA

Naisyin Wang*, Texas A&M University
Annie Qu, Oregon State University

The analysis of hierarchical biomedical data sometimes requires more modeling flexibility than that which can be provided by standard parametric approaches. It is commonly believed that the effect of ignoring covariance structure is mainly the loss of efficiency. There are situations wherein estimation biases could also be concerns. We argue that the modeling of covariance in a multi-layered covariate process is in fact a very important task in joint modeling approaches. I will use some recent applications as examples to illustrate the potential problems to be considered and provide some semiparametric solutions which are easily computable. I will use data from on-going biomedical studies to illustrate a few key points in the modeling strategy.

email: nwangstat@gmail.com

GENERALIZATION ERROR ESTIMATION IN SEMISUPERVISED LEARNING

Junhui Wang*, Columbia University
Xiaotong Shen, University of Minnesota

In classification, an estimated generalization error is often used to quantify a classifier's generalization ability. As a result, quality of estimation of the generalization error becomes crucial in tuning classifiers. In this talk, we propose an estimation methodology for the generalization error for semisupervised learning, where a large amount of unlabeled data is available with only a small number of labeled data. To leverage unlabeled data for enhancing estimation accuracy, we introduces an novel loss function for unlabeled data, which seeks efficient extraction of the information from unlabeled data for estimating the generalization error. In particular, I will discuss the motivating idea and methodology development as well as data perturbation technique. Numerical examples will be provided to demonstrate the advantage of our proposed methodology against other estimating techniques using labeled data alone.

email: jwang@stat.columbia.edu

26. SPATIAL AND SPATIO-TEMPORAL LATENT STRUCTURE MODELING

A SPATIAL LATENT CLASS MODEL FOR MULTIVARIATE SPATIAL DATA

Melanie M. Wall*, University of Minnesota

The motivating data for this talk come from soil samples collected across a large geographic area. At each site, eight different heavy metals were measured and the data represent whether the level of each of the heavy metals is over the legal threshold or not. The primary goal is to identify which locations are more polluted than others and to explore whether the different make-up of heavy metals can be used to identify different pollution sources. The latent class model is a popular multivariate finite mixture model for binary data that explains dependences between variables by identifying clusters (i.e. latent classes) within which variables are independent. For the soil data, the eight variables are not only correlated within location by also across locations. A spatial latent class model is introduced with a hierarchical structure where at one level the relationships between the observed indicators and the latent classes are defined, and the second level a multinomial probit is used to model the latent classes by introducing a spatially varying continuous random variables underlying the latent classes, thus directly modeling spatial correlations between the latent class at different sites. Implementation is demonstrated using a fully Bayesian framework.

email: melanie@biostat.umn.edu

SPATIAL DYNAMIC FACTOR ANALYSIS

Hedibert F. Lopes, University of Chicago
Esther Salazar, Universidade Federal do Rio de Janeiro
Dani Gamerman*, Universidade Federal do Rio de Janeiro

A new class of space-time models derived from standard dynamic factor models is proposed. The temporal dependence is modeled by latent factors while the spatial dependence is modeled by the factor loadings. Factor analytic arguments are used to help identify temporal components that summarize most of the spatial variation of a given region. The temporal evolution of the factors is described in a number of forms to account for different aspects of time variation such as trend and seasonality. The spatial dependence is incorporated into the factor loadings by a combination of deterministic and stochastic elements thus giving them more flexibility and generalizing previous approaches. The new structure implies nonseparable space-time variation to observables, despite its conditionally independent nature, while reducing the overall dimensionality, and hence complexity, of the problem. The number of factors is treated as another unknown parameter and fully Bayesian inference is performed via a reversible jump Markov Chain Monte Carlo algorithm. The new class of models is tested against one synthetic dataset and applied to real data obtained from the Clean Air Status and Trends Network (CASTNet). The factor model decomposition is shown to capture important aspects of spatial and temporal behavior of the data.

email: dani@im.ufrj.br

MIXTURE- AND GAM-BASED MODELS FOR SPACE-TIME LATENT STRUCTURE IN HEALTH DATA

Bo Cai, University of South Carolina
Andrew B. Lawson*, University of South Carolina
Kun Huang, University of South Carolina

The belief in underlying structure in spatio-temporal data leads to the consideration of two novel approaches to the structural modeling. The first approach is a mean mixture approach where a set of latent temporal components are considered within a linear combination with weight components. The temporal effects are temporal-only dependent while the weights are spatially dependent. We consider entry variables with beta priors for component selection. The second model assumes a GAM-like structure with ZI-mixture priors with unobserved components defined by basis functions. This approach deals well with variable selection and also allows for a flexible non-parametric form for the components. Both approaches are evaluated via simulation studies and application to real data is also provided.

email: alawson@gwm.sc.edu

27. NEW ADVANCES IN SURVIVAL ANALYSIS FOR LARGE DIMENSIONAL BIOMEDICAL DATA

INVERSE REGRESSION FOR CENSORED DATA

Nivedita Nadkarni, University of North Carolina-Chapel Hill
Michael R. Kosorok*, University of North Carolina-Chapel Hill

An inverse regression methodology for right censored data is developed along with inference procedures and a computational algorithm. A low-dimensional representation of a high-dimensional covariate space is estimated without assuming models for either the failure or censoring times. The implementation is nonparametric and computationally fast. The approach can be very useful as a diagnostic tool in the model selection process. In addition to theoretical justification of consistency, both a simulation study and data analysis are provided to illustrate the practical utility of the procedure.

email: kosorok@unc.edu

ANALYZING EVENT TIME DATA IN THE PRESENCE OF HIGH-DIMENSIONAL CONFOUNDERS

Donglin Zeng*, University of North Carolina
Danyu Lin, University of North Carolina

We propose to fit the proportional hazards model to analyze event time data when high-dimensional confounders are present. To control for the confounders, a single-index function of the confounders is used in the regression. Penalized maximum likelihood estimation is used for inference where a non-convex penalty is used for selecting the number of important confounders and a sobolev-type penalty is used to control the smoothness of the unknown single-index function. We have derived the large sample properties of the proposed estiamtors and obtained the oracle property in selecting important confounders. The small-sample performance of our approach is demonstrated via numerical studies.

email: dzeng@bios.unc.edu

SURVIVAL ANALYSIS OF CASE-CONTROL FAMILY DATA WITH GENERAL SEMI-PARAMETRIC SHARED FRAILTY MODEL AND MISSING GENETIC INFORMATION

Anna Graber*, Technion – Israel Institute of Technology
Malka Gorfine, Technion – Israel Institute of Technology
Li Hsu, Fred Hutchinson Cancer Research Center

Case-control family studies are now widely used to study the role of gene-environment interactions in the etiology of complex diseases. A typical case-control family study includes a random sample of independent diseased individuals (case-probands) and non-diseased individuals (control-probands), along with their family members. An array of genetic and environmental risk factor measures is collected on these individuals. Frequently, however, some of the risk factors are observed on the probands but not on their relatives (e.g. genetic information). In this work we consider correlated failure time data arising from population-based case-control family studies with missing genotypes of relatives. A new method for estimating the risk factors' effect, the within family dependence parameter and the cumulative baseline hazard function is presented. We present simulation studies to assess performance of the proposed methodology, and illustrate its utility on a real data example.

email: agraber@tx.technion.ac.il

SURVIVAL ANALYSIS WITH LARGE DIMENSIONAL COVARIATES: AN APPLICATION IN MICROARRAY STUDIES

David Engler*, Brigham Young University
Yi Li, Dana-Farber Cancer Institute

Use of microarray technology often leads to high-dimensional, low-sample size data settings. Over the past several years, a variety of approaches have been proprosed for variable selection in this context. However, only a small number of these are applicable to time-to-event data where censoring is present. Among standard variable selection methods shown both to have good predictive accuracy and to be computationally efficient is the elastic net penalization approach. In this paper, adaptation of the elastic net penalized estimation approach is presented for variable selection both under the Cox proportional hazards model and under an accelerated failure time (AFT) model. Assessment of the two methods is conducted through simulation studies and through analysis of microarray data obtained from a set of patients with diffuse large B-cell lymphoma where time to survival is of interest. The approaches are shown to be an improvement over existing methods in terms of both computational efficiency and prediction accuracy.

email: engler@byu.edu

28. ON THE UTILITY OF DETERMINISTIC MODELS FOR CAUSAL EFFECTS

CAUSAL INFERENCE FOR CONTINUOUS TIME PROCESSES OBSERVED ONLY AT DISCRETE TIMES

Mingyuan Zhang*, University of Pennsylvania
Marshall M. Joffe, University of Pennsylvania
Dylan Small, University of Pennsylvania

Most current researches on and applications of Structural Nested Model and g-estimation for causal inference in longitudinal data assume a discrete time underlying data generating process. However, in some observational studies, it is more reasonable to assume that the data are generated from a continuous time process, and only observable at discrete time points. Under these circumstances, sequential randomization in the observed discrete time data, the key assumption for the discrete time g-estimation, may not be reasonable. With a deterministic model, we discuss one set of conditions on the underlying continuous time processes under which the discrete time g-estimation can continue to work on the observed discrete time data. In more general cases, when the discrete time g-estimation gives inconsistent estimation for causal parameters of interest, we propose a new method that provides at least as good performance as g-estimation in most scenarios, and provides consistent estimation in some cases in which g-estimation is severely inconsistent. We use our method to study the effect of diarrhea on children's height, using a data set collected following a massive flood in Bangladesh.

email: zhangmi@wharton.upenn.edu

COMPARISON OF DETERMINISTIC VERSUS STOCHASTIC MONOTONICITY ASSUMPTIONS FOR THE INSTRUMENTAL VARIABLES METHOD

Dylan S. Small*, University of Pennsylvania

The instrumental variables (IV) method is a method for making causal inferences about the effect of a treatment when there are unmeasured confounding variables. The method requires a valid IV, a variable that is associated with the treatment but is independent of unmeasured confounding variables and has no effect on the outcome beyond its effect on the treatment. An additional assumption that is often made is deterministic monotonicity, which is that for each subject, the level of treatment that a subject would take if given a level of IV is a monotonic increasing function of the level of IV. Under deterministic monotonicity, the IV method identifies the sign of the average treatment effect when the sign of each subject's treatment effect is the same (the no sign reversal property). However, deterministic monotonicity is sometimes not realistic. We introduce a stochastic monotonicity condition which does not require that a monotonic increasing relationship hold within subjects between the levels of the IV and treatment, but only that a monotonic increasing relationship hold across subjects between the IV and treatment in a certain manner. We show that under stochastic monotonicity, the no sign reversal property of the IV method still holds.

email: dsmall@wharton.upenn.edu

MIMICKING COUNTERFACTUAL OUTCOMES TO AVOID DETERMINISTIC TREATMENT EFFECTS

Judith J. Lok*, Harvard School of Public Health

In observational studies the treatment may be adapted to covariates at several times without a fixed protocol. The treatment influences covariates, which in turn influence the treatment, and so on. This dynamic confounding may happen in continuous time. Then, even time-dependent Cox-models, which specify the hazard, do not estimate the net treatment effect consistently. Robins has proposed the so-called Structural Nested Models to estimate treatment effects in this setting. This methodology compares counterfactual distributions of outcomes that patients would have experienced had the treatment been withheld after different times. Previous work on these models assumed that counterfactuals depend deterministically on observed data, conjecturing that this assumption could be relaxed. However, the assumption of deterministic treatment effects has frequently been criticized as unrealistic in most practical contexts. In this talk, I demonstrate that the assumption of deterministic treatment effects can indeed be relaxed. I prove that we can construct mimicking variables that have the same distribution as the counterfactuals, even given past observed data. These variables can be used to estimate treatment effects without assuming them to be deterministic. This is possible even in a continuous time setting. I hope that this talk will contribute to the discussion about causal reasoning.

email: jlok@hsph.harvard.edu

CAUSAL MODELS FOR THE EFFECTS OF WEIGHT GAIN ON MORTALITY

James M. Robins*, Harvard University

Suppose, contrary to fact, in 1950, we had put the cohort of 18 year old non-smoking American men on a stringent mandatory diet that guaranteed that no one would ever weigh more than their baseline weight established at age 18. How would the counter-factual mortality of these 18 year olds have compared to their actual observed mortality through 2007? We describe in detail how this counterfactual contrast could be estimated from longitudinal data by applying g-estimation of structural nested models. Our analytic approach adjusts for (i) measured time-varying confounders that are simultaneously intermediate variables , (ii) unmeasured confounding by undiagnosed preclinical disease (i.e reverse causation) [provided an upper bound can be specified for the maximum length of time a subject may suffer from a subclinical illness severe enough to affect his weight without the illness becomes clinically manifest], and (iii) the prescence of identifiable subgroups, such as those suffering from serious renal, liver, pulmonary, and/or cardiac disease, in whom confounding by unmeasured prognostic factors is believed to be so severe as to render useless any attempt at direct analytic adjustment for confounding.

email: robins@hsph.harvard.edu

29. ESTIMATING THE DISTRIBUTION OF USUAL INTAKES OF NUTRIENTS AND FOODS, AND RELATING USUAL INTAKE TO HEALTH PARAMETERS IN A NATIONAL HEALTH SURVEY

CHALLENGES IN THE ESTIMATION OF USUAL INTAKE OF FOODS AND NUTRIENTS

Patricia M. Guenther*, U.S. Department of Agriculture
Dennis W. Buckman, Information Management Services, Inc.
Raymond J. Carroll, Texas A&M University
Kevin W. Dodd, National Cancer Institute
Laurence S. Freedman, Gertner Institute for Epidemiology and Health Policy Research
Victor Kipnis, National Cancer Institute
Susan M. Krebs-Smith, National Cancer Institute
Douglas Midthune, National Cancer Institute
Amy F. Subar, National Cancer Institute
Janet A. Tooze, Wake Forest University School of Medicine

The setting of effective nutrition and food safety policies requires knowledge of what target populations are eating. Estimates of distributions of usual, or long-term average, intakes are needed both when setting standards, such as the Dietary Reference Intakes for nutrients; when assessing compliance with standards, such as those found for food groups in MyPyramid, the USDA food guide; and when conducting risk assessments related to food additives or contaminants. Because standards are established regarding intakes over time, the estimate of the population's usual intake distribution is needed. Federal nutrition surveys employ interviewer-administered 24-hour recalls of dietary intake because of the detailed information they provide. This approach is based on the assumption that an individual can more accurately recall and describe the foods eaten yesterday than over a longer period of time. However, because people do not eat the same thing every day and because it is feasible to administer only a small number of recalls, statistical models have been developed that remove day-to-day variability. Cohort or case-control studies that seek to establish relationships between diet and disease, on the other hand, typically employ self-administered food frequency questionnaires that ask participants how often they consume a pre-determined list of foods.

email: patricia.guenther@cnpp.usda.gov

A TWO-PART MIXED EFFECTS MODEL WITH CORRELATED RANDOM EFFECTS TO MODEL FOOD AND NUTRIENT INTAKE

Janet A. Tooze*, Wake Forest University School of Medicine
Doug Midthune, National Cancer Institute
Kevin W. Dodd, National Cancer Institute
Laurence S. Freedman, Gertner Institute for Epidemiology and Health Policy Research
Susan M. Krebs-Smith, National Cancer Institute
Amy F. Subar, National Cancer Institute
Patricia M. Guenther, U.S. Department of Agriculture
Raymond J. Carroll, Texas A&M University
Victor Kipnis, National Cancer Institute

Statistical methods must be used to estimate usual, or long-term, intake of foods when only a small number of dietary assessments are available on an individual. These methods must account for the within-person error that arises in data of this type, as well as for the right skewness of the data. Additionally, for foods that are not consumed every day, the method must accommodate the presence of excess zeroes, and the correlation between the probability of consuming a food and the amount consumed on days when it is consumed. Finally, researchers are interested in incorporating covariates into models to facilitate subpopulation estimates and comparisons and adjust for these variables. This talk will discuss the development of a two-part mixed effects model with correlated random effects, which can accommodate the large number of non-consumption days that arise for episodically consumed foods, the correlation that usually exists between the probability of consuming a food and the amount consumed, incorporate covariates, and accommodate non-normal error distributions. The model will be presented in a measurement error framework, and applications of the model will be introduced.

email: jtooze@wfubmc.edu

ESTIMATION OF USUAL FOOD AND NUTRIENT INTAKE DISTRIBUTIONS IN THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

Kevin W. Dodd*, National Cancer Institute
Dennis W. Buckman, Information Management Services, Inc.
Raymond J. Carroll, Texas A&M University
Laurence S. Freedman, Gertner Institute for Epidemiology and Health Policy Research
Patricia M. Guenther, U.S. Department of Agriculture
Victor Kipnis, National Cancer Institute
Susan M. Krebs-Smith, National Cancer Institute
Douglas Midthune, National Cancer Institute
Amy F. Subar, National Cancer Institute
Janet A. Tooze, Wake Forest University School of Medicine

Statistical modeling methods to estimate usual intake, defined as the long-run average daily intake of a food or nutrient by an individual, are required to conduct dietary surveillance appropriately. Tooze, et al. (ADAJ, 2006) developed a nonlinear mixed model for usual intake in terms of 24HR measurements and supplementary covariates. This model is used to estimate quantiles of the population distributions of usual intake, accounting for the measurement error induced by within-person variability. The model also accommodates the presence of zeros and the skewed distributions typical of 24HR data. We cover the extension of the Tooze, et al. methodology to the survey sampling context, specifically using the National Health and Nutrition Examination Survey (NHANES). The NHANES survey is unique in that it combines interviews with physical examinations and laboratory studies, including a dietary assessment portion with up to two 24-hour recalls (24HR) per respondent. We incorporate sampling weights into the model fitting (using weighted pseudo-maximum likelihood) and in the Monte Carlo estimation of distributions. The balanced repeated replication method is used to approximate standard errors of model parameters and estimated usual intake quantiles.

email: doddk@mail.nih.gov

MODELING THE RELATIONSHIP BETWEEN USUAL FOOD INTAKE AND HEALTH PARAMETERS IN THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

Laurence S. Freedman*, Gertner Institute for Epidemiology and Health Policy Research
Douglas Midthune, National Cancer Institute
Dennis W. Buckman, Information Management Services, Inc.
Kevin W. Dodd, National Cancer Institute
Raymond J. Carroll, Texas A&M University
Janet A. Tooze, Wake Forest University School of Medicine
Patricia M. Guenther, U.S. Department of Agriculture
Susan M. Krebs-Smith, National Cancer Institute
Amy F. Subar, National Cancer Institute
Victor Kipnis, National Cancer Institute

Tooze et al (J Am Diet Assoc, 2006) describe a general statistical approach (NCI method) for analyzing intake of episodically consumed foods reported on two or more 24-hour recalls (24HRs) and show how it can be used to estimate the distribution of usual intake of such foods in the general population. We show how this method can be utilized to predict individual usual intake of such foods and to use the predicted value for evaluating diet and health outcome relationships. Following the regression calibration approach for correcting for the measurement error in 24HR reports, individual usual intake is predicted as its conditional mean given the 24HR-reported intake and other covariates, including the covariates in the health model. Applying the method to data from the Eating at America's Table Study, we show that precision of estimation can be improved considerably by including the report from a food frequency questionnaire as a covariate in the measurement error model. We demonstrate the method in evaluating the relationship between fish intake and blood mercury levels using data collected in the National Health and Nutrition Examination Survey. We also demonstrate in simulations that the method provides nearly unbiased estimates of exposure effects in this context.

email: lsf@actcom.co.il

30. MULTIPLE TESTING

GENERAL GATEKEEPING PROCEDURES WITH LOGICAL RESTRICTIONS

Alex Dmitrienko, Eli Lilly & Company
Lingyun Liu, Northwestern University
Ajit C. Tamhane*, Northwestern University

Recently Dmitrienko, Tamhane and Wiens (2007) have shown how stepwise procedures can be constructed to test ordered families of hypotheses subject to parallel or serial gatekeeping restrictions without the need for application of the closed testing principle and the consequent need to test all nonempty intersections of hypotheses. It is based on two new concepts, called the error rate function of a multiple test procedure and the separability condition, which implies that the Type I error rate for any rejected hypotheses can be carried forward to test hypotheses in the next ordered family. In this paper we extend this approach to deal with logical restrictions which are necessary when testability of any hypothesis is conditional on rejection of certain hypotheses in previous families. In particular, this includes tree-structured gatekeeping problems studied in Dmitrienko, Wiens, Tamhane and Wang (2007). The proposed procedure is illustrated by a clinical trial example.

email: ajit@iems.northwestern.edu

A UNIFIED APPROACH FOR CONSTRUCTING A CLOSED MULTIPLE TESTING PROCEDURE FOR A FIXED SEQUENCE OF FAMILIES WITH MULTIPLE NULL HYPOTHESES

Hanjoo Kim*, University of Pennsylvania School of Medicine
Richard A. Entsuah, Wyeth Research
Justine Shults, University of Pennsylvania School of Medicine

We present a multiple comparison procedure (MCP) for testing hierarchically ordered, fixed a priori, families of null hypotheses with the strong control of family-wise error rate at a pre-specified level of significance. In particular we focus on the application to clinical trials when the secondary family will not be examined unless some or all primary hypotheses exhibit significance. Under such hierarchical testing design, we show that the secondary family can be examined at the full significance level as long as at least one of the primary hypotheses exhibits significance at the pre-specified significance level using the closed testing principle which easily allows to compute adjusted pvalues for the proposed MCP. We illustrate our methodology using a clinical trial example.

email: hanjoo@mail.med.upenn.edu

EXTENSION OF PIEGORSCH AND CASELLA SIMULTANEOUS INTERVALS TO GENERALIZED LINEAR MODELS AND FUNCTIONS OF THEIR PARAMETERS

Amy E. Wagler*, Oklahoma State University
Melinda McCann, Oklahoma State University

Generalized linear models are utilized in a variety of statistical applications. Many times the estimated quantities from the models are of primary interest. These estimated quantities may include the mean response, odds ratio, relative risk, or attributable proportion. In these cases, overall conclusions about these quantities may be desirable. Most existing methods involving simultaneous estimation in the generalized linear model setting emphasize estimation of the expected response; few consider estimation of the regression parameters or functions of these parameters, such as the odds ratio or relative risk. Previously, Piegorsch and Casella (1988) proposed a restricted-Scheffé interval for the logistic regression model. These intervals provide a less conservative solution to the usual Scheffé intervals. In this project, the restricted-Scheffé intervals are also extended to any generalized linear model scenario. Additionally, restricted-Scheffé intervals are utilized to simultaneously estimate any set of the regression parameters. Simulation studies demonstrate that the restricted-Scheffé intervals provide a less conservative solution to the usual Scheffé intervals.

email: amy.wagler@okstate.edu

A BAYESIAN APPROACH TO LARGE SCALE SIMULTANEOUS INFERENCE

Bing Han*, RAND Corporation
Steven F. Arnold, Penn State University
Naomi Altman, Penn State University

We discuss Bayesian decision rules for highly multiple comparisons in the context of differential expression in microarray studies. Some of the advantages of our Bayesian approach in this context include: flexible modeling of gene expression, many options for decision rules for detecting differential expression while controlling either expectation or non-expectation error rate, insensitivity to weak dependencies in the data, and pooling of results across collections of models while retaining control of expected error rate. Due to the convenient Bayesian interpretation of hypothesis test, derivations for error rates can be greatly simplied by the concept of posterior probability of differential expression. We provide a MCMC-based procedure to estimate the posterior joint distribution of hypotheses and also apply a Poisson approximation under weak dependence in the posterior.

email: bhan@rand.org

CONSIDERING P-VALUE DEPENDENCE IN A STEPWISE MULTIPLICITY ADJUSTMENT METHOD

Richard E. Blakesley*, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh
Patricia R. Houck, University of Pittsburgh

Multiple hypothesis testing with correlated outcomes is a challenge. Resampling-based multiplicity adjustment methods have shown good properties, but implementation remains an obstacle. The powerful Bonferroni-derived methods control type I error too conservatively with increasing correlation between outcomes. Sidak-derived methods incorporate correlation measures, but with unstable type I error control. We propose a method that combines and refines elements of existing methods to control type I error while considering correlation. These elements are the Sidak functional form, the Hochberg stepwise component, and a refined adjustment component that uses a measure of p-value dependence. In a simulation study, we estimated the type I error and power rates of the proposed method and ten existing methods across many conditions, with the threshold alpha = 0.05. The proposed method exhibited type I error between [0.047, 0.057] with power rates similar to, or exceeding, all methods with conservative type I error performance. While not proven theoretically to control type I error, the proposed method has shown, through simulation, the desired properties of a multiplicity adjustment method.

email: reb18@pitt.edu

ON CONSONANCE OF CLOSED TESTING IN COMBINATION DRUG EFFICACY TRIALS

Julia Soulakova*, University of Nebraska

I discuss three multiple testing procedures for identifying the minimum efficacious doses in a balanced factorial combination drug trial. All of these procedures utilize a closed testing principle hence, strongly control the overall error rate and satisfy the coherence property. While coherence is an essential requirement for any multiple testing procedure, consonance is a highly desirable characteristic. A procedure is said to be consonant if whenever any null intersection-hypothesis is rejected, at least one of its components is also rejected. In our settings if a testing procedure is consonant then it always provides a set of all minimum efficacious combinations as a result, otherwise, it may lead to ambiguity in terms of the estimated set. I show that in the considered setting, whether (or not) a certain procedure satisfies the consonance property depends entirely on the nature of the test statistic. I also present the simulation probability of ambiguities which arise due to nonconsonance.

email: jsoulakova2@unlnotes.unl.edu

FAST FSR AND INFERENCE IN REGRESSION

Dennis D. Boos*, North Carolina State University
Leonard A. Stefanski, North Carolina State University

The False Selection Rate (FSR) method for selecting variables in regression is based on simulation of phony explanatory variables and monitoring their entry into a fitted model. Fast FSR is a recent advance that avoids the simulation step and allows resampling to be used for inference. Here we explore inference after selection in a variety of contexts.

email: boos@stat.ncsu.edu

31. BIOMARKERS

IDENTIFYING HIGH-DIMENSIONAL BIOMARKERS FOR PERSONALIZED MEDICINE VIA VARIABLE IMPORTANCE RANKING

Hojin Moon*, California State University-Long Beach
Songjoon Baek, National Center for Toxicological Research, U.S. Food and Drug Administration
Hongshik Ahn, Stony Brook University
Ralph L. Kodell, University of Arkansas for Medical Sciences
Chien-Ju Lin, National Center for Toxicological Research, U.S. Food and Drug Administration
James J. Chen, National Center for Toxicological Research, U.S. Food and Drug Administration

We apply robust classification algorithms to high-dimensional genomic data in order to find biomarkers, by analyzing variable importance, that enable a better diagnosis of disease, an earlier intervention, or a more effective assignment of therapies. The goal is to use variable importance ranking to isolate a set of important genes which can be used to classify life-threatening diseases with respect to prognosis or type in order to maximize efficacy or minimize toxicity in personalized treatment of such diseases. We present a ranking method and several other methods to select a set of important genes to use as genomic biomarkers, and we evaluate the performance of the selected genes in patient classification by external cross-validation. The various selection algorithms are applied to published high-dimensional genomic data sets using several well-known classification methods. We show that classification algorithms like ours are competitive with other selection methods for discovering genomic biomarkers underlying both adverse and efficacious outcomes for improving individualized treatment of patients for life-threatening diseases.

email: hmoon@csulb.edu

CONVEX HULL ENSEMBLE METHOD FOR CLASS PREDICTION WITH APPLICATION TO PERSONALIZED MEDICINE

Songjoon Baek*, National Center for Toxicological Research, U.S. Food and Drug Administration
Ralph L. Kodell, University of Arkansas
Hojin Moon, California State University-Long Beach
James J. Chen, National Center for Toxicological Research, U.S. Food and Drug Administration

Personalized medicine is defined by the use of genomic signatures of patients in a target population for assignment of more effective therapies as well as better diagnosis and earlier interventions that might prevent or delay disease. Classification algorithms are required to be highly accurate for optimal treatment on each patient. Typically, there are numerous genomic and clinical variables over a relatively small number of patients, which presents challenges for most traditional classification algorithms to avoid over-fitting the data. We developed a robust classification algorithm for high-dimensional data based on ensembles of two-dimensional convex hull classifiers. The proposed algorithm is applied to genomic data sets including leukemia data and the colon cancer data to distinguish disease subtypes for optimal treatment. The cross-validated classification accuracy, sensitivity, specificity, positive predictive value and negative predictive value are compared to corresponding values for several other classification procedures. The proposed classification method is expected to play a critical role in developing safer and more effective therapies that replace one-size-fits-all drugs with treatments that focus on specific patient needs.

email: SongJoon.Baek@fda.hhs.gov

DYNAMIC OPTIMAL STRATEGY FOR MONITORING DISEASE RECURRENCE

Hong Li*, Brown University

Surveillance to detect cancer recurrence is an important part of care for cancer survivors. In this paper we discuss the design of optimal strategies for early detection of disease recurrence based on each patient's distinct biomarker trajectory and periodically updated risk in the setting of a prospective cohort study. We adopt a latent class joint model which considers a longitudinal biomarker process and an event process jointly, to address heterogeneity of patients and disease, to discover distinct biomarker trajectory pattern, to classify patients into different risk groups, and to predict the risk of disease recurrence. The model is used to develop a monitoring strategy that dynamically modifies the monitoring intervals according to patients' current risks associated with periodically updated biomarker measurements and other indicators of disease spread. The optimal biomarker assessment time is derived using a utility function. In addition, we develop an algorithm to apply the proposed strategy to monitor any new diseased patients after initial treatment. We illustrate the models and the derivation of the optimal strategy using simulated data from monitoring for prostate cancer recurrence over a 5-year period.

email: hong_li@brown.edu

IDENTIFYING GENES THAT RESPOND TO ABIOTIC STRESS

Haiyan Wang*, Kansas State University

Plant response to abiotic stress (cold/freezing, salt, drought,wounding, heat) has been an area of intensive research for many years, but the molecular and cellular mechanisms of thermotolerance are not well understood. There is strong evidence of many genes and pathways in response to stresses yet to be discovered. In addition, membrane lipids play both important structural and signalling roles in stress responses and regulate lipid composition and fatty acid saturation levels to optimize these functions. Here I will illustrate a use of a recently developed nonparametric clustering analysis on gene expression data to aid in uncovering roles for membrane lipids in plant response to stresses and to identify in vivo functions of genes involved in lipid metabolism.

email: hwanggo@gmail.com

ENHANCED ENDPOINT ANALYSIS USING AUXILIARY INFORMATION IN CLINICAL TRIALS

Linda Sun*, Merck & Company
Cong Chen, Merck & Company

In clinical trials, data on the primary endpoint may be missing or subject to right censoring at the time of analysis. This will cause loss of information and consequently loss of efficiency in statistical analyses. For example, most of the patient survival data will be censored at the time of interim analyses in oncology survival trials. Can we increase the efficiency of survival analysis so that a better Go/No Go decision can be made at interim analyses? Sometimes, it is possible to recover some of this lost information by using auxiliary information. An auxiliary marker (e.g. time to disease progression in oncology trials) is a variable correlated with the true primary endpoint. Information on an auxiliary marker can usually be collected earlier and more complete than the true endpoint. This presentation will review several methods to enhance true endpoint analysis by incorporating auxiliary information. One particular method will be recommended with a focus in time-to-event analysis.

email: linda_sun@merck.com

A STUDY OF LOGIC REGRESSION WITH APPLICATION TO BLADDER CANCER

Bethany J. Wolf*, Medical University of South Carolina
Omar Moussa, Medical University of South Carolina
James Klein, Medical University of South Carolina
Dennis K. Watson, Medical University of South Carolina
Elizabeth H. Slate, Medical University of South Carolina

In the US, bladder cancer is the 5th most common cancer and diagnosis and treatment for patients is invasive and costly. Identifying biomarkers for early detection would significantly impact disease management and patients' outcomes. We address the problem of identifying which of multiple biomarkers and interactions among biomarkers best predict disease. A method for accommodating the complex interactions in biologic systems is logic regression, which finds Boolean combinations (AND, OR, NOT) of binary markers that best distinguish disease status. We performed a simulation study to quantify the ability of logic regression to correctly identify a known Boolean expression for varying levels of noise, sample size and complexity of the logic expression representing the data. We then applied logic regression to a study of bladder cancer biomarkers. Total mRNA extracted from urine sediment samples from 82 bladder cancer cases and 47 controls was used to make cDNA analyzed by PCR for the presence of each of five potential biomarkers. Molecular data were compared with cystoscopy as the standard for bladder cancer detection. Logic regression yielded two Boolean combinations of the markers having low misclassification rates. We place this result in the context of our simulation study.

email: wolfb@musc.edu

USE OF A PSEUDO MAXIMUM LIKELIHOOD ESTIMATOR TO SIMPLIFY COMPUTATIONS FOR A MULTIVARIATE LEFT-CENSORED LONGITUDINAL MODEL

Ghideon S. Ghebregiorgis*, University of Pittsburgh
Lisa Weissfeld, University of Pittsburgh

A mixed effects model based on a full likelihood is one of the few methods available to model longitudinal data subject to left-censoring. However, a full likelihood approach is complicated algebraically due to the large dimension of the numeric computations, and maximum likelihood estimation can be computationally prohibitive when the data are heavily censored. Moreover, for mixed models, the complexity of the computation increases as the dimension of the random effects in the model increases. We propose a pseudo likelihood method that simplifies the computational complexities, allows all possible multivariate models, and that can be used for any data structure including settings where the level of censoring is high. A robust variance-covariance estimator is used to adjust and correct the variance-covariance estimate. We perform a simulation study to evaluate and compare the performance of the proposed method for efficiency, simplicity and convergence with existing methods. The proposed methodology is illustrated in the analysis of Genetic and Inflammatory Markers for Sepsis study (GenIMS).

email: ghg2@pitt.edu

32. JOINT MODELS- SURVIVAL AND LONGITUDINAL

JOINT MODEL OF LONGITUDINAL PROCESS AND STATE-CHANGE PROCESS

Caitlin Ravichandran*, McLean Hospital and Harvard University
Victor DeGruttola, Harvard University

This work investigates joint modeling of a longitudinal process and a state-change process in the presence of missing data when the probability of missingness in the state-change process depends on the unobserved state. The investigation is motivated by examination of the association between changes in substance use and monitoring by resident father in adolescents using data from the National Longitudinal Surveys of Youth 1997. We classify substance use into the categories of no history of use, current use, and past use to describe the sequence of changes in substance use, or 'path', over time. Characterization of the path requires observed data at multiple observation times, but state is sometimes missing for a subset of observation times with probability that may depend upon substance use category. We model the mean for monitoring as a function of summary features of the path using a linear mixed effect model. Factorization of the joint likelihood into the distributions of the longitudinal response and the missingness process conditional on the path and the marginal distribution of the path allows all observed data to inform estimation in the presence of missing states. Application to the National Longitudinal Surveys data and a simulation study demonstrate the method's usefulness.

email: cravichandran@mclean.harvard.edu

JOINT MODELING OF LONGITUDINAL OUTCOMES AND EVENT TIME DATA IN A RHEUMATOID ARTHRITIS STUDY

Li Zhu*, University of California-Davis
Juan Li, Amgen Inc.
Eric Chi, Amgen Inc.

Often in clinical trials, certain variables of interest are measured repeatedly on each individual. Some of these individuals could withdraw from studies prematurely. These withdrawals may be informative (Rubin, 1976; Little and Rubin, 1987) if the dropout probability depends on the longitudinal outcome process. Ignoring this relationship may generate biased results (Tsiatis and Davidian, 2004) when comparing the rate of change of a continuous longitudinal response variable between treatment groups. Consequently, modeling the longitudinal outcomes and event time data jointly has gained a lot of attention in the literature recently (Hogan and Laird, 1997a). In this article, we propose an extension of the shared parameter model (Wu and Carroll, 1988; Schlucher, 1992; Follmann and Wu, 1995; Wulfsohn and Tsiatis, 1997; etc.), where each patient's repeated measurements are jointly modeled with two different informative dropout processes, and then use this method to analyze the data from TEMPO, a rheumatoid arthritis study.

email: lizhu@ucdavis.edu

FULLY EXPONENTIAL LAPLACE APPROXIMATIONS FOR THE JOINT MODELING OF SURVIVAL AND LONGITUDINAL DATA

Dimitris Rizopoulos*, Catholic University, Leuven-Belgium
Geert Verbeke, Catholic University, Leuven-Belgium
Emmanuel Lesaffre, Catholic University, Leuven-Belgium

A common objective in longitudinal studies is the joint modeling of a longitudinal response with a time-to-event outcome. Random effects are typically used in the joint modeling framework to explain the interrelationships between these two processes. However, estimation in the presence of random effects involves intractable integrals requiring therefore numerical integration. In this paper we propose a new computational approach for fitting such models based on the fully exponential Laplace method for integrals that makes the consideration of high dimensional random effects structures feasible. Contrary to the common Laplace approximation, our method requires much less repeated measurements per individual in order to produce reliable results. In addition and in order to facilitate the estimation of standard errors, we consider a flexible but parametric model for the cumulative baseline hazard function by expanding it into B-splines basis functions. The proposed model is exemplified in a study on 407 patients that underwent a primary renal transplantation and interest lies in using longitudinal haematocrit measurements as a prognostic factor for the time to graft failure.

email: dimitris.rizopoulos@med.kuleuven.be

JOINT ANALYSIS OF GENERALIZED LONGITUDINAL MEASUREMENTS AND SURVIVAL DATA WITH MULTIPLE FAILURE TYPES

Ning Li*, University of California Los Angeles
Robert Elashoff, University of California Los Angeles
Gang li, University of California Los Angeles

Existing joint models for longitudinal and survival data adopt a linear mixed effects model for the longitudinal measurements, which is not applicable for discrete longitudinal outcomes such as count data. In a study conducted at UCLA School of Medicine, longitudinal measurements of the number of relapses are observed in relapsing remitting multiple sclerosis subjects. Missing data caused by dropout or treatment failure can be non-ignorable since patients with a higher relapse rate would be more likely to experience treatment failure or dropout. We propose a joint model in which the generalized linear mixed model is linked to the event times by the association between the random effects for the two endpoints. This general model is able to handle not only the count data, but other discrete longitudinal outcomes such as binary and negative binomial random variables. In addition, our model permits multiple failure types for the event times, offering a framework to model informatively censored events as a competing risk. The proposed joint model is illustrated by simulations and the aforementioned study for relapsing remitting multiple sclerosis.

email: ningli@ucla.edu

PREDICTION OF EVENT RISK USING JOINT MODELS FOR LONGITUDINAL MEASUREMENTS AND EVENT TIMES

Nicholas J. Salkowski*, University of Minnesota
Melanie M. Wall, University of Minnesota

Clinical trials often generate both longitudinal and event time data. Joint models that simultaneously consider the longitudinal measurements and the event time often introduce latent variables that account for the associations between the multivariate measurements. One potential application of this sort of model is to estimate the risk of a future event for a new subject using a series of longitudinal measurements, if the event has not yet occurred. Data from the Valsartan Heart Failure Trial (Val-HeFT) are used to illustrate difficulties arising from attempting to predict survival using longitudinally measured quality of life. In Val-HeFT, participants who died typically have fewer quality of life measurements than censored participants due to good follow-up. When participants have few quality of life measurements, the survival time strongly influences the estimates of the latent variables and the estimated true quality of life trajectory. Participants with similar quality of life data may have dramatically different estimated quality of life trajectories when survival times differ. Thus despite observing a strong relationship between longitudinal data and event times for the fitted data, it is demonstrated that future predictions of event times based solely on longitudinal data can be rather poor.

email: salk0008@umn.edu

SIMULTANEOUS MODELING OF LONGITUDINAL DATA AND DROPOUT PROCESS IN CLINICAL STUDIES

Qinfang Xiang*, Endo Pharmaceuticals

It is not unusual in practice for some sequence of measurements to terminate prematurely due to dropout in a longitudinal clinical trial. Although several statistical packages are available for analyzing the resulting unbalanced longitudinal data, they only yield valid inferences under specific assumptions with respect to the dropout process. To obtain valid inferences, it is necessary to accommodate dropout in the modeling process. Several approaches have been proposed to deal with non-ignorable dropout including selection models, pattern mixture models and joint models. All three models account for the dropout process and simultaneously model the measurement and dropout processes. The objective of this research work is to demonstrate and compare these three modeling strategies using the data from a clinical trial and provide some practical guidance from several points of view, i.e., the difficulty in the model parameter estimation, the interpretation of model results, the sensitivity of the model assumptions, and the availability of the computational utilities.

email: xiang.qinfang@endo.com

BAYEISAN JOINT ANALYSIS OF LONGITUDINAL MEASUREMENTS AND COMPETING RISKS FAILURE TIME DATA VIA MODELING OF MULTIVARIATE RANDOM EFFECTS COVARIANCES

Xin Huang*, University of California, Los Angeles
Gang Li, University of California, Los Angeles
Robert M. Elashoff, University of California, Los Angeles
Jianxin Pan, University of Manchester-UK

Most existing methods for joint modeling of longitudinal measurements and survival data assume that random effects in longitudinal models are very low-dimensional or simply univariate and the random effects covariance matrix is typically constant across subject. However, it may not be always true in practice. In this paper we propose a unified approach to jointly model the repeated measurements and competing risks survival data by modeling covariance structures of random effects using regression models. The models considered include a linear mixed effects sub-model for the longitudinal outcome and proportional cause-specific hazards frailty sub-models for the competing risks survival data, which are linked together through some latent random effects. The proposed approach allows for high dimensional random effects and heterogeneous covariance matrices of the multivariate random effects, and the resulting estimated covariance matrices are guaranteed to be positive definite. Bayesian analysis is made for statistical inferences of the joint models. Real data analysis and simulation studies are made to illustrate the usefulness of the proposed approach.

email: xinhuang@ucla.edu

33. DATA MINING AND MACHINE LEARNING

CARRYING PREDICTION MODELS ACROSS MICROARRAY DATA SETS GENERATED BY DIFFERENT LABS AND DIFFERENT PLATFORMS

Chunrong Cheng*, University of Pittsburgh
George C. Tseng, University of Pittsburgh

Reproducibility of microarray experiment has been greatly improved in the past decade and its application in biomedical research is more and more prevalent. Multiple literatures investigating an identical disease are found with different array platform and implemented in different labs. Similar high disease prediction accuracies are often reported in these studies; however, applying a prediction model established in one study to the other usually generates poor performance. We investigated the application of gene-wise normalization following the commonly practiced global sample-wise normalization. The proposed gene-wise normalization often dramatically increases the prediction accuracies in the cross-dataset prediction. We further propose a bootstrapping and an alternative analytical method to adjust for differential sample ratios of disease groups that may affect the performance of gene-wise normalization. Simulation result and application to three lung cancer data sets show significant and robust improvement of our method. A simple calibration scheme is developed to apply our method to future clinical trials. The number of calibration samples needed is estimated from existing studies and suggested for application to future studies.

email: crcwxd@yahoo.com

VARIABLE SELECTION IN PENALIZED MODEL-BASED CLUSTERING VIA REGULARIZATION ON GROUPED PARAMETERS

Benhuai Xie*, University of Minnesota
Wei Pan, University of Minnesota
Xiaotong Shen, University of Minnesota

Penalized model-based clustering has been proposed for high-dimensional but small sample-sized data, such as arising from genomic studies; in particular, it can be used for variable selection. A new regularization scheme is proposed to group together multiple parameters of the same variable across clusters, which is shown both analytically and numerically to be more effective than the conventional L_1 penalty for variable selection. In addition, we develop a strategy to combine this grouping scheme with grouping structured variables. Simulation studies and applications to microarray gene expression data for cancer subtype discovery demonstrate the advantage of the new proposal over several existing approaches.

email: bhxie@umn.edu

PRINCIPAL COMPONENT ANALYSIS FOR INTERVAL-VALUED SYMBOLIC DATA

Jennifer Le-Rademacher*, University of Georgia

In this presentation, we propose a new approach to principal component analysis (PCA) for interval-valued data. Interval-valued data fall into the class of symbolic data which was first introduced by Edwin Diday in 1987. Unlike an observation in classical data which is represented by a single point in a p-dimensional space, an observation in interval-valued data is represented by a hypercube with 2^p vertices. The two common adaptations of PCA for interval-valued data are the centers method and the vertices method. The centers method computes the principal components using the centers of the original intervals. Although the order of computation using this method is low, it ignores the internal variations within each observation. The vertices method computes the principal components using all vertices, treating each of them as an independent observation. An advantage of this method is its accounting for internal variations. A drawback of this method is the dependency among the vertices of the same observation is ignored. We propose a new method using a so-called symbolic variance-covariance structure. This method avoids the drawbacks associated with the other two PCA methods. The proposed method accounts for the internal variation of interval-valued variables and the dependency among vertices of the same observation. We will illustrate with an example.

email: jle1@uga.edu

LASSO-PATTERNSEARCH ALGORITHM

Weiliang Shi*, University of Wisconsin-Madison
Grace Wahba, University of Wisconsin-Madison
Stephen Wright, University of Wisconsin-Madison
Kristine Lee, University of Wisconsin-Madison
Ronald Klein, University of Wisconsin-Madison
Barbara Klein, University of Wisconsin-Madison

The LASSO-Patternsearch is proposed, as a two-stage procedure to identify clusters of multiple risk factors for outcomes of interest in large demographic studies, when the predictor variables are dichotomous or take on values in a small finite set. Many diseases are suspected of having multiple interacting risk factors acting in concert, and it is of much interest to uncover higher order interactions when they exist. In the first stage a LASSO is used to select a subset of patterns from a possibly extremely large set of potential patterns of various orders via a pattern selection criteria. Then the patterns selected by the LASSO are further selected in the framework of (parametric) generalized linear models. Notably, the patterns are those that arise naturally from the log linear expansion of the multivariate Bernoulli density. A novel tuning procedure is proposed and a novel computational algorithm for the LASSO step is developed to handle a large number of unknowns in the problem. The method is applied to data from the Beaver Dam Eye Study and is shown to expose physiologically interesting interacting risk factors.

email: shiw@stat.wisc.edu

IDENTIFYING REPRESENATIVE TREES IN RANDOM FOREST FOR SURVIVAL DATA

Mousumi Banerjee, University of Michigan
Ying Ding*, University of Michigan
Anne-Michelle Noone, University of Michigan

Tree-based methods have become popular for analyzing right censored survival data where the primary goal is prognostic grouping of patients. Ensemble techniques such as random forest improve the accuracy in prediction and address the instability in a single tree by growing an ensemble of trees and aggregating. However, individual trees are lost in the forest. In this paper, we propose a methodology for selecting the most representative trees in a forest for survival data, based on three tree similarity metrics. For any two trees, the metrics are chosen to (1) measure similarity of the covariates used to split the trees; (2) reflect similar clustering of patients in the terminal nodes of the trees; and (3) measure similarity in predictions from the two trees. While the latter focuses on prediction, the first two metrics focus on the architectural similarity between two trees. The most representative trees in the forest are chosen based on the average similarity score assigned to each tree corresponding to each of the three metrics. Out of bag estimates of error are computed for the most representative trees using a neighborhood of similar trees. Finally the methods are illustrated using data from a cohort study of breast cancer patients to model recurrence free survival.

email: yingding@umich.edu

DATA MINING FOR MEDICAL RESEARCH: IDENTIFYING RELATIONSHIPS THAT CANNOT BE IDENTIFIED IN ANY OTHER WAY

Shenghan Lai*, Johns Hopkins University Medical School

Dr. Lai will discuss his use of data mining techniques to identify relationships followed by his use of standard statistical techniques to prove that such relationships are indeed true. Dr. Lai will use several examples including his research related to the Association between Vitamin E and the development of myocardial infarction and the association between regional heart function and coronary calcification.

email: lisasolomon@yahoo.com

SPARSE DISTANCE WEIGHTED DISCRIMINATION

Lingsong Zhang*, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

In the High Dimension Low Sample Size situation, Marron et al (2007) proposed a new classification method, Distance Weighted Discrimination (DWD), which is similar to Support Vector Machine (SVM) when the number of objects is larger than the number of features, but perform better than SVM on high dimension low sample situation. However, in the high dimensional case, the noise in the data may still dominate in finding the separating hyperplane. In this paper, we proposed a Sparse DWD (SDWD) method, which incorporates the variable selection along with the classification. Theoretical properties are explored under some special conditions. Applications to proteomics and genetic pathways are used to illustrate the SDWD method.

email: zhang@hsph.harvard.edu

34. GENETIC EPIDEMIOLOGY-ASSOCIATIONS

ESTIMATING ODDS RATIOS IN GENOME SCANS: AN APPROXIMATE CONDITIONAL LIKELIHOOD APPROACH

Arpita Ghosh*, 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

In genome-wide studies we are interested in identifying genetic variants and once such a variant is detected, we proceed to quantify the genetic effect of that variant on the phenotype of interest based on the same data. In modern whole genome scans, use of stringent thresholds for significance testing to control the genome-wide error distorts the estimation process and produces estimates of effect sizes which may be, on average, far greater in magnitude than the true effect sizes. We refer to this phenomenon as significance bias. In this paper we introduce a method, based on the odds ratio estimate and its standard

error as reported by standard statistical software, to correct for the significance bias in case-control association studies. We compare our approach to recently proposed methods and find that it performs well, and is far easier to implement. We also develop a rigorous method for constructing confidence intervals for the odds ratio. We evaluate the performance of our approach via extensive simulations for a range of genetic models, minor allele frequencies and odds ratio values. Compared to the naïve estimation procedure, our approach reduces the bias and the mean-squared error, especially for modest effect sizes.

email: aghosh@bios.unc.edu

RANGES OF MEASURES OF ASSOCIATION FOR PEDIGREE BINARY VARIABLES

Yihao Deng*, Indiana University-Purdue University-Fort Wayne
N. Rao Chaganty, Old Dominion University

Analysis of pedigree binary data plays an important role in genetic studies, linkage analysis, and epidemiologic research. Several measures are commonly employed to study the association between the pedigree binary variables. These measures include correlations, odds ratios, kappa statistics and relative risks. In this talk we discuss permissible ranges of these measures of association. Understanding these ranges is a first step to develop efficient parameter estimation methods for statistical models for real life pedigree binary data.

email: dengy@ipfw.edu

SCORE STATISTICS FOR FAMILY-BASED ASSOCIATION MAPPING OF QUANTITATIVE TRAITS

Samsiddhi Bhattacharjee*, University of Pittsburgh
Eleanor Feingold, University of Pittsburgh

Most family based tests of association for quantitative traits are extensions of the Transmission Disequilibrium Test (TDT). They condition upon parental genotypes to protect against population stratification whereas parental phenotypes are generally ignored. Both of these factors contribute to loss of power of these tests relative to population-based or unconditional family-based tests. To improve power these tests should use all the available data including parental phenotypes. They should also extract information from parental genotype-phenotype correlation. We derive novel likelihood-based score statistics which have optimal power to detect association in families, while protecting against population sub-structure as well as phenotype-based ascertainment. We also discuss extensions of these statistics for handling non-normally distributed traits. Finally, we compare the performance of the proposed statistics relative to some of the standard family-based tests of association.

email: samsiddhi@hgen.pitt.edu

A SEMIPARAMETRIC ASSOCIATION TEST IN STRUCTURED POPULATIONS

Meijuan Li*, University of Minnesota
Tim Hanson, University of Minnesota

Although Population-based Linkage-Disequilibrium (LD) mapping may be subject to bias caused by population stratification, alternative methods that are robust to population stratification such as family-based association designs may have lower mapping resolution. Recently, various statistical methods robust to population stratification were proposed for association studies, using unrelated individuals to identify associations between candidate genes and traits of interest. Here, we propose a semiparametric test for association (SPSA) in structured populations. SPSA controls for population stratification by first deriving genetic background variables using multidimensional scaling analyses (MDS) on the genetic distance matrix of the sampled individuals derived from a set of independent genetic markers, and then modeling the relationship between trait values, genotypic scores at the candidate gene, and genetic background variables through a semiparametric model. We model the error distribution in the association test as a mixture of absolutely continuous Polya trees constrained to have median 0 centered around a normal family of distributions. We apply the proposed method to a previously published public available data set of association mapping in 95 Arabidopsis thaliana varieties. We also conduct a simulation study to demonstrate the power of the method.

email: meijuanl@biostat.umn.edu

A BAYESIAN SEMIPARAMETRIC FRAMEWORK FOR GENETIC ASSOCIATION STUDIES IN THE PRESENCE OF POPULATION STRUCTURE

Nicholas M. Pajewski*, Medical College of Wisconsin
Purushottam W. Laud, Medical College of Wisconsin

There has been considerable discussion amongst genetic researchers about the impact of population structure on association studies. When a sample of unrelated individuals consists of subpopulations differing in disease risk and allele frequencies, estimates of disease association with a particular locus can be exaggerated or attenuated. Numerous approaches have been proposed to address this. Foremost amongst these is the Bayesian model-based clustering approach of Pritchard and coauthors (2000,2001), developed within the context of candidate gene studies. They proposed modeling the subpopulations, classifying individuals accordingly, and essentially pooling to produce stratified inference. As the focus of research has moved towards genome-wide association studies, recent emphasis is on computationally efficient methods (Epstein et.al. 2007; Kimmel et.al. 2007). We consider a unified Bayesian semiparametric framework for association studies of quantitative traits and case-control designs. The model appropriately integrates population structure in making association inference. It also uses a nonparametric sparsity prior to incorporate the prior belief that most loci are not phenotypically associated (Dunson et.al. 2007). Effectiveness of the proposed model is demonstrated via a simulated study of a quantitative trait based on the HapMap data (www.hapmap.org).

email: npajewsk@mcw.edu

A JOINT TEST FOR HAPLOTYPE-BASED ASSOCIATION IN CASE-CONTROL STUDIES

Tao Wang*, Medical College of Wisconsin
Howard Jacob, Medical College of Wisconsin
Soumitra Ghosh, Medical College of Wisconsin
Xujing Wang, Medical College of Wisconsin
Zhao-Bang Zeng, North Carolina State University

With a dense set of genetic markers such as single nucleotide polymorphisms (SNPs) for genetic association studies within a candidate locus, a haplotype-based approach is attractive in reducing the data complexity and increasing the statistical power of detecting genetic risk

factors. However, a comprehensive haplotype-based association test for a set of markers requires consideration of all possible combination of the markers, which often leads to severe multiple testing problems. In this study, we propose a latent variable approach for haplotype-based association testing in case-control studies. First, a logistic mixture regression model is introduced with genetic effects contributed by a putative disease susceptible locus (DSL). Next, using a retrospective likelihood to adjust for the case-control sampling ascertainment and hold the Hardy-Weinberg equilibrium constraint, we develop an EM-based algorithm to fit the model and estimate the joint haplotype frequencies of the DSL and markers simultaneously. With the latent DSL being a potential genetic risk factor that may consist of an allele, a haplotype or sub-haplotype of the markers, or a putative locus between the markers, the likelihood ratio statistic can then provide a joint test for haplotype-based association between a set of markers and a disease phenotype.

email: taowang@mcw.edu

EFFICIENT MULTIVARIATE POLYGENIC AND ASSOCIATION ANALYSIS

Wei-Min Chen*, University of Virginia

Family-based genomewide association (GWA) studies have the advantage of using existing linkage scan data and have already been successfully applied to map genetic variants of complex traits. Reliable polygenic analysis of phenotype data is required prior to performing a score test, an efficient procedure allowing millions of SNPs to be scanned in family-based GWA. Further, multivariate polygenic analysis is crucial for mapping common genetic variants responsible for multiple correlated traits. I implemented a computationally efficient score method to perform likelihood maximization for multivariate polygenic analysis. Redundancy in the modeling of multiple traits is dramatically reduced by pre-calculating the derivatives and inverses of the phenotype variance-covariance matrix. Compared to a standard implementation in SOLAR, my proposed method yielded identical parameter estimates with much less computation time. For example, the time for a bivariate analysis was reduced from 7 minutes to 3 seconds in one dataset with 198 phenotyped individuals in the largest pedigree, and from 45 minutes to 79 seconds in another dataset with 625 phenotyped individuals in the largest pedigree. Applications of this work are discussed.

email: wmchen@virginia.edu

35. GENERALIZED LINEAR MODELS

PSEUDO-SCORE INFERENCE FOR PARAMETERS IN DISCRETE STATISTICAL MODELS

Alan Agresti*, University of Florida
Euijung Ryu, Mayo Clinic

We discuss ‘pseudo-score’ inference for parameters in models for discrete data that uses the Pearson statistic comparing fitted values under null and alternative cases. The inferences are alternatives to likelihood-ratio tests and profile likelihood confidence intervals. Much research has shown that score-test-based inference methods are often better than likelihood-ratio-test-based methods, in terms of achieving error rates closer to the nominal level. The pseudo-score method for multinomial response models simplifies to the score method for some inferences and otherwise is a simple approximation for that method. In cases in which ordinary score methods are impractical, such as when the likelihood function cannot be explicitly expressed in terms of the model parameters, the pseudo-score method is still simple to implement. A quasi-likelihood implication of the method is that high-quality inference is possible for comparing two models based solely on the fitted values for the models and the variance of the observations under the simpler model. This suggests a generalization of the method to cases in which the likelihood function is not fully specified, such as an alternative to the GEE method for marginal modeling of clustered, correlated categorical responses that does not use Wald methods or require using the sandwich covariance estimator.

email: aa@stat.ufl.edu

SIMULTANEOUS CONFIDENCE INTERVALS FOR COMPARING BINOMIAL PARAMETERS

Alan Agresti, University of Florida
Matilde Bini, University of Florence-Italy
Bruno Bertaccini, University of Florence-Italy
Euijung Ryu*, Mayo Clinic

To compare proportions with several independent binomial samples, we recommend a method of constructing simultaneous confidence intervals that uses the studentized range distribution with a score statistic. It applied with a variety of measures, including the difference of proportions, odds ratio, and relative risk. For the odds ratio, a simultaneous study suggests that the method has coverage probability closer to the nominal value than ad hoc approaches such as the Bonferroni implementation of Wald or exact small-sample pairwise intervals. It performs well even for the problematic but practically common case in which the binomial parameters are relatively small. For the difference of proportions, the proposed method has performance comparable to a method proposed by Piegorsch (1991).

email: eryu@stat.ufl.edu

BAYESIAN GENERALIZED LINEAR MODELS FOR GENOMEWIDE QTL ANALYSIS

Nengjun Yi*, University of Alabama at Birmingham
Samprit Banerjee, University of Alabama at Birmingham

Mapping quantitative trait loci (QTL) is to identify molecular markers or genomic loci that influence the variation of complex traits. The problem is complicated by the facts that QTL data usually contain a large number of highly correlated markers across the entire genome and most of them have little or no effect on the phenotype. In this paper, we propose a unified Bayesian generalized linear models framework for mapping multiple QTL for various types of complex traits (e.g., continuous, binary, ordinal, Poisson traits). The proposed models use prior distributions for the genetic effects that are scale mixtures of normal distributions with mean zero and variances distributed to give each effect a high probability of being near zero. We consider two types of priors for the variances, exponential and scaled inverse-distributions. We treat all hyperparameters as unknowns and estimate them along with other parameters. We fit generalized linear models in R with the proposed prior distributions by incorporating an approximate EM algorithm into the usual iteratively weighted least squares. The methods are illustrated using several real and simulated data sets.

email: nyi@ms.soph.uab.edu

REGRESSION SPLINES FOR THRESHOLD SELECTION WITH APPLICATION TO A RANDOM EFFECTS LOGISTIC DOSE-RESPONSE MODEL

Daniel L. Hunt*, St. Jude Children's Research Hospital
Chin-Shang Li, University of California-Davis

A flexible parametric method is proposed for smoothing the dose effect in a general threshold dose-response model with random effects. This method uses a linear combination of linear B-splines in which an interior knot, regarded as a free parameter for a piecewise linear spline, indicates a break point in the dose-response curve. The knot will be estimated along with the model parameters. The proposed method is applied to data from developmental toxicity studies.

email: daniel.hunt@stjude.org

DIAGNOSIS OF MODEL MISSPECIFICATION FOR RANDOM EFFECTS IN GENERALIZED LINEAR MIXED MODELS

Xianzheng Huang*, University of South Carolina

Generalized linear mixed models (GLMMs) are widely used in the analysis of clustered data. However, the validity of likelihood-based inference in such analyses can be greatly affected by the assumed model for the random effects. We propose a diagnostic method for random-effects model misspecification in GLMMs for clustered binary response data. We provide a theoretical justification of the proposed method and investigate its finite sample performance via simulation. The proposed method is applied to data from a longitudinal respiratory infection study.

email: huang@stat.sc.edu

ESTIMATION OF TREATMENT EFFECTS BY COMBINING TWO OR MORE DESIGN PLANS

Yvonne M. Zubovic*, Indiana University Purdue University-Fort Wayne
Chand K. Chauhan, Indiana University Purdue University-Fort Wayne

Consider an experiment in which observations are taken on N subjects to investigate potential differences between c different treatments. Suppose that all c treatments are administered to n of the subjects. If these were the only subjects under study, the resulting data set could be analyzed using the procedures for a randomized complete block design. However, in this experiment the remaining N-n subjects are administered exactly k of the c treatments, where k < c, following the constraints of a balanced incomplete block design. In this paper we consider a method to estimate the different treatment effects by combining estimates from the complete block and balanced incomplete block designs. Properties and the applications of these combined estimators are investigated.

email: zubovic@ipfw.edu

APPROACHES FOR ESTIMATING AND TESTING CONDITIONAL CORRELATIONS

Xueya Cai*, SUNY-Buffalo
Gregory, E. Wilding, SUNY-Buffalo
Alan Hutson, SUNY-Buffalo

Current literature on testing the equality of correlations between two random variables across another variable is limited to the case where the covariate is categorical. In this study, we propose two approaches to performing such hypothesis testing across a numeric covariate under different model assumptions. The first approach assumes bivariate normal distribution of the data, and the correlation coefficient is modeled as a function of the covariate. The restricted log likelihood based on a generalized linear model is maximized and parameter estimates obtained. In the second approach, the bivariate normal assumption is relaxed. Standardized residuals obtained from classical least squares analyses are used to estimate correlation coefficients, which are then tested in the generalized linear model. Both approaches are extended to accommodate heteroscedastic observations, in which conditional variances of the two random variables are either modeled as functions of the covariate, or estimated via the least squares method. Hypothesis testing of equal correlation in both approaches are equivalent to testing whether the coefficient in the correlation function is zero. Monte Carlo simulations show that the proposed methods have higher power and smaller bias than the existing methods.

email: xueyacai@buffalo.edu

36. STATISTICAL ISSUES IN ANALYSIS OF GENOMICS DATA

OVERCOMING ADVERSE EFFECTS OF CORRELATIONS IN MICROARRAY DATA ANALYSIS

Lev Klebanov, Charles University-Czech Republic
Andrei Yakovlev*, University of Rochester

The currently practiced methods of significance testing in microarray gene expression profiling are highly unstable and their power tends to be very low. These undesirable properties are due to the nature of multiple testing procedures, as well as extremely strong and long-ranged correlations between gene expression levels. Resorting to normalization procedures does not provide a satisfactory solution to the problem because of their distorting effects on the true expression signals. Such effects are especially pronounced in large sample studies where control of type 1 errors may be entirely lost. We have identified a special structure in gene expression data that produces a sequence of weakly dependent random variables. This structure, termed the delta-sequence, lies at the heart of a new methodology for selecting differentially expressed genes in non-overlapping gene pairs. The proposed method has two distinct advantages: (1) it leads to dramatic gains in terms of the mean numbers of true and false discoveries, as well as in stability of the results of testing; (2) its outcomes are entirely free from the log-additive array-specific technical noise. We demonstrate the usefulness of this approach in conjunction with the nonparametric empirical Bayes methodology.

email: andrei_yakovlev@urmc.rochester.edu

A GENE EXPRESSION BARCODE FOR MICROARRAY DATA

Rafael A. Irizarry*, Johns Hopkins University
Michael J. Zilliox, Johns Hopkins University

The ability to measure genome-wide gene expression holds great promise for characterizing cells and distinguishing diseased from normal tissues. Thus far, microarray technology has only been useful for measuring relative expression between two or more samples, which has handicapped the ability of microarrays to classify tissue types. This paper presents the first method that can successfully predict tissue type based on data from a single microarray hybridization. We achieved this by developing a statistical procedure that is able to accurately demarcate expressed from unexpressed genes and therefore defines a unique gene expression barcode for each tissue type. The utility of the method is demonstrated by defining a barcode-based classification algorithm with better predictive power than the best existing algorithms. Hundreds of publicly available human and mouse arrays were used to define and assess the performance of the barcode. With clinical data, we find near perfect predictability of normal from diseased tissue for three cancer studies and one Alzheimer's disease study. The barcode method also discovers new tumor subsets in previously published breast cancer studies that can be used for the prognosis of tumor recurrence and survival time. A preliminary web-tool, that when given a raw data file predicts tissue type, is available at http://rafalab.jhsph.edu/barcode/.

email: rafa@jhu.edu

A SYSTEMATIC FRAMEWORK OF WHOLE-GENOME SNP SCREENING SEEKING A PREDICTIVE GENOMIC BIOMARKER FOR POTENTIAL TREATMENT INDIVIDUALIZATION

Sue-Jane Wang*, U.S. Food and Drug Administration

With the completion of the whole genome scan project and the high failure rate observed in late phase drug development program, there has been a rising interest to incorporate patients' genomic (composite) biomarker for assessing the therapeutic risk/benefit. Consider a genomic biomarker based randomized controlled trial. The genuine intent is the ability of the biomarker to predict sensitive patients for individualization of medical treatment. Microarray gene expression data have been used as a vehicle to explore patient subpopulation that may have enhanced response to treatment. It is envisaged that the whole-genome single nucleotide polymorphism (SNP) genotyping may be more akin to the diagnostic utility of therapeutic response (efficacy and/or safety). In this presentation, the utility and challenges for a systematic framework to seek a genomic biomarker using the whole genome SNP scan approach will be presented. The accompanied statistical issues, such as ascertainment bias and statistical uncertain, in nested case-control association studies and randomized clinical trials will be discussed and elucidated with typical study examples. We consider a valid adaptive design approach for inference of biomarker's clinical utility in which biomarker classification is performed via a companion diagnostic.

email: suejane.wang@fda.hhs.gov

37. RECENT DEVELOPMENTS IN NON-SMOOTH ESTIMATING FUNCTIONS FOR CENSORED DATA

ESTIMATION IN THE SEMIPARAMETRIC ACCELERATED FAILURE TIME MODEL WITH MISSING DATA

Bin Nan, University of Michigan
John D. Kalbfleisch*, University of Michigan
Menggang Yu, Indiana University-School of Medicine

We consider a class of doubly weighted rank based estimating equations for the accelerated failure time model with missing data as arises, for example, in case-cohort studies. The weights in the equation are allowed to be non predictable in a stochastic process formulation. We outline proofs of asymptotic properties for the weighted estimators using the empirical process theory where the martingale theory may fail. Simulations show that the outcome-dependent weighted method works well and improves efficiency compared to methods with predictable weights. Different weighting schemes are considered and their usefulness explored across a variety of error models.

email: jdkalbfl@umich.edu

ACCELERATED RECURRENCE TIME MODELS

Yijian Huang*, Rollins School of Public Health-Emory University
Limin Peng, Rollins School of Public Health-Emory University

For the analysis with recurrent events, we propose a generalization of the accelerated failure time model to allow for evolving covariate effects. These so-called accelerated recurrence time models postulate that time to expected recurrence frequency, upon transformation, is a linear function of covariates with frequency-dependent coefficients. This modeling strategy shares the same spirit as quantile regression. An estimation and inference procedure is developed by generalizing the celebrated Powell's (1984, 1986) estimator for censored quantile regression. Consistency and asymptotic normality of the proposed estimator are established. An algorithm is devised to attain good computational efficiency. Simulations demonstrate that this proposal performs well under practical settings. This methodology is illustrated in an application to the well-known bladder cancer study.

email: yhuang5@emory.edu

ON COMBINING MULTIPLE ESTIMATING EQUATIONS

Lu Tian*, Northwestern University

In semiparametric setting, inferences about the unknown parameters may be made based on nonsmooth estimating functions. It is often difficult to numerically solve the resulting estimating equations. Furthermore, the corresponding statistical inference may involve difficult and subjective nonparametric smoothing step. In this talk, we will propose an efficient method to combine multiple non-smooth estimating equations when there is a 'good' initial estimator. The combined estimator will be asymptotically more efficient than any of those based on individual estimating equations and the associated asymptotical distribution can be easily estimated. We will use accelerated failure time and quantile regression models as illustrative examples.

e-mail: lutian@northwestern.edu

VARIANCE ESTIMATION IN CENSORED LINEAR REGRESSION

Zhezhen Jin*, Columbia University

In semiparametric censored linear regression, estimating functions for regression parameters are often nonsmooth and non-monotone, and variance estimates are difficult. We discuss the issues and present recently developed new approaches for the variance estimation. We show general theory, impementation, simulation studies and demonstrate the methods with examples.

e-mail: zj7@columbia.edu

38. BAYESIAN METHODS IN EPIDEMIOLOGY

NOVEL BAYESIAN APPROACHES TO MODEL-ROBUST INFERENCE

Kenneth M. Rice*, University of Washington
Adam Szpiro, University of Washington
Thomas Lumley, University of Washington

Epidemiological studies typically involve data on many hundreds of individuals; sample sizes of several thousand are not uncommon. The datasets involved are therefore large, but although this makes accurate estimation feasible, standard parametric models almost always fit these data poorly, detracting from the prima facie validity of model-based inference. A modern frequentist approach to this problem is to dispense altogether with the requirement of a correctly-specified parametric model: by stating, non-parametrically, what we want to know about the data-generating mechanism, estimating equations and associated 'robust' or 'sandwich'-based intervals provide accurate large-sample inference, at no more computational effort than fitting a GLM. The standard Bayesian analog to this model-light approach ('Non-Parametric Bayes') takes a rather different approach, specifying a fixed, highly-parameterized model: in practice this requires cutting-edge computational skills. We investigate alternative Bayesian approaches to model-robust inference, taking a decision-theoretic approach. We show that simple Bayesian methods can provide intervals with the same model-robustness as the frequentist large-sample approach, and go on to show where the addition of prior information may additionally provide better small-sample properties.

email: kenrice@u.washington.edu

BAYESIAN MODELING OF COMPLEX TRAITS

Paola Sebastiani*, Boston University

Discovering the genetic basis of common diseases is one of the major challenges of biomedical sciences that has been limited in the past to candidate gene studies. Nowadays, advances in high throughput technology can provide genetic information of almost the entire genome and open the opportunity to discover the genetic basis of complex disease. However, a challenge of this discovery process is building complex models with many variables from massive amount of data. This talk will describe some of the issues related to modeling complex traits, the new challenges posed by the analysis of genome-wide data, and the feasibility of network modeling. In particular, we will present a hierarchical and modular approach to screen genome wide genotype data that incorporates quality control filters, linkage disequilibrium, physical distance and gene ontology to build prognostic models of complex traits. We will use examples of genetic dissection of complex traits using data from a cohort study of the complications of sickle cell anemia, and data from a cohort study of exceptional longevity.

email: sebas@bu.edu

A BAYESIAN APPROACH FOR ESTIMATING THE EFFECT OF EXPOSURE TO AIR POLLUTION ON CARDIOVASCULAR DISEASE RISK FACTORS

Trivellore E. Raghunathan*, University of Michigan
Yun Bai, University of Michigan

There is growing public health concern about the effect of exposure to airborne particles on cardiovascular disease. Prospective cohort studies have demonstrated associations between cardiovascular mortality and exposure to PM10 or PM2.5 (particles less than or equal to 10 or 2.5 microns, respectively, in diameter). However the mechanisms underlying these associations and the extent to which exposures over long periods are related to the development of atherosclerosis are still poorly understood. Furthermore, there are no studies that track individual level exposure to these particles over long periods. The present project combines information from two sources, data on PM10 and PM2.5 over a 20 year period obtained from EPA monitors at various locations and the residential history from a cohort study over the same period and providing disease outcome. A Bayesian approach is used for predicting residential level exposure time series and then the features of these time series are related to disease outcome. The predictions are based on spatio-temporal model that decomposes the time and location effects using parametric and semiparametric approaches.

email: teraghu@umich.edu

EMPIRICAL BAYES-TYPE SHRINKAGE ESTIMATION IN GENETIC EPIDEMIOLOY

Bhramar Mukherjee*, University of Michigan
Nilanjan Chatterjee, National Cancer Institute

In analyzing data from case-control genetic association studies, plausible constraints on the exposure distribution like gene-environment independence, Hardy-Weinberg equilibrium are being exploited to yield efficient estimation strategies based on the retrospective likelihood. However, these modern retrospective methods may incur bias under departures from the assumed constraints. The Bayesian paradigm is a natural alternative to build uncertainty around the assumed constraints, but comes with computational challenges, especially for large-scale association studies. We propose an empirical Bayes-type shrinkage estimation strategy for such problems, striking a balance between efficiency and model robustness, boosted with simple computation. The problem of estimating gene-environment interaction from case-control data and that of combining related and unrelated controls in family based studies will serve as illustrative examples of the proposed approach.

email: bhramar@umich.edu

39. LONGITUDINAL DATA ANALYSIS IN THE PRESENCE OF MORTALITY

ANALYSIS OF LONGITUDINAL DATA SUBJECT TO TRUNCATION DUE TO DEATH

Patrick J. Heagerty*, University of Washington

Longitudinal studies typically collect repeated measures of health status through a fixed follow-up time period. In studies of certain high-risk populations a large number of study participants have their follow-up stopped due to death. Statistical analysis of longitudinal data truncated by death requires selection of a specific analysis goal, and this talk will review current options with discussion of appropriate methods of inference, and caveats regarding the naive use of standard repeated measures methods.

email: heagerty@u.washington.edu

IDENTIFICATION AND ESTIMATION OF THE SURVIVOR AVERAGE CAUSAL EFFECT

Brian L. Egleston*, Fox Chase Cancer Center

Evaluation of the causal effect of an exposure on a non-mortality outcome in panel data is often complicated when study participants die before non-mortality outcomes are measured. In this setting, the causal effect is only well-defined for the principal stratum of subjects who would live regardless of the exposure. The Survivor Average Causal Effect (SACE) is an estimand of the effect in this stratum. We introduce a set of assumptions to identify SACE and embed our methodology within a sensitivity analysis framework. We provide data examples in which SACE was used to provide meaningful estimates of exposure effects when death was a significant competing risk.

email: brian.egleston@fccc.edu

USING AN INTERVENTION-BASED FRAMEWORK TO ADDRESS INPUT DATA MISSING DUE TO DEATH

Constantine E. Frangakis*, Johns Hopkins University
Donald B. Rubin, Harvard University
Ming-Wen An, Johns Hopkins University
Ellen MacKenzie, Johns Hopkins University

Missing data due to death occur often in research of high risk populations. Standard methods to address this problem assume ignorability of the missing data mechanism, that is, essentially comparability between observed and missing data within levels of covariate information. This assumption of course is often unrealistic. We have recently proposed a framework for addressing data missing due to death through explicit assumptions about interventions that would have caused these data to be observed. The framework allows for non-ignorable missing data mechanisms, and, as we demonstrate, provides more plausible answers to research questions.

email: cfrangak@jhsph.edu

40. IMPROVING MEASUREMENT IN PROFILING PROVIDERS OF HEALTHCARE

FINDINGS FROM THE HCAHPS MODE EXPERIMENT

Marc N. Elliott*, Rand Corporation
Elizabeth Goldstein, Centers for Medicare and Medicaid Services
William G. Lehrman, Centers for Medicare and Medicaid Services
Alan M. Zaslavsky, Harvard University
Katrin Hambarsoomians, Rand Corporation
Mary Anne Hope, Health Services Advisory Group
Laura A. Giordano, Health Services Advisory Group

Within each of 45 randomly sampled hospitals, a total of 27,229 patients were randomized in equal proportions to four modes of survey administration: Mail Only, Telephone Only, Mixed Mode (mail with telephone follow-up), or Active IVR (interactive voice response, in which patients respond via telephone keypads). These patients completed the CAHPS (Consumer Assessment of Healthcare Plans and Systems) Hospital Survey, in which recently discharged patients evaluate aspects of their hospital care. All surveys were administered by a single vendor. Linear regression was used to model each outcome from fixed effects for mode, hospital identifiers, and patient characteristics. Response rates varied strongly by randomized mode ($p<0.0001$), ranging from 41.2% for Mixed Mode to 20.7% for Active IVR; these patterns were consistent across hospitals. Substantial and statistically significant ($p<0.05$) mode effects were found for both 1 of 2 overall ratings and 6 of 7 composites. Patients provided more positive evaluations in the Telephone Only and Active IVR modes than in the Mail Only and Mixed modes. Differences between Telephone Only and Active IVR responses were small, and there were very few differences in Mail Only and Mixed Mode responses. When measured in terms of hospital-level standard deviations, the Telephone Only and Active IVR scores for the 7 most affected outcomes were at least 0.4-0.5 standard deviations higher and sometimes as much as 1 standard deviation higher than scores from Mail Only and Mixed modes. These mode effects varied little by hospital. Because a hospital's choice of vendor or survey mode may be confounded with factors related to underlying quality, an external mode experiment is necessary to estimate mode effects for subsequent fieldings of the survey; in the absence of such adjustments a hospital that would have ranked at the 50th percentile in the Mail Only mode would be ranked at the 66th to 84th percentile in the Telephone Only mode for a majority of outcomes. CMS will employ adjustments derived from this experiment, in conjunction with adjustment for patient-level characteristics, that allow hospitals to administer the HCAHPS Survey in any of four modes without affecting the comparability of their scores with other hospitals, national or regional averages, or their own past performance.

email: elliott@rand.org

ENSURING THE RELIABILITY OF CONSUMER ASSESSMENTS OF CLINICIANS AND GROUPS IN THE CAHPS 4.0 CLINICIAN AND GROUP SURVEY

Dana Safran*, Tufts University
Marc N. Elliott, Rand Corporation
Julie Brown, Rand Corporation
Ron D. Hays, Rand Corporation

Measuring consumer-reported healthcare experiences at the level of the clinician or medical group is a challenging proposition in that the necessity of ensuring adequate reliability for each of a larger number of providers must be weighed against the expense of a large number of clinician-specific surveys. In the context of pay-for-performance, physicians are likely to be especially interested in the reliability of a determinant of their compensation. In this talk we consider issues of follow-up and response rates, and sample size requirements in this context. We discuss the recommendation of 45 completes per physician in the context of empirical results regarding intraclass correlations of CAHPS outcomes by physician, standards for physician-level reliability, and margins of error expressed as percentiles of the distribution of physician-level means.

email: elliott@rand.org

IMPROVING SUBGROUP COMPARISONS OF CONSUMER REPORTS BY ADJUSTING FOR DIFFERENCES IN EXTREME RESPONSE TENDENCY

Amelia Haviland*, Rand Corporation
Marc N. Elliott, Rand Corporation
Katrin Hambarsoomians, Rand Corporation

Consumer evaluations of healthcare are a vital source of information for understanding racial/ethnic and other disparities in health and healthcare. Some previous analyses of CAHPS and other surveys of patient experience have found patterns between subgroups that are surprising and counterintuitive, even after case-mix adjustment (CMA). Such patterns include less positive ratings for those with supplementary insurance and higher income and more positive evaluations for African-Americans and Hispanics than Whites. We explore the role of extreme response tendency (ERT) in these patterns and propose a modeling adjustment for ERT that attenuates many previously counterintuitive findings.

email: elliott@rand.org

OPTIMAL SURVEY DESIGN WHEN NONRESPONDENTS ARE SUBSAMPLED FOR THE FOLLOW-UP

A. James O'Malley*, Harvard University
Alan M. Zaslavsky, Harvard University

Surveys often first mail questionnaires to sampled subjects and then follow up mail nonrespondents by phone. The high unit costs of telephone interviews make it cost-effective to subsample the followup. We derive optimal subsampling rates for the phone subsample for comparisons of health plans or other units. Computations under design-based inference depart from the traditional formulae for Neyman allocation because the phone sample size at each plan is constrained by the number of mail non-respondents and multiple plans are subject to a single cost constraint. Because plan means for mail respondents are highly correlated with those for phone respondents, more precise estimates (at fixed overall cost) for potential phone respondents are obtained by combining the direct estimates from phone followup with predictions from the mail survey using small-area estimation (SAE) models.

email: omalley@hcp.med.harvard.edu

41. BAYESIAN VARIABLE SELECTION WITH HIGH DIMENSIONAL COVARIATE DATA

A STOCHASTIC PARTITIONING METHOD TO ASSOCIATE HIGH-DIMENSIONAL DATASETS

Stefano Monni, University of Pennsylvania
Mahlet G. Tadesse*, Georgetown University

Several procedures have been developed to associate high-dimensional covariate data to univariate outcomes. In recent years, however, there has been a growing interest in relating data sets in which both the number of regressors and response variables are substantially larger than the sample size. For example, in an attempt to gain new insights into molecular processes, many efforts are being carried out to integrate data from various high-throughput genomic sources. We propose a Bayesian stochastic partioning method to identify sets of covariates associated with correlated outcomes. The procedure provides a unified framework for determining subsets of predictors that act jointly and uncovering response variables with similar expression patterns. We illustrate the method with an application to eQTL analysis, in which gene expression microarray data are related to genotype data from thousands of SNP markers.

email: mgt26@georgetown.edu

BAYESIAN VARIABLE SELECTION AND MONTE CARLO COMPUTATION FOR HIGH DIMENSIONAL DATA IN REGRESSION MODELS

Faming Liang, Texas A&M University
Ming Hui Chen, University of Connecticut
Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Mayetri Gupta*, Boston University

Prior elicitation and Monte Carlo computation are perhaps two major challenges in the modern Bayesian variable selection problem for general regression models when the number of covariates p in the regression model greatly exceeds the number of subjects n. For prior elicitation, ridge priors are developed and an entropy based criterion is proposed for eliciting the hyper-parameters. For Monte Carlo computation, we propose a novel Adaptive Stochastic Approximation Monte Carlo (ASAMC) algorithm. In the ASAMC algorithm, the sample space is adaptively partitioned to make sample more and more focused on the high posterior density regions, while maintaining the samples correctly weighted with respect to their importance weights. Our numerical results show that the ASAMC algorithm works effectively for the high dimensional variable selection problem. In addition, we show that the conditional probability of interest, P(sampled models | all models with the posterior probability greater than a specified value), can be conveniently estimated via ASAMC.

email: gupta@bu.edu

VARIABLE SELECTION VIA A BAYESIAN ENSEMBLE

Hugh A. Chipman, Acadia University
Edward I. George*, University of Pennsylvania
Robert E. McCulloch, University of Chicago

Bayesian methods provide an attractive and comprehensive approach to learning ensemble models such as forests of trees. We consider a particular case: the flexible and fast Bayesian Additive Regression Trees (BART) approach. BART can be used to screen for relevant predictors, providing an essentially nonparametric approach to variable selection. As the BART algorithm runs, different potential predictors enter the sum-of-trees model with different frequencies. Those that enter rarely or not at all are candidates for elimination, and those that enter frequently are candidates for inclusion. By varying the size of the sum-of-trees model, BART can identify subsets of predictors containing the strongest predictive information. BART also provides an omnibus test: the absence of any relationship between y and the predictors is indicated when BART posterior intervals for the response reveal no signal.

email: edgeorge@wharton.upenn.edu

42. CLUSTERED AND HIERARCHICAL MODELS

VACCINE EFFICACY TRIALS USING STEPPED WEDGE DESIGN

JoAnna Scott*, University of Washington

Cluster randomized trials are useful trial designs for evaluating infectious diseases. They have been used to evaluate the public health effects of vaccines. The Stepped Wedge design is a trial design that can be used to evaluate vaccine efficacy, especially in trials where it is unethical to assign placebo or have limitations making it difficult to distribute vaccines to all clusters simultaneously. In the Stepped Wedge design, time when an intervention is introduced is randomized and all clusters eventually receive the intervention. Total vaccine efficacy is the combined direct and indirect vaccine effects found by comparing outcomes in vaccinated participants from partially vaccinated clusters to outcomes in participants from totally unvaccinated clusters. This is the natural measure of vaccine efficacy found in a Stepped Wedge trial. This talk evaluates methods for estimating and testing total vaccine efficacy from a Stepped Wedge trial. Specifically, it will compare power in the Stepped Wedge design to that in a parallel design, how power is affected by VES, VEI, VEP, and the coverage fraction, and compare a GEE analysis to permutation tests and that of two modified GEE analyses, first a small sampled adjusted GEE and second, a GEE using the score function.

email: elorra@u.washington.edu

A GENERAL CLASS OF AGREEMENT COEFFICIENTS FOR CATEGORICAL RESPONSES

Wei Zhang*, The Pennsylvania State University
Vernon M. Chinchilli, The Pennsylvania State University

In this paper, we propose a general class of agreement coefficients for categorical responses. An agreement coefficient is used to measure the interrater agreement. Motivated by the traditional Cohen's kappa and the recent random marginal agreement coefficient (RMAC), we formulate this task using a parameter 'a', which reflects the distance between marginal distributions. Our approach generalizes Cohen's kappa as the lower bound and RMAC as the upper bound in a class of appropriate measurements of interrater agreement based on the discrepancy of marginal distributions. We study the large sample properties for the estimators of members of this class and conduct simulation studies to assess and compare the accuracy and precision of the estimators.

email: wuz108@stat.psu.edu

COMPARISON OF PARAMETRIC, NONPARAMETRIC AND SMOOTH NONPARAMETRIC MIXED EFFECT MODELS

Tihomir Asparouhov*, Mplus
Bengt Muthen, University of California Los Angeles

We conduct simulation studies to evaluate the effect of non-normality of the random effects in mixed models and to critically evaluate the advantages of the non-parametric and smooth non-parametric mixed effect models. We consider the effect of non-normality on the efficiency of the estimators, standard error estimation, likelihood ratio testing and model fit. Mixed effect models with continuous and categorical variables are considered. We also describe the estimation process as implemented in the Mplus software.

email: tihomir@statmodel.com

ANALYSIS OF GROUP RANDOMIZED TRIALS WITH MULTIPLE BINARY ENDPOINTS AND SMALL NUMBER OF GROUPS

Ji-Hyun Lee*, H. Lee Moffitt Cancer Center & Research Institute
Michael J. Schell, H. Lee Moffitt Cancer Center & Research Institute
Richard Roetzheim, University of South Florida

The group randomized trial (GRT) is a common study design to assess the effect of an intervention program aimed at health promotion or disease prevention. In GRTs, groups rather than individuals are randomized into intervention or control arms. Then, responses are measured on individuals within those groups. A number of analytical problems beset GRT designs. The major problem emerges from the likely positive intra-class correlation among observations of individuals within a group. Generalized linear mixed models (GLMMs) provide a suitable framework for handling a GRT design with binary data. In this study, we show how to use GLMMs to analyze GRT data with an application of a randomized cancer prevention trial, where multiple binary primary endpoints were obtained. The marginal p-values for the multiple endpoints were adjusted by the stepdown Bonferroni method. Use of small number of groups in a GRT raises additional concerns regarding the statistical analysis of the intervention effect. In this study, we will develop an index of extra variability to investigate group-specific effects on response. The index is designed to gauge the heterogeneity of response among the individual groups and provide insights into theses influence on the GRT study.

email: ji-hyun.lee@moffitt.org

HIERARCHICAL SPATIAL MODELING OF ADDITIVE AND DOMINANCE GENETIC VARIANCE FOR LARGE SPATIAL TRIAL DATASETS

Andrew O. Finley*, Michigan State University
Sudipto Banerjee, University of Minnesota
Patrik Waldmann, Swedish University of Agricultural Sciences
Tore Ericsson, Swedish University of Agricultural Sciences

This talk expands upon recent interest in Bayesian hierarchical models in quantitative genetics by developing spatial process models for inference on additive and dominance genetic variance within the context of large spatially referenced trial datasets. Direct application of such models to large spatial datasets are, however, computationally infeasible because of cubic order matrix algorithms involved in estimation. The situation is even worse in Markov chain Monte Carlo (MCMC) contexts where such computations are performed for several iterations. Here, we discuss and illustrate approaches that help obviate these hurdles without sacrificing the richness in modeling. For genetic effects, we demonstrate how an initial spectral decomposition of the relationship matrices negate the expensive matrix inversions required in previously proposed MCMC methods. For spatial effects, we outline two approaches for circumventing the prohibitively expensive matrix decompositions: the first leverages analytical results from Ornstein-Uhlenbeck processes that yield computationally efficient tridiagonal structures, while the second derives a predictive process model from the original model by projecting its realizations to a lower-dimensional subspace, thereby reducing the computational burden.

email: finleya@msu.edu

BAYESIAN GROUNDWATER CONTAMINATION MODEL

Yongsung Joo*, University of Florida
Keunbaik Lee, Louisiana State University Health Science Center
Donald Mercante, Louisiana State University Health Science Center

In this paper, we develop a Bayesian contamination model that clusters the sampling locations of groundwater into polluted and unpolluted groups and simultaneously estimates the average amount of human impact. Among major dissolved ions in groundwater, NO3, Ca, SO4, Cl, and Na were documented as useful variables describing the hydrochemical characteristics of anthropogenically polluted groundwater. Increased concentrations of these ions indicate that overused agrochemicals (particularly nitrogen fertilizers) and domestic sewage are the most important causes of groundwater pollution to considerable depths in the studied region. Our proposed model can be used to identify effective measures for groundwater quality management, such as source control.

email: yjoo@phhp.ufl.edu

A POSITIVE STABLE FRAILTY MODEL FOR CLUSTERED FAILURE TIME DATA WITH COVARIATE DEPENDENT FRAILTY

Dandan Liu*, University of Michigan
Jack D. Kalbfleisch, University of Michigan
Doug E. Schaubel, University of Michigan

In this article, we propose a positive stable shared frailty Cox model for clustered failure time data where the frailty distribution varies with cluster level covariates. The proposed model accounts for covariate dependent intra-cluster dependence and permits both conditional and marginal inferences. We use a stratified Cox type partial likelihood approach for regression parameter estimation. The proposed estimator is consistent and asymptotically normal with a covariance matrix for which a consistent estimator is provided. Furthermore, we establish the uniform consistency and joint weak convergence of the Breslow type estimator for the conditional cumulative baseline hazard function. Simulation studies show that the proposed estimation procedure is appropriate for practical use with a realistic number of clusters. Finally, we present a real application of the proposed method to kidney transplantation data from California.

email: dandanl@umich.edu

43. STATISTICAL GENETICS

BINARY TRAIT MAPPING IN EXPERIMENTAL CROSSES WITH SELECTIVE GENOTYPING

Ani Manichaikul*, Johns Hopkins University
Karl W. Broman, University of Wisconsin-Madison

Selective genotyping is well established as an efficient strategy for mapping quantitative trait loci. In the study of binary traits, one may consider focusing the genotyping on affected individuals, with some or none of the unaffecteds being genotyped. Under selective genotyping of this sort, the usual method for binary trait mapping, which conditions phenotypes on genotypes, is not appropriate. We consider an alternative approach, instead conditioning genotypes on phenotypes, and compare this to the more standard method of analysis, both analytically and by example. Based on our investigation, we recommend performing an initial genome scan with affecteds only, followed by genotyping of some unaffected individuals in genomic regions of interest to confirm results from the initial screen.

email: amanicha@jhsph.edu

A TIME-DEPENDENT POISSON RANDOM FIELD MODEL FOR POLYMORPHISM WTHIN AND BETWEEN TWO RELATED SPECIES

Amei Amei*, University of Nevada Las Vegas
Stanley Sawyer, Washington University in St. Louis

Characterizing the various forces that shape patterns of genetic polymorphism within and between species is the central goal of population genetics. Statistical inference using Poisson random field models of Sawyer and Hartl (1992) provides powerful likelihood and Bayesian methods for estimating some of these forces, such as mutation and selection. In this talk, I will derive a time-dependent Poisson random field model from which we can make inferences about the amounts of mutation and selection that have occurred in the history of observed aligned DNA sequences.

email: amei.amei@unlv.edu

AN INCOMPLETE-DATA QUASI-LIKELIHOOD FRAMEWORK WITH APPLICATION TO GENETIC ASSOCIATION STUDIES ON RELATED INDIVIDUALS

Zuoheng Wang*, University of Chicago
Mary Sara McPeek, University of Chicago

We propose an incomplete-data quasi-likelihood (IQL) framework to accommodate both missing and dependent data. The motivation of this method comes from genetic association studies, where we consider the problems of estimating haplotype frequencies and testing association between a disease and haplotypes based on multiple tightly-linked genetic markers, using case-control samples containing arbitrary combinations of related and unrelated individuals with relationships specified by known pedigrees. Current routine genotyping methods typically do not provide haplotype information; only the unphased genotype is directly observable. Statistically, this is a missing-data problem. Sampling multiple members from a pedigree provides additional haplotype information, but it also creates dependence among the genotypes of related individuals. This presents a dependent-data problem. By forming a quasi-likelihood score function based on the conditional expectation of the marginal natural minimal sufficient statistic, we draw inference without specifying the full joint distribution. The resulting IQL score function retains the optimality properties of the quasi-likelihood approach. The consistency and asymptotic normality of the IQL estimator are established. The application of our method to haplotype association analysis is developed. Simulation studies are conducted to evaluate the power and type I error of our proposed method for genetic association testing with related individuals.

email: zwang@galton.uchicago.edu

AN EPISTATIC MODEL FOR MAPPING PHENOTYPIC PLASTICITY OF A COUNT TRAIT

Arthur Berg*, University of Florida
Derek Drost, University of Florida
Evandro Novaes, University of Florida
Matias Kirst, University of Florida
Rongling Wu, University of Florida

Different expression of the same genotype in morphology, physiology and anatomy over changing environments is called phenotypic plasticity. The understanding of the genetic basis of phenotypic plasticity has been one of the most important challenges facing modern biology. In this study, we develop a statistical model for detecting specific quantitative trait loci (QTLs) that stimulate the phenotypic plasticity of a count trait through genetic regulations. The model was derived with a bivariate Poisson distribution with one variable for phenotypic means over different environments and the other for phenotypic differences between the same pair of environments. A multi-QTL model was implemented into a general mixture model framework, allowing the tests of how the same QTL affects the means and differences and whether the epistasis between different QTLs contribute to phenotypic plasticity. A real example for the number of sylleptic branches on the main stems of poplar hybrids was used to elucidate the interpretation and utility of our model. The new model will provide a useful tool for dictating the picture of genome by environment interactions that determine the final phenotypes of complex count traits.

email: berg@ufl.edu

GENOTYPING ERROR DETECTION IN SAMPLES OF UNRELATED INDIVIDUALS

Nianjun Liu*, University of Alabama at Birmingham
Dabao Zhang, Purdue University
Hongyu Zhao, Yale University

Data with genotyping errors are still common in genetic studies, since there are many situations where genotyping errors can be induced. Although many studies showed that genotyping errors could cause severe problems in genetic studies, almost all analytic methods assume that the inputs of the genotype data are without errors. The identification of genotyping errors still remains neglected, especially for studies with unrelated individuals where very few methods have been developed, besides HWE checking. These methods mainly rely on external 'validation' study, or replicates to get the estimates of error rates. We evaluate several models for genotyping error detection in unrelated population samples with SNPs genotype data. We show that the parameters of the models we evaluated are not identifiable; but with some restrictions on the parameter spaces, the parameters of some of the models are identifiable. Simulation study shows that one of the models performs well. We apply that model on HapMap data to show its usability in practice. The results show that even for high quality HapMap data, there are still SNPs which would have been overlooked by deviations from HWE alone but are suspicious to genotyping errors. Our work may provide a useful tool in genetic studies.

email: nliu@uab.edu

GENETIC MAPPING BY MINIMIZING INTEGRATED SQUARE ERRORS

Song Wu*, University of Florida
Guifang Fu, University of Florida
Yunmei Chen, University of Florida
Rongling Wu, University of Florida

Maximum likelihood estimation (MLE) is a very popular statistical tool in genetic mapping because of its many good properties. However, to maintain these good properties, we need to correctly specify the underlying model for our observed data. In real data analysis, we often encounter the situation in which the empirical density of our data is close to, but not quite the same as, a certain parametric distribution, such as a normal distribution. This motivates us to find a more robust approach that is flexible enough to accommodate a certain degree of misspecification of the true model. Recent studied on the integrated square error (or L2E) have shown that it has a good robustness property in parametric modeling. In this talk, we will present a new approach for genetic mapping by incorporating the idea of integrated square errors into the mixture setup. We will formulate hypothesis testing by defining a test statistic -- energy difference (ED).Simulation studies obtain the following results: (1) If the proposed model is true for the data, both the MLE and L2E can give consistent estimators although the MLE is more efficient. (2) If the proposed model is different from the true model (e.g. with outliers), the MLE is biased but the L2E provide consistent estimates. We applied the new approach to a real dataset of mouse growth, demonstrating its usefulness and utilization in a practical genetic mapping project.

email: swu@stat.ufl.edu

MULTILOCUS ESTIMATION OF THE RECOMBINATION FRACTION, OUTCROSSING RATE AND LINKAGE DISEQUILIBRIUM IN WILD POPULATIONS

Wei Hou*, University of Florida
Jiahan Li, University of Florida
Kun Han, Zhejiang Forestry University
Song Wu, University of Florida
Yanchun Li, Zhejiang Forestry University
Rongling Wu, University of Florida

Analysis of population structure and organization with DNA-based markers can provide important information regarding the history and evolution of a species. Linkage disequilibrium (LD) analysis based on allelic associations between different loci is emerging as a viable tool to unravel the genetic basis of population differentiation. We derive the EM algorithm estimate the linkage disequilibria between dominant markers, aimed to study the patterns of genetic diversity for a diploid species. The algorithm was expanded to estimate and test linkage disequilibria of different orders among three dominant markers and can be extended to manipulate an arbitrary number of dominant markers. The feasibility of the proposed algorithm is validated by an example of hickory trees, using dominant random amplified polymorphic DNA markers. The estimates of linkage disequilibrium between dominant markers were compared with that between codominant markers. Simulation results suggest that three-locus LD analysis displays increased power of LD detection relative to two-locus LD analysis. This algorithm will be useful for studying the pattern and amount of genetic variation within and among populations.

email: whou@biostat.ufl.edu

44. CLINICAL TRIAL ADAPTIVE DESIGN AND RANDOMIZATION

AN EFFICIENT CLINICAL TRIAL DESIGN INVESTIGATING TREATMENT BY COVARIATE INTERACTION

Ayanbola O. Elegbe*, Bristol Myers Squibb
David T. Redden, University of Alabama at Birmingham

When the effectiveness of a treatment depends upon the characteristics of the patient, it is important to detect this treatment by covariate interaction. Appropriately handling this interaction when it truly exists enables a clear inference of the true treatment effect. We propose a two-stage adaptive design that investigates the presence of a covariate by treatment interaction, and adjusts the second stage of the design conditional on the significance of the test of interaction. The information in the first stage is used to test for the statistical significance of the interaction which determines whether the clinical trial proceeds to a second stage. We examine the statistical properties of the procedure using a binary response, a dichotomous covariate and two treatments. Specifically, we investigate an interaction that exists when the treatment effect occurs solely in one level of the covariate. For trials that proceed beyond the first stage, the design allows for early termination of any stratum that shows little evidence of a treatment effect conditional on the information from the first stage. We achieve this by incorporating an early stopping rule based on a conditional power approach at the beginning of the second stage.

email: ayanbola.elegbe@bms.com

ESTIMATION OF TREATMENT DIFFERENCE IN PROPORTIONS IN CLINICAL TRIALS WITH BLINDED SAMPLE SIZE REESTIMATION

Xiaohui Luo*, Merck & Co.
Peng-Liang Zhao, Sanofi-Aventis

Shih and Zhao (1997) proposed a design with a simple stratification strategy for clinical trials with binary outcomes to re-estimate the required sample size during the trial without unblinding the interim data. The naïve point estimator for the between group difference in proportions is positively biased. Two methods are proposed to correct the bias in the naïve estimator: i.e., conditional bias correction, and bootstrap bias correction. Simulation studies show that our proposed methods compare favorably with the other unbiased estimators based on fixed weights.

email: edmund_luo@merck.com

NONLINEAR BAYESIAN PREDICTION MODEL

Haoda Fu*, Eli Lilly & Company
David Manner, Eli Lilly & Company

In phase II clinical trials, a central task is to select the dose from different dose levels which demonstrates efficacy and is well-tolerated. In this work, we propose a new nonlinear hierarchical Bayesian model for prediction of the final dose response during an interim analysis which facilitates dose selection. Furthermore, based on the predicted dose response, we can make several adaptive clinical decisions. Some example of decisions include: stop dosing a particular dose level; reallocating future patients to power the study, stopping a study for futility or positive efficacy. The model has a general structure which can be applied to different therapeutic areas. Also it naturally handles incomplete longitudinal data.

email: fu_haoda@lilly.com

INTERNAL PILOT WITH INTERIM ANALYSIS FOR SINGLE DEGREE OF FREEDOM HYPOTHESIS TESTS

John A. Kairalla*, University of Florida
Keith E. Muller, University of Florida
Christopher S. Coffey, University of Alabama at Birmingham

An internal pilot with an interim analysis (IPIA) design for Gaussian linear models is introduced and defined. The design allows for early stopping for efficacy and futility while also re-estimating sample size needs based on an interim variance estimate. In order for accurate study planning in small samples, exact theory is derived for single degree of freedom hypothesis tests within the general linear univariate model framework, which includes one or two group t-tests as special cases. Exact and computable forms of distributions allow fast and accurate calculations of power, type I error rate, and expected sample size. Examples compare study characteristics with a fixed sample design as well as with the internal pilot and two-stage group sequential designs, all of which can be seen as special cases within the IPIA framework.

email: jak@biostat.ufl.edu

OPTIMAL AND ADAPTIVE DESIGNS IN DOSE RANGING STUDIES BASED ON BOTH EFFICACY AND SAFETY RESPONSES

Olga V. Marchenko*, i3 Statprobe

Traditionally, most designs for Phase I studies gather safety information, aiming to determine the maximum tolerated dose (MTD). Then Phase II designs would evaluate the efficacy of doses in the assumed toxicity range. Recently, several designs for dose selection have been proposed that are based on both efficacy and safety. While a majority of designs provide appropriate, safe and efficacious dose or doses with some precision, few of them gain the sufficient information on all doses in the range studied. In this talk, I will compare several adaptive and optimal designs that are based on both efficacy and safety information with different response distributions in dose ranging studies.

email: olga.marchenko@i3statprobe.com

ADAPTIVE DESIGNS FOR DOSE FINDING BASED ON THE Emax MODEL

S. Krishna Padmanabhan*, Wyeth
Francis Hsuan, Temple University
Vladimir dragalin, Wyeth

We propose an adaptive procedure for dose-finding in clinical trials when the primary efficacy endpoint is continuous. We model the mean of the efficacy endpoint, given the dose, according to a four-parameter logistic function. The efficacy endpoint at each dose is distributed according to either a normal or a gamma distribution. We consider the cases of fixed variance and fixed coefficient of variation assuming them to be both known and unknown. The analytic formulae for the Fisher information matrix are obtained, which are used to build the locally and adaptive D-optimal designs. We show that this procedure of learning about the dose-response for a drug leads to greater efficiencies (translating to smaller sample sizes required) and precision of estimation

of points on the dose-response curve. We also present results of simulation studies conducted to investigate these designs.

email: skp@temple.edu

OPTIMAL ADAPTIVE DESIGNS FOR BINARY RESPONSE TRIALS

Youngsook Jeon*, University of Virginia
Feifang Hu, University of Virginia

One fundamental question in response adaptive randomization is what allocation proportion we should target to achieve requisite power while resulting in fewer treatment failures. For comparing two treatments, such optimal allocations are well studied in literature. Generalization to K (>2) treatments is necessary in practice. We are interested in finding the optimal allocation proportion, which achieves a requisite power of a multivariate test of homogeneity in binary response experiment while minimizing expected treatment failures at the same time. We propose new optimal allocation without completing the proof. We also present simulation studies which show response adaptive randomization procedure that targets this new optimal allocation performs better over complete randomization. Future research topics also are discussed.

email: yj6k@virginia.edu

45. GENOME-WIDE ASSOCIATION STUDIES

A BAYESIAN APPROACH FOR INCORPORATING PRIOR KNOWLEDGE IN GENOME-WIDE ASSOCIATION STUDIES

Haojun Ouyang*, North Carolina State University
Jung-Ying Tzeng, North Carolina State University

Genome-wide association studies (GWAS) have become possible with the advance of modern technologies. In GWAS, hundreds of thousands of tests are performed simultaneously, and one main challenge faced in multiple testing adjustments is to optimize the false positives and power. Here we introduce a method to measure genome-wide statistical significance based on the Bayesian decision framework and the false discovery rate (FDR). We propose a loss function and construct the corresponding optimal decision rule. The decision rule of significance is based on posterior probabilities, and aims to control for the average FDR. One advantage of the proposed method is that prior knowledge obtained from other resources, such as linkage studies, microarray analysis and others, can be readily integrated into the testing framework to up-weight or down-weight genomic regions. We illustrate how to specify the prior distributions that incorporate other information sources through a linkage-GWAS example. Our simulations show that the power can be substantially improved with correct prior information while the FDR is controlled at the desired level. When prior information is mis-specified, our method can still improve the power of detecting signals and keeps FDR under control. This proposed procedure is directly applicable to other high-throughput screening studies.

email: houyang@ncsu.edu

STEPWISE FORWARD MULTIPLE REGRESSION FOR COMPLEX TRAITS IN HIGH DENSITY GENOME-WIDE ASSOCIATION STUDIES

Xiangjun Gu*, M.D. Anderson Cancer Center
Christopher I. Amos, M.D. Anderson Cancer Center
Gary Rosner, M.D. Anderson Cancer Center
Mary Relling, St. Jude Children's Research Hospital
Ralph F. Frankowski, University of Texas School of Public Health at Houston

Most currently used methods for genome-wide association studies are based on separate single-nucleotide polymorphism (SNP) analyses. In simulation studies from a 115K SNP data set, methods based on separate SNP analyses were found to require either too stringent criteria to detect weak genetic effects or yield an excess of false positive results. To increase the power of detecting multiple weak genetic factors and reduce false positive results caused by multiple tests or dependence among test statistics, a modified stepwise forward multiple regression (SFMR) approach is proposed. Simulation studies showed that for detecting moderate and weak genetic effects, SFMR has significantly higher power than the Bonferroni and false discovery rate (FDR) procedures and that SFMR retains an acceptable false positive rate no matter whether causal SNPs are correlated with many SNPs in the genome; for detecting strong genetic effects, SFMR has a lower familywise error rate than the Bonferroni and FDR procedures when causal SNPs are correlated with many SNPs across the genome. When the Bonferroni significance criterion is adopted by SFMR, SFMR has a higher power and a lower familywise error rate than Bonferroni and FDR procedures.

email: xgu@mdanderson.org

PROBABILITY OF DETECTING DISEASE-ASSOCIATED SNPs IN CASE-CONTROL GENOME-WIDE ASSOCIATION STUDIES

Ruth Pfeiffer*, National Cancer Institute, National Institutes of Health
Mitchell Gail, National Cancer Institute, National Institutes of Health

Some case-control genome-wide association studies (GWASs) select promising single nucleotide polymorphisms (SNPs) by ranking corresponding p-values, rather than by applying the same p-value threshold to each SNP. For such a study, we define the detection probability (DP) for a specific disease-associated SNP as the probability that the SNP will be T-selected, namely have one of the top T largest chi-square values for trend tests of association. The corresponding proportion positive (PP) is the fraction of selected SNPs that are true disease-associated SNPs. We study DP and PP analytically and via simulations, for fixed and random effects models of genetic risk. DP increases with genetic effect size and case-control sample size, and decreases with the number of non-disease SNPs, mainly through the ratio of T to N, the total number of SNPs. We show that DP increases very slowly with T, and the increment in DP per unit increase in T declines rapidly with T. DP is also diminished if the number of true disease SNPs exceeds T. For a genetic odds ratio per minor allele of 1.2 or less, even GWAS with 1000 cases and 1000 controls require T to be impractically large to achieve an acceptable DP, leading to PP values so low as to make such studies futile.

email: pfeiffer@mail.nih.gov

SHRINKAGE ESTIMATION FOR ROBUST AND EFFICIENT SCREENING OF SINGLE SNP ASSOCIATION FROM CASE-CONTROL GENOME-WIDE ASSOCIATION STUDIES

Sheng Luo*, Johns Hopkins University
Nilanjan Chatterjee, National Cancer Institute
Bhramar Mukherjee, University of Michigan

In the genome wide association studies, the evaluation of the association between a disease and every single-nucleotide polymorphisms (SNPs) is often conducted as the initial step. Since the very large number of SNP comparisons in a whole-genome scan is likely to generate many false positive signals, only he top candidate SNPs (15,000-20,000) are assessed. Therefore, it is necessary to develop a powerful method for the preliminary screening of associations. We have developed a novel estimator using an empirical Bayes framework and a corresponding test for disease-SNP association. In the simulation studies and in applications to the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer project, sponsored by the National Cancer Institute, the proposed test achieves an excellent balance between bias and efficiency, depending on the true nature of the Hardy-Weinberg equilibrium among controls. It also has substantially better overall performance than the other existing methods, especially when the effect of the SNP is recessive. This methodology would be particularly useful for analyzing data from the initial screening stages of multi-stage association studies.

email: sluo@jhsph.edu

GENOME-WIDE ASSOCIATION STUDIES WITH RELATED INDIVIDUALS

Weihua Guan*, University of Michigan
Liming Liang, University of Michigan
Michael Boehnke, University of Michigan
Gonçalo R. Abecasis, University of Michigan

Recently, genome-wide association (GWA) studies have drawn great interest as a promising tool to dissect complex diseases such as hypertension, diabetes, and bipolar disorder. Population stratification is a major concern that can lead to spurious disease-marker association or mask a true association. We have proposed a similarity score matching method that matches cases and controls based on their genetic similarity, which can be accurately estimated using the large amount of genotype data in GWA studies. We now extend our method to the analysis of related case and control samples. We apply a new test statistic, based on the work of Thornton et al. 2007, to account for the correlation between the related individuals and matched case-control pairs. Through computer simulations, we demonstrate that our method correctly controls type I errors, and has improved power compared to genomic control in the presence of stratification. We illustrate our method with data from the Pritzker Consortium Bipolar GWA study.

email: wguan@umich.edu

A MIXTURE OF EXPERT MODEL FOR POPULATION STRATIFICATION IN GENOMIC ASSOCIATION STUDY

Yulan Liang*, University at Buffalo, The State University of New York

Many genetic association approaches have been developed to investigate contributions of polymorphisms in disease susceptibility. However, most of them are prone to spurious associations (false positive), which often cause non-replications in subsequent independent studies. Failure of replication is often blamed on confounding due to population stratification. In this paper, we propose a hierarchical mixture of expert model for addressing hidden confounding and controlling population stratifications. The high level of expert model is a mixture model while low level models include HMM model for the mixture proportions and density function of subpopulation with a generalized linear model, which can be further extended with generalized additive model or neural network model if continuous time covariate or spatial effect are included. This proposed mixture of expert model can be applied to various design scenarios of genetic association data such as model continuous traits or binary case-control or longitudinal studies. Monte Carlo simulations were carried out to assess both the Type I Error rate and the power of the proposed method. Finally we demonstrate and apply our model to cardiovascular study and showed that the new method is not susceptible to spurious associations.

email: yliang@buffalo.edu

A HIERARCHICAL BAYESIAN MODEL FOR GENOME-WIDE ASSOCIATION STUDIES OF SNPs WITH MISSINGNESS

Zhen Li*, University of Florida
George Casella, University of Florida

By the boost of technology, high through-put Single nucleotide polymorphism (SNP) dataset has been easily available. Association testing which aims at finding the causal genes or SNPs for the interesting traits or disease has been popular in the past decade. Typically the micro-array SNP datasets would have some missingness in it, the typical percentage of missingness is 5\% to 10\%. Valid statistical methods are in demand and some progress has been made, but still more advancement is in perspective. We propose a Bayesian hierarchical model to explain the SNP effects for the interesting traits, meanwhile the family structure information for the observations has been incorporated. Dealing with missing SNPs and fully family structure based genome-wide SNP association is a novel idea and we expect it gives better power for the association test. For this association test, not much information is expected to be known beforehand and missing SNP is imputed based on all the available information. We use Gibbs sampling to estimate the parameters and proved that just updating one SNP at a iteration still preserves the ergodic property of the Markov chain while it greatly improves computation speed.

email: zhenli79@ufl.edu

46. SPATIAL MODELING APPLICATIONS

REPEATED MEASURES METHODOLOGY FOR SPATIAL CLUSTER DETECTION WHILE ACCOUNTING FOR MOVING LOCATIONS

Andrea J. Cook*, University of Washington
Diane R. Gold, The Channing Laboratory, Brigham and Women's Hospital and Harvard University
Yi Li, Harvard University and The Dana Farber Cancer Institute

Spatial cluster detection has become an important methodology in quantifying the effects of hazardous exposures. Previous methods have focused on outcomes that are binary or continuous. There are no spatial cluster detection methods proposed for repeated measured outcomes. This manuscript provides an extension of the cumulative geographic residual method proposed for censored outcomes to repeated measured outcomes. A major advantage of this method is its ability to readily incorporate information on study participants relocation, which most spatial cluster detection statistics cannot. Application of these methods will be illustrated by the Home Allergens and Asthma prospective cohort study evaluating the relationship between environmental exposures and repeated measured outcome, persistent wheezing in the last 6 month, while incorporating moving.

email: cook.aj@ghc.org

SEASONAL/REGIONAL EFFECT ESTIMATES FOR PM2.5 AND HOSPITAL ADMISSIONS RATES

Keita Ebisu*, Yale University
Michelle L. Bell, Yale University
Roger D. Peng, Johns Hopkins University
Francesca Dominici, Johns Hopkins University

We investigated the spatio-temporal associations between PM2.5 levels and cause-specific hospital admission rates. We applied: (1) a pollutant-seasonal indicator model; and (2) a model to smooth seasonal patterns of county-specific log relative rates with a sine/cosine function. Both these models were fit using a generalized additive model with adjustment of possible confounders. In order to pool the county-specific estimates to form national average estimates, we fit a Bayesian hierarchical model using Two-Level Normal independent sampling estimation with non-informative priors. To assess the regional difference, we applied the hierarchical model stratified by region. Overall, the results of the pollutant-seasonal indicator model showed an association between PM2.5 and increased risk of hospital admissions during the winter. For example, for total respiratory infection, at lag 0, the estimated national average percentage increase in admission rate due to an increase of 10¼g/m ^ 3 in PM2.5 in winter was 1.05% (95% posterior interval 0.29, 1.82). Total cerebrovascular disease at lag 0 also showed significantly strong effects in winter. The results from the harmonic smoothing model produced similar results: the association between PM2.5 and hospitalization rates increases in the winter. We also found regional difference between eastern and western U.S.

email: keita.ebisu@yale.edu

MODELS FOR SPATIAL BIVARIATE BINARY DATA

Petrutza C. Caragea, Iowa State University
Emily J. Berg*, Iowa State University

A common situation encountered in environmental monitoring is to observe a multitude of (binary) characteristics at each site (spatial unit). One could examine the spatial structure individually for each of these attributes, using, for example, several centered auto-logistic models. Another traditional approach is to treat one of these characteristics as the primary response variable and the others as fixed covariates in the centered auto-logistic model. We propose a third approach, which conceptualizes two binary attributes as primary response variables and relates them to each other through additional dependence parameters in a bivariate auto-logistic model. We demonstrate that this approach renders interpretable results in an application rife with small sample sizes and missing observations.

email: emilyb@iastate.edu

WEIGHTED NORMAL SPATIAL SCAN STATISTIC FOR HETEROGENEOUS POPULATION DATA

Lan Huang*, Information Management Services, Inc.
Ram C. Tiwari, National Cancer Institute
Zhaohui Zou, Information Management Services, Inc.
Martin Kulldorff, Harvard University and Harvard Pilgrim Health Care
Eric J. Feuer, National Cancer Institute

In geographical disease surveillance and cluster detection, all the existing spatial methods for detecting clusters in continuous data are designed for individual-level data. Some cluster detection methods, including the Poisson based spatial scan statistic, can handle aggregated regional data, but are only designed for count data. Motivated by growing demands to study the spatial heterogeneity of continuous measures in population data, such as mortality rates, survival rates, body mass indexes (BMI) and pollution at state, county, and census tract levels, we propose a weighted normal scan statistic for searching the clusters of unusual high/low continuous regional measures, where the weights reflect the uncertainty of the regional measures, or sample size (number of observed cases) in the geographic units (cells). Power and precision, the selection of the weights, and the sensitivity of the proposed test statistic to data from various distributions are investigated through intensive simulation. The method is applied to 1988-2002 stage I and II lung cancer survival data in Los Angeles County and to 1999-2003 breast cancer age-adjusted mortality rate data in the US collected by the Surveillance, Epidemiology and End Results (SEER) program. The proposed weighted normal scan statistic will be included in the existing SaTScan software (www.satscan.org).

email: huangla@mail.nih.gov

SPATIO-TEMPORAL MODEL FOR IRREGULARLY SPACED AEROSOL OPTICAL DEPTH DATA

Jacob J. Oleson*, University of Iowa
Naresh Kumar, University of Iowa

It is common for space-time data to be sparsely collected over both space and time. Analyzing a very sparse data set is a challenge in such scenarios. We construct a square grid over the region and average measurements within each grid for measurements at spatial locations. A conditional auto-regressive model is used to model spatial correlation between the grids. Time intervals are also irregular for each grid and we use an exponential correlation function to model temporal relationships. The proposed model is a separable Bayesian hierarchical model. The model can be implemented in WinBUGS. Using the proposed methodology aerosol optical depth (AOD) from satellite data will be corrected for meteorological conditions and spatial-temporal structure to predict air quality for Delhi, India and its neighboring areas. We set up a 46 by 42 5km square grid and collect daily AOD measurements from 2003-2006. This methodology will have greater applications for estimating air quality at unprecedented spatial-temporal resolutions for air quality surveillance and management, epidemiological and environmental justice research, because the existing network of air pollution monitoring stations has limited spatial-temporal coverage.

email: jacob-oleson@uiowa.edu

INTRODUCING THE S-VALUE: AN EXPLORATORY TOOL FOR DETECTING SPATIAL DEPENDENCE ON A LATTICE

Petrutza C. Caragea*, Iowa State University
Mark S. Kaiser, Iowa State University

The application of Markov random fields to problems involving spatial data on lattice systems (as often desirable in the environmental and ecological sciences, agriculture, and other areas of biology) requires decisions regarding a number of important aspects of model structure. Existing exploratory techniques appropriate for spatial data do not provide direct guidance to an investigator about these decisions. We introduce a diagnostic quantity useful in situations for which one is contemplating the application of a Markov random field model based on conditional one parameter exponential family distributions. This exploratory diagnostic is shown to be a meaningful statistic that can inform decisions involved in modeling spatial structure with statistical dependence terms. We illustrate its use in guiding modeling decisions with simulated examples and demonstrate that these properties have use in applications.

email: pcaragea@iastate.edu

SPATIAL PROCESSES WITH STOCHASTIC HETEROSCEDASTICITY

Wenying Huang*, Colorado State University
Ke Wang, Colorado State University
Frank J. Breidt, Colorado State University
Richard A. Davis, Columbia University

Stationary Gaussian processes are widely used in spatial data modeling and analysis, which aim at the description, explanation and prediction of a spatial process based on a sample of observations. Stationarity is a relatively restrictive assumption regarding spatial association. By introducing stochastic volatility into a Gaussian process, we propose a stochastic heteroscedastic process (SHP) with conditional nonstationarity. That is, conditional on a latent Gaussian process, the SHP is a Gaussian process with non-stationary covariance structure. Unconditionally, the SHP is a stationary non-Gaussian process. The realizations from SHP are versatile and can represent spatial inhomogeneities. The unconditional correlation of SHP offers a rich class of correlation functions which can also allow for a smoothed nugget effect. For maximum likelihood estimation, we propose to apply importance sampling in the likelihood calculation and latent process estimation. Empirical results with simulated and real data show improved out-of-sample prediction performance of SHP modeling over stationary Gaussian process fit.

email: wenying@lamar.colostate.edu

47. MISSING OR INCOMPLETE DATA

DYNAMIC GRAPHICS AND SENSITIVITY ANALYSIS FOR DROPOUT DATA

Edward C. Chao*, Data Numerica Institute

In the analysis of longitudinal data, dropout problems are quite common. But the data might be dropout not at random, or, non-ignorable. The analysis for such data often relies on some strong assumptions, while such modeling methods might not be verifiable. A good approach is to investigate the sensitivity of the departure from the modeling assumption. It is of interest to study the sensitivity of influential cases, dependence of covariates, prior knowledge, or variance structures. Graphical methods, especially dynamic graphics, are quite useful to display the combination of impacts from various dimensions. We will present a new environment for such analyses. The visualization methods are also applicable to data without missing data problems.

email: ez.ap@163.com

NEW METHODS FOR ESTIMATING STAGE DISTRIBUTION IN CANCER REGISTRY DATA

Guoliang Tian*, University of Maryland Greenebaum Cancer Center
Ming T. Tian, University of Maryland Greenebaum Cancer Center

The goal of this paper is to develop new statistical methodology to improve the Surveillance, Epidemiology and End Results (SEER) reporting on cancer staging by utilizing supplemental SEER data. Cancer cases are staged as either localized, regional, distant or unstaged according to SEER Summary Stage 1977. The unstaged data represents a substantial portion of the patients in SEER. Current estimates of stage distribution, i.e., the true proportions of the localized, the regional and the distant stages, ignore the unstaged data and are biased. Thus they may provide misleading summary for cancer investigations and health policy decision making. Therefore, new sampling distributions are needed to describe such data and make statistical inference and comparisons. Having such an inferential structure will provide more accurate information for cancer investigations and decision making in health policy and services research.

email: gtian2@umm.edu

WEIGHTED ESTIMATING EQUATIONS FOR LONGITUDINAL STUDIES WITH DEATH AND NON-MONOTONE MISSING TIME-DEPENDENT COVARIATES AND OUTCOMES

Michelle Shardell*, University of Maryland School of Medicine
Ram R. Miller, University of Maryland School of Medicine

We propose a marginal modeling approach to estimate the association between a time-dependent covariate and an outcome in longitudinal studies where some study participants die during follow-up and both variables have non-monotone response patterns. The proposed method is an extension of weighted estimating equations that allows the outcome and covariate to have different missing-data patterns. We present methods for both random and non-random missing-data mechanisms. A study of functional recovery in a cohort of elderly female hip-fracture patients motivates the approach.

email: mshardel@epi.umaryland.edu

NONPARAMETRIC REGRESSION WITH MISSING OUTCOMES USING WEIGHTED KERNEL ESTIMATING EQUATIONS

Lu Wang*, Harvard University
Xihong Lin, Harvard School of Public Health
Andrea Rotnitzky, Harvard School of Public Health

We consider nonparametric regression on a covariate when the outcome is missing at random (MAR). We show that nonparametric kernel regression estimation based only on complete cases is generally inconsistent. We propose inverse probability weighted (IPW) kernel estimating equations and a class of augmented inverse probability weighted (AIPW) kernel estimating equations for nonparametric

regression under MAR. Both approaches do not require a parametric model for the error distribution and the estimators are consistent when the probability that a sampling unit is observed, i.e., the response probability, is known by design or is estimated using a correctly specified model. We show that the AIPW kernel estimator is double-robust in that it yields valid inference if either the model for the response probability is correct or the model for the conditional mean of the outcome given covariates and auxiliary variables is correct. In addition, adequate choices of AIPW kernel estimating equations help increase the efficiency with which we estimate the nonparametric regression function. We study the asymptotic properties of the proposed IPW and AIPW kernel estimators. We perform simulations to evaluate their finite sample performance, and apply the proposed methods to the analysis of the AIDS Costs and Services Utilization Survey data.

email: luwang@hsph.harvard.edu

MARGINALIZED SEMI-PARAMETRIC SHARED PARAMETER MODELS FOR INCOMPLETE ORDINAL RESPONSES

Roula Tsonaka*, Catholic University-Belgium
Dimitris Rizopoulos, Catholic University-Belgium
Geert Verbeke, Catholic University-Belgium
Emmanuel Lesaffre, Catholic University-Belgium

In this work we are concerned with the analysis of longitudinal ordinal responses subject to informative non-monotone missingness. A Shared Parameter Model (SPM) is used that postulates a random effects component to account for the dependence between the longitudinal responses and the missingness process. However, with SPMs there are mainly two issues that should be carefully considered. First, certain assumptions about the random effects distribution can possibly affect the inference. Second, when categorical responses are considered the use of a random effects model leads to a conditional interpretation of the model parameters, which is often not desirable. In this work, we propose a new SPM that simultaneously deals with the above issues. In particular, to avoid the impact of parametric assumptions for the random effects distribution to inference, we leave this distribution completely unspecified. In addition, to achieve marginal interpretation for the parameters, we extent the work of Heagerty and Zeger (2000) on marginalized multilevel models to the analysis of incomplete longitudinal ordinal responses. The estimation of the proposed model is based on the Vertex Exchange Method (Bohning, 1999). Finally, our model is exemplified in a study with patients suffering from Rheumatoid Arthritis.

email: spyridoula.tsonaka@med.kuleuven.be

INFORMATION ATTAINABLE IN SOME RANDOMLY INCOMPLETE MULTIVARIATE RESPONSE MODELS

Tejas A. Desai*, The Indian Institute of Management at Ahmedabad
Pranab K. Sen, The University of North Carolina at Chapel Hill

In a general parametric setup, a multivariate regression model is considered when responses may be missing at random while the explanatory variables and covariates are completely observed. Asymptotic optimality properties of maximum likelihood estimators for such models are linked to the Fisher information matrix for the parameters. It is shown that the information matrix is well defined for the missing-at-random model and that it plays the same role as in the complete-data linear models. Applications of the methodologic developments in hypothesis-testing problems, without any imputation of missing data, are illustrated. Some simulation results comparing the proposed method with Rubin's multiple imputation method are presented.

email: tdesai@iimahd.ernet.in

WHEN NMAR IS ALMOST MAR

Yan Zhou*, University of Michigan
Roderick Little, University of Michigan
Jack Kalbfleisch, University of Michigan

Missing values are common issues in empirical studies. The performance of the methods analyzing missing data strongly depends on the missing-data mechanism. If missing mechanism is ignorable, the likelihood based inferences will only depend on the observed data. For two variables X and Y where both of X and Y are missing, however, missing-data mechanism may not be ignorable, since the missingness of Y can possibly depend on the values of X which are missing. We propose a nonignorable missing-data mechanism, in which X is MCAR and Y is MAR given the value and missingness of X. The non-iterative maximum likelihood estimates exist and data are excluded for estimating certain parameters. Extensions of this type of mechanism will be also discussed.

email: yzhouz@umich.edu

48. FDA ADVISORY COMMITTEES: A STATISTICIAN'S ROLE IN DECIDING PUBLIC POLICY (A PANEL DISCUSSION)

AN INTRODUCTION TO FDA ADVISORY COMMITTEES

Gregory Campbell*, Center for Devices and Radiological Health, U.S. Food and Drug Administration

The Food and Drug Administration (FDA) has a number of Advisory Committees, all of which operate under the Federal Advisory Committee Act (FACA), which provides a mechanism for the U.S. government agencies to obtain expert input on a variety of subjects. Experts can volunteer to serve on the FDA Advisory Committees; if accepted they become Special Government Employees. As a science-based agency, FDA welcomes this expert input and uses Advisory Committee recommendations for a large number of difficult decisions involving the marketability of medical products, including not only drugs but also biologicals and medical devices. At the FDA the drug center (CDER) has 16 such committees, the biological center (CBER) 5, and the device center (CDRH) has 18 Advisory Panels, all organized according the medical product specialty. Statisticians play a key role as members or consultants on these committees, in terms of study design, analysis and interpretation of clinical trials. It is not uncommon that statisticians from the company, from the agency and from the panel weigh-in concerning the communication and discussion of the issues. The role of FDA statisticians in the preparation of these meetings is discussed.

email: greg.campbell@fda.hhs.gov

ROLE OF AN INDUSTRY STATISTICIAN IN AN FDA ADVISORY COMMITTEE MEETING

Frank W. Rockhold*, GlaxoSmithKline R&D

FDA Advisory Committee Meetings (Adcoms) are a critical event in the development of a medicine. The most important factors in determining the success of an adcom are the quality of the medicine and the research performed. In addition to an adcom being an important part of the FDA review and approval process, they are in a public forum. The industry statistician's role is to help the medical and regulatory staffs present the data in a manner that is true to the design and analysis of the trials and in a way clear to an audience with a wide range of backgrounds. The statistician serves as a critical bridge because he or she is often the closest to the data but rarely the one making the presentation. This is aided by the fact that analyses will already have been extensively reviewed by FDA staff so the issues should be clear ahead of time. The statistician should influence the presentation in a way that focuses on the facts in a clear and concise manner. A truly successful statistician in an adcom will have influenced the background material and presentations so that many of the technical questions from the committee can often be answered by the adcom and FDA statisticians with no intervention from the industry staff. Adcoms are a resource intensive and stressful exercise, but one where statisticians have a chance to display their technical and communication skills in a visible manner.

email: frank.w.rockhold@gsk.com

A STATISTICIANS ROLE IN FOOD AND DRUG ADMINISTRATION ADVISORY COMMITTEES

David A. Schoenfeld*, Massachusetts General Hospital

A statistician can play an important and unique role in determining public health policy as a participant at a Food and Drug Administration Advisory Committee. My talk will draw on my experience as a member of the Pulmonary-Allergy Drugs Advisory Committee and as a consultant for industry. A statistician's role is to communicate the statistical issues to the committee and to the public in order to help them understand the uncertainty that affects the question being considered. I will describe several of the statistical issues that I encountered at these meetings. I will describe barriers to effective communication and discuss how they might be overcome.

email: dschoenfeld@partners.org

SEEKING ADVICE FROM STATISTICIANS: A CLINICIAN'S PERSPECTIVE

Celia M. Witten*, U.S. Food and Drug Administration

FDA's mission is to promote the public health by promptly and efficiently reviewing clinical research and taking appropriate action on the marketing of regulated products in a timely manner. FDA Advisory Committees provide independent expert advice and assist in external review of FDA's intramural research programs. Advisory Committees are Product and/or Center Related. Topics brought before Advisory Committees can include: review of cutting edge scientific issues, product approval issues, adverse event problems, labeling issues, guidance documents, and peer review of intramural research. The statistician member of the Advisory Committee plays an important role in interpreting scientific and clinical data and communicating that interpretation both to the Advisory Committee to assist in its deliberations, and to FDA. Key areas that are important in addition to interpretation of effectiveness results in clinical studies are input into interpretation of safety data, and discussion of study design for future studies.

email: celia.witten@fda.hhs.gov

49. NEW DIRECTIONS IN SAFETY PLANNING AND ANALYSIS FOR CLINICAL DEVELOPMENT

REGULATORY PERSPECTIVES ON PLANNING FOR PRE-MARKETING SAFETY

George C. Rochester*, U.S. Food and Drug Administration

In the current climate of heightened expectation from consumers, legislative bodies and other stakeholders, there is a need to improve pre-marketing risk assessment. Risk evaluation must be systematically considered throughout the life-cycle of a product. This presentation will discuss 2 new ideas. The first is the Pre-approval Risk Identification Model (PRIM) which aims to streamline safety evaluation throughout the life-cycle to meet regulatory requirements for good safety assessment which forms the basis for good risk management. The second is the Safety Analysis Plan, presented as a tool for documenting the risk identification processes and analytical framework for regulatory submissions, data analysis and presentation. A case study will be presented.

email: george.rochester@fda.hhs.gov

PLANNING FOR META-ANALYSIS

Jesse A. Berlin*, Johnson & Johnson Pharmaceutical Research and Development

Within individual studies, unless they are specifically focused on a safety endpoint, little thought is generally given to precision for estimating treatment effects on safety endpoints. The Integrated Summary of Safety (ISS), is essentially a meta-analysis of individual patient data, and can (and should) be planned to address questions of interest. Planning the safety meta-analysis should involve consideration of statistical power and precision, anticipated analyses, both descriptive and inferential, anticipated subgroup analyses, and how the study design elements will complement each other across the program. With respect to subgroup analyses, planned meta-analyses offer precision that won't generally be available within individual studies. With regard to design elements, the issue of confounding of study design elements with each other should be explicitly addressed. For example, a low dose study in one subgroup (e.g., women) and a high dose study in another (e.g., men), completely confounds dose and sex, making it impossible to disaggregate the effects of either on the study findings. Examples will be presented to illustrate these points.

email: jberlin@prdus.jnj.com

ANALYSIS OF CLINICAL ADVERSE EVENT DATA USING FALSE DISCOVERY RATE METHODS

Devan V. Mehrotra*, Merck Research Laboratories
Joseph F. Heyse, Merck Research Laboratories

Adverse experience (AE) data in randomized clinical trials are routinely evaluated using between-group p-values for every AE encountered within each of several body systems. If the p-values (or corresponding confidence intervals) are interpreted without multiplicity considerations, there is a potential for excessive false positive findings that can needlessly cloud the safety profile of the drug or vaccine. We propose a novel method for addressing multiplicity in the evaluation of 'tier 2' AEs that achieves an ICH-recommended balance between type I and type II errors. The method involves a two-step application of adjusted p-values based on the Benjamini and Hochberg false discovery rate (FDR) procedure. Data from three moderate to large vaccine trials are used to illustrate the proposed "Double FDR" approach, and a fourth example serves to reinforce the potential consequences of failing to account for multiplicity. This work was in collaboration with the late Professor John W. Tukey.

email: devan_mehrotra@merck.com

DETECTING SAFETY SIGNALS IN CLINICAL TRIALS: A BAYESIAN PERSPECTIVE

H. Amy Xia*, Amgen, Inc.
Haijun Ma, Amgen, Inc.

Detection of safety signals from routinely collected adverse event data in clinical trials is critical in drug development. How to deal with the multiplicity issue and rare adverse event (AE) data in such a setting is a challenging statistical problem. Bayesian hierarchical mixture modeling [Berry and Berry (2004)] is appealing in the following aspects: 1) in contrast to considering each type of AE independently in a typical classical approach, it allows for explicitly modeling the AE data with the existing coding structure (e.g., system organ class (SOC) and preferred terms (PT) within the MedDRA dictionary) so that AEs within a SOC or across SOCs can borrow strength from each other; 2) it is attractive in dealing with rare AE data because the model modulates the extremes and 3) it is straightforward and flexible to assess the posterior probability of a clinically important difference with different scales (such as risk difference, odds ratio, or relative risk). In this presentation, we illustrate the use of Bayesian hierarchical binomial and Poisson mixture models for binary and subject-year adjusted outcomes, respectively. We also show some effective graphics for displaying flagged signals when analyzing hundreds or thousands of AEs.

email: hxia@amgen.com

50. VALIDATION OF GENOMIC CLASSIFIERS

THE MICROARRAY QUALITY CONTROL CONSORTIUM: EXPERIENCES IN THE DEVELOPMENT AND VALIDATION OF GENOMIC PREDICTIVE MODELS FOR CLINICAL AND TOXICOGENOMIC DATA SETS

Russell D. Wolfinger*, SAS Institute Inc.

Phase II of the Microarray Quality Control Consortium (MAQC) has focused on the development and validation of genomic predictive models suitable for regulated use. From an experimental design perspective, a key challenge has been determining effects of factors such as data pre-processing, predictor standardization, predictor reduction, and classification technique, representing millions of different ways of forming a predictive model. Further complexity arises from various combinations of assessment criteria and cross-validation methods, with a goal of determining generalizability performance and uncertainty. We discuss some experiences and lessons learned from a few of the MAQC data sets.

email: russ.wolfinger@sas.com

DEVELOPING AND VALIDATING GENOMIC CLASSIFIERS

Kevin K. Dobbin*, National Cancer Institute

Moving genomic classifiers from initial research laboratory discovery towards clinical utility requires validation of both the classification performance and of the ability to extend the assay technology to a clinically realistic setting. The biomarker literature is littered with markers that seemed to work in an initial laboratory study but failed to "scale up" or be reproducible. We discuss these issues in the context of a pair of publications from the NCI Director's Challenge Lung Study Consortium: a study of the reproducibility of assays from different laboratories following a common protocol, and a related study to assess a microarray based prognostic predictor. Various lessons learned are discussed, as well as issues in designing these types of studies including recent advances in determining the sample size needed to develop a predictor from high throughput data.

email: dobbinke@mail.nih.gov

GENE EXPRESSION PROFILING DEVICES FOR CANCER PROGNOSIS: FDA CLEARANCE PROCESS

Reena Philip*, U.S. Food and Drug Administration

A gene expression profiling device for cancer prognosis is a device that measures the RNA expression level of multiple genes and combines this information to yield a signature (pattern or classifier or index) to aid in prognosis of previously diagnosed breast cancer. In gene expression test systems for breast cancer prognosis, an algorithm is applied to such measurements to yield a result that can be used by physicians as a prognostic marker, in combination with clinicopathological factors, to assess the risk of cancer recurrence (e.g., distant metastasis). This type of prognostic device is one for which test results explain the variation in outcomes for patients who are otherwise alike in terms of a predefined set of characteristics such as biological features (e.g., women over age 50 at a specific stage of disease) or a previously defined treatment (e.g., women receiving no adjuvant therapy). FDA cleared the first gene expression profiling device for breast cancer prognosis in February 2007. This talk will discuss the FDA clearance process for such devices and also will provide an update on FDA's current recommendations for assessing clinical and analytical performance of these devices.

email: reena.philip@fda.hhs.gov

DRUG-DIAGNOSTIC CO-DEVELOPMENT: SIMULTANEOUS VALIDATION OF A NEW THERAPEUTIC AND A PHARMACOGENETIC DIAGNOSTIC BIOMARKER FOR SELECTING PATIENTS MOST LIKELY BENEFIT FROM IT

Gene A. Pennello*, U.S. Food and Drug Administration

Drug/test combinations have the potential to provide many clinical benefits to patients, allowing for more refined treatment choices in selecting, avoiding, or dosing drugs. Because the safety and efficacy of a new drug are dependent on the ability of a test to properly identify patients for drug treatment or avoidance, the safety and efficacy of the drug becomes inextricably linked to the safety and effectiveness of the test. To validate both, we discuss several study designs, including the randomized marker by treatment design, marker-based and partially marker-based strategy designs, and retrospective studies of banked samples. We also discuss statistical analysis, including adjustments for multiplicity when assessing both overall drug efficacy and efficacy within the test positive subpopulation.

email: gene.pennello@fda.hhs.gov

51. METHODS FOR ANALYZING MULTI-READER RECEIVER OPERATING CHARACTERISTIC (ROC) DATA

USING MARGINAL ANOVA MODELS TO MOTIVATE, GENERALIZE, AND DERIVE PROPERTIES FOR THE OBUCHOWSKI-ROCKETTE PROCEDURE FOR MULTI-READER ROC DATA ANALYSIS

Stephen L. Hillis*, Iowa City VA Medical Center

The Dorfman-Berbaum-Metz (DBM) method has been the most popular method for analyzing multireader ROC studies. Although it has performed well in simulations, a drawback of this method is that statistical properties are difficult to derive since it is based on a model with pseudovalues as outcomes. Recently it has been shown that equivalent results can be obtained using the Obuchowski-Rockette (OR) ANOVA model with correlated errors. Thus the DBM method has the same statistical properties as the OR procedure. Since statistical properties can be much more easily derived from the OR model, the OR model rather than the DBM model provides the foundation for further generalizations, for example, to split-plot and ANCOVA designs. However, a drawback of the OR model is that it is not a familiar model and properties are tedious to derive. In this talk I show how the OR model can be viewed as a marginal ANOVA model resulting from a conventional ANOVA model with independent errors. Viewing the OR model in this way provides an intuitive motivation for the OR model and generalizations of it, an easy way to derive the appropriate OR test statistics and their properties, and a unifying framework that clearly shows the relationship between the DBM and OR methods.

email: steve-hillis@uiowa.edu

ENSEMBLE VARIANCE FOR BINORMAL MODEL OF MRMC AUC

Brandon D. Gallas*, U.S. Food and Drug Administration

An expression for the multi-reader multi-case MRMC variance of the reader-averaged area under the empirical ROC curve has been recently derived. These expressions allowed for simple moment approaches to variance estimation under different study designs: designs where every reader reads every case, where every reader reads their own cases, and mixtures of the two. The derivations also gave a foundation for comparing the bias of two popular MRMC variance estimation methods: the jackknife and the bootstrap. The derivations also led to numerical integration solutions of the MRMC variance when the ROC scores are modeled as a linear combination of Gaussian random effects from the reader, case, modality, and interactions. This random effect model for the scores is heavily used to validate MRMC variance methods of all kinds (parametric, nonparametric, regression, and Bayesian methods) and the numerical solutions should assist in this validation.

email: brandon.gallas@fda.hhs.gov

COMPARISONS OF METHODS FOR ANALYSIS OF MULTIREADER MULTI-MODALITY ROC STUDIES

Xiao-Hua A. Zhou*, University of Washington, VA Puget Sound Health Care System

The receiver operating characteristic (ROC) curve is a popular tool to characterize the capabilities of diagnostic tests with continuous or ordinal responses. One common design for assessing the accuracy of diagnostic tests is to have each patient examined by multiple readers with multiple tests; this design is most commonly used in a radiology setting, where the results of diagnostic tests depend on a radiologist's subjective interpretation. The most widely used approach for analyzing data from such a study is the Dorfman-Berbaum-Metz (DBM) method (Dorfman, Berbaum and Metz, 1992) which utilizes a standard analysis of variance (ANOVA) model for the jackknife pseudovalues of the AUCs. In recent years, several authors have proposed some extensions to improve the BDM method. In this talk, focusing on continuous outcomes, we will introduce a marginal model approach based on the AUCs which can adjust for covariates as well. We compare the performance of our approach with the DBM method and its extensions via simulation and by a real data set.

email: azhou@u.washington.edu

52. RECENT ADVANCES IN THE MODELING OF COMPETING RISKS

CAUSE-SPECIFIC RELATIVE RISK MODELS FOR ESTIMATING ABSOLUTE RISK

Mitchell H. Gail*, National Cancer Institute

Absolute risk is the probability that a patient with specific risk factors and free of the disease of interest when evaluated initially will be diagnosed with that disease over a subsequent specified time interval. The term absolute risk is sometimes used synonymously with "cumulative incidence" or "crude risk", because absolute risk is reduced by the chance of dying of some other cause during the interval. One approach to modeling the effects of covariates on absolute risk is through cause-specific relative risk models. Cause-specific models display the effects of covariates separately on the hazard of the disease of interest and the hazard from competing causes of mortality; this representation facilitates an understanding of covariate effects. This approach has the added advantage that the models can be estimated not only from cohort data but also by combining case-control data with national age-specific disease rates and mortality rates. Such cause-specific models will be compared with models that relate absolute risk to covariates directly.

email: gailm@mail.nih.gov

PARAMETRIC INFERENCE ON COMPETING RISKS DATA

Jong-Hyeon Jeong*, University of Pittsburgh

In this talk, recent development in parametric inference on cumulative incidence function in K-sample and regression cases will be reviewed.

Flexible parametric distribution families will be introduced for modeling the cumulative incidence function. The results will be compared relative to existing nonparametric approaches. The methods will be demonstrated via real datasets from phase III breast cancer clinical trials performed by National Surgical Adjuvant Breast and Bowel Project (NSABP).

email: jeong@nsabp.pitt.edu

QUANTILE INFERENCE FOR COMPETING RISKS DATA

Jason P. Fine*, University of Wisconsin, Madison

A conceptually simple quantile inference procedure is proposed for cause specific failure probabilities with competing risks data. The quantiles are defined using the cumulative incidence function, which is intuitively meaningful in the competing risks set-up. We establish the uniform consistency and weak convergence of a nonparametric estimator of this quantile function. These results form the theoretical basis for extensions of standard one-sample and two-sample quantile inferences for independently censored data. This includes the construction of confidence intervals and bands for the quantile function, and two sample tests. Simulation studies and a breast cancer example illustrate the practical utility of the methodology.

email: fine@biostat.wisc.edu

53. ANALYSIS OF MARKED POINT PATTERNS WITH SPATIAL AND NONSPATIAL COVARIATE INFORMATION

ANALYSIS OF MARKED POINT PATTERNS WITH SPATIAL AND NONSPATIAL COVARIATE INFORMATION

Shengde Liang, University of Minnesota
Bradley P. Carlin*, University of Minnesota
Alan E. Gelfand, Duke University

Hierarchical modeling of spatial point process data has historically been plagued by computational difficulties. Likelihoods feature intractable integrals that are themselves nested within a Markov chain Monte Carlo (MCMC) algorithm. We extend customary spatial point pattern analysis in the context of a log-Gaussian Cox process model to accommodate spatially referenced covariates, individual-level risk factors, and individual-level covariates of interest that mark the process. We also use multivariate process realizations to capture dependence among the intensity surfaces across the marks. We illustrate using a collection of breast cancer case locations collected over the mostly rural northern part of the state of Minnesota that are marked by their treatment selection, mastectomy or breast conserving surgery (“lumpectomy”). The key substantive covariate (driving distance to the nearest radiation treatment facility) is spatially referenced, but other important covariates (notably age and stage) are not. Our approach facilitates mapping of marginal log-relative intensity surfaces for the two treatment options, and resolves the issue of whether women who face long driving distances are significantly more likely to opt for mastectomy while still accounting for all sources of spatial and nonspatial variability in the data.

email: brad@biostat.umn.edu

BAYESIAN SPATIAL SCAN STATISTIC ADJUSTED FOR OVERDISPERSION AND SPATIAL CORRELATION

Deepak Agarwal*, Yahoo! Research

Spatial scan statistic has become the method of choice for detecting spatial clustering after adjusting for inhomogeneity. The method is particularly suitable in applications where the goal is to find the actual location of spatial clusters or ”hotspots” as opposed to testing for global clustering. In this talk, we propose a Bayesian solution to the problem. A Bayesian solution has several advantages in this scenario. First, hotspot detection is based on posterior probabilities of models corresponding to each sub-region and hence there is no need to conduct the expensive randomization procedure. Second, compared to the classical approach where multiple hotspots are generally detected using a conservative test, detecting multiple hotspots in the Bayesian framework is automatic and does not require any additional machinery. Furthermore, we provide a framework to adjust for additional characteristics like overdispersion and spatial correlation using a Cox process formulation. We illustrate our method on datasets that have been previously analyzed in the literature.

email: dagarwal@yahoo-inc.com

ESTIMATING FUNCTIONS FOR INHOMOGENEOUS SPATIAL POINT PROCESSES WITH INCOMPLETE COVARIATE DATA

Rasmus P. Waagepetersen*, Aalborg University

Inhomogeneous spatial point processes find many applications in biology. In forest ecology for example, it may be of interest to study the relation between the intensity of locations of trees and environmental covariates describing topography and soil properties. The R package spatstat provides a very flexible and useful framework for fitting spatial point process models to point pattern data with spatial covariates. However, in practice one often faces incomplete observation of the covariates and this leads to parameter estimation error which is difficult to quantify. In this talk we consider a Monte Carlo version of the estimating function used in spatstat for fitting inhomogeneous Poisson processes and certain inhomogeneous cluster processes. For this modified estimating function it is feasible to obtain the asymptotic distribution of the parameter estimates in the case of incomplete covariate information. This allows a study of the loss of efficiency due to the missing covariate data.

email: rw@math.aau.dk

54. STATISTICAL INFERENCE

APPROXIMATE P-VALUES FROM MONTE CARLO HYPOTHESIS TESTING

Allyson M. Abrams*, Harvard Pilgrim Health Care and Harvard Medical School
Martin Kulldorff, Harvard Pilgrim Health Care and Harvard Medical School
Ken P. Kleinman, Harvard Pilgrim Health Care and Harvard Medical School

Monte Carlo hypothesis testing can be used to construct null distributions of statistics which are analytically unapproachable. Data are reassigned under the null and the test statistic recalculated. A drawback to this powerful technique is that to obtain an additional digit of p-value precision, ten times as many reassignments are required. P-values as small as 0.00001 (which require 100,000 reassignments) are desirable when many parallel tests are performed, as is the case when using the scan statistic for health surveillance. In such cases, the computation time for necessary precision is non-trivial. We propose a technique for obtaining precise p-values with fewer reassignments. We treat an observed set of statistics from reassignments as a sample from the null distribution, then fit a functional distribution to these observations. From this distribution, we calculate a continuous p-value. This technique can be used in any Monte Carlo hypothesis testing setting. For a scan statistic, we show that the Gumbel distribution fits the reassignment statistics well, and demonstrate that the resulting alpha levels are conservative. Other tested distributions resulted in anti-conservative alphas. The Gumbel-based p-values also have smaller variability than Monte Carlo p-values, suggesting that the proposed approach may result in greater power.

email: allyson_abrams@harvardpilgrim.org

AN APPLICATION OF EXACT TESTS USING TWO CORRELATED BINOMIAL VARIABLES IN CLINICAL TRIALS

Jihnhee Yu*, University at Buffalo
James L. Kepner, American Cancer Society

New therapy strategies for the treatment of cancer are emerging at a rapid pace owing to the recent technology advances in genetics and molecular biology. However, the corresponding ways in which we conduct clinical trials have remained nearly unchanged. When potentially efficacious therapies are tested, current clinical trial design and analysis methodologies may not be suitable for detecting therapeutic effects. The exact method using correlated bivariate binomial random variables can provide an alternative approach in clinical trials. The method can incorporate the two primary outcomes in the study design without a substantial increase of the sample size compared to univariate binomial distribution approach. The consequence of this strategy is that the significance level upper bound to test individual endpoint is relaxed so that the tests achieve a desired power even with relatively small improvement over the standard-of-care primary endpoints. Thus, the study design demands relatively less improvement (i.e. delta) of the objective response.

email: jinheeyu@buffalo.edu

IMPROVING EFFICIENCY OF INFERENCES IN RANDOMIZED CLINICAL TRIALS USING AUXILIARY COVARIATES

Min Zhang*, North Carolina State University
Anastasios, A. Tsiatis, North Carolina State University
Marie Davidian, North Carolina State University

The primary goal of a randomized clinical trial is to make comparisons among two or more treatments. For example, in a two-arm trial with continuous response, the focus may be on the difference in treatment means; with more than two treatments, the comparison may be based on pairwise differences. With binary outcomes, pairwise odds-ratios or log-odds ratios may be used. In general, comparisons may be based on meaningful parameters in a relevant statistical model. Standard analyses for estimation and testing in this context typically are based on the data collected on response and treatment assignment only. In many trials, auxiliary baseline covariate information may also be available, and it is of interest to exploit these data to improve the efficiency of inferences. Taking a semiparametric theory perspective, we propose a broadly-applicable approach to adjustment for auxiliary covariates to achieve more efficient estimators and tests for treatment parameters in the analysis of randomized clinical trials. Simulations and applications demonstrate the performance of the methods.

email: mzhang4@ncsu.edu

INFERENCE FOLLOWING AN ADAPTIVE DESIGN -PRACTICAL ISSUES ARISING FROM A LARGE STUDY IN ONCOLOGY

Lothar T. Tremmel*, Cephalon

Study 02CLLIII was planned as a randomized, open-label study in chronic lymphocytic leukemia, following a group sequential design with adaptive recalculation of sample size. At the third of four planned interim analyses, the study was stopped due to overwhelming efficacy, after n=300 patients were enrolled. Beyond determining the statistical significance, complete statistical inference comprises the calculation of the level of evidence ('p value'), point estimates, and confidence intervals. Challenges in deriving those measures were introduced both by the nonstandard nature of this design as well as by the way the trial was actually conducted. This talk will show how adjusted significance levels, adjusted p values, bias-corrected point estimates and confidence intervals that are adjusted for repated testing can be derived in this setting. The talk will conclude with an assessment whether the advantages of the design justify the considerable increase in complication of statistical inference.

email: lothar@tremmel.net

MODELING INFECTIVITY RATES AND ATTACK WINDOWS FOR TWO VIRUSES

Jian Wu, Miami University
A. John Bailer*, Miami University
Stephen E. Wright, Miami University

Cells exist in an environment in which they are simultaneously exposed to a number of viral challenges. Infection by one virus may preclude infection by other viruses. Under the assumption that the times until infection by two viruses are independent, a likelihood-based procedure is presented to estimate the infectivity rates and the size of the window during which a cell might be susceptible to infection by multiple viruses. A test for equal infectivity rates is proposed, along with interval estimates of parameters. The performance of this test and estimation procedure is explored in a small simulation study.

email: baileraj@muohio.edu

COMPARING STATISTICAL INTERVAL ESTIMATES

Michelle Quinlan*, University of Nebraska-Lincoln
James Schwenke, Boehringer Ingelheim Pharmaceuticals, Inc.
Walt Stroup, University of Nebraska-Lincoln

Statistical interval estimates are constructed to estimate parameters or quantify characteristics of a population. To correctly interpret each interval estimate, it must be clearly defined what each interval is estimating. Confidence and prediction intervals are well understood. However, the definition of a tolerance interval varies among literature sources. Because tolerance intervals are being used with more

frequency, a consistent definition of a tolerance interval needs to be established. To fully understand tolerance intervals and their relationship among other interval estimates, a computer simulation was conducted. In addition to tolerance intervals, other interval estimates considered in the simulation include simultaneous tolerance intervals, beta-equivalent tolerance intervals, beta-content tolerance intervals, and confidence/prediction intervals on confidence and prediction interval endpoints. Each interval estimate is evaluated and compared to determine the relationship among the estimates. Simulation results are presented in addition to the characteristics of each interval estimate. The estimates are quantified to establish the appropriate use of each interval across various situations.

email: mquinlan22@yahoo.com

EXACT BAYESIAN INFERENCE IN 2 BY 2 CONTINGENCY TABLES

Yong Chen*, Johns Hopkins University
Sining Chen, Johns Hopkins University
Haitao Chu, The University of North Carolina at Chapel Hill

The relative risk, odds ratio and risk difference are the most commonly used measures of the association between a binary exposure and a binary outcome. Using conjugate Dirichlet priors and Beta priors, we derive the exact posterior distributions of these measures for paired and independent binary endpoints. Bayesian inference based on these exact distributions is discussed, particularly when the sample size is small. The proposed method is applied to several biomedical studies.

email: yonchen@jhsph.edu

55. STATISTICAL METHODS- GENERAL

A NOVEL MOMENT-BASED DIMENSION REDUCTION APPROACH IN MULTIVARIATE REGRESSION

Jae Keun Yoo*, University of Louisville

Recently, a moment based sufficient dimension reduction methodology in multivariate regression, focusing on the first two moments, was introduced. We present in this article a novel approach of the earlier method in roughly the same context. This novel method possesses several desirable properties that the earlier method did not have such as dimension tests with chi-squared distributions, predictor effects test without assuming any model, and so on. Simulated and real data examples are presented for studying various properties of the proposed method and for a numerical comparison to the earlier method.

email: peter.yoo@louisville.edu

ONE-SIDED COVERAGE INTERVALS FOR A PROPORTION ESTIMATED FROM A STRATIFIED SIMPLE RANDOM SAMPLE

Phillip S. Kott*, National Agricultural Statistics Service, U.S. Department of Agriculture

Using an Edgeworth expansion to speed up the asymptotics, we develop one-sided coverage intervals for a proportion based on a stratified simple random sample. To this end, we assume the population units are independent random variables with a common mean within each stratum. These stratum means, in turn, may either be free to vary or are assumed to be constant. The more general assumption is equivalent to a model-free randomization-based framework when finite population correction is ignored. Unlike when an Edgeworth expansion is used to construct one-sided intervals under simple random sampling, it is necessary to estimate the variance of the estimator for the population proportion when the stratum means are allowed to differ. As a result, there may be accuracy gains from replacing the Normal z-score in the Edgeworth expansion with a t-score.

email: pkott@nass.usda.gov

ROBUST ESTIMATION OF THE SPECTRAL ENVELOPE

Mark A. Gamalo*, University of Missouri – Kansas City

In many applications, time series are collected with the interest of knowing whether signals are approximately the same without having to make restrictive assumptions. One way of determining this is whether any - and how many - have common cyclic components. This can be accomplished through the spectral envelope which is a principal components based approach first discussed in Stoffer et al. (Biometrika, 80, 611-622). As principal components approach, its accuracy is contingent upon the correct estimate of the spectral matrix. But this quantity can be severely biased and oscillatory in the presence of outliers; thus suggesting a need for a robust estimation procedure for the said technique.

email: gamalom@umkc.edu

RESTRICTED LIKELIHOOD RATIO TESTING FOR ZERO VARIANCE COMPONENTS IN LINEAR MIXED MODELS

Sonja Greven*, Ludwig-Maximilians-Universität Munich
Ciprian M. Crainiceanu, Johns Hopkins University
Helmut Küchenhoff, Ludwig-Maximilians-Universität Munich
Annette Peters, GSF-National Research Center for Environment and Health

The goal of our paper is to provide a transparent, robust, and computationally feasible statistical platform for restricted likelihood ratio testing (RLRT) for zero variance components in linear mixed models. This problem is non-standard because under the null the parameter is on the boundary of the parameter space. Our proposed approach is different from the asymptotic results of Stram and Lee (1994) who assume that the outcome vector can be partitioned into many independent subvectors. Thus, our methodology applies to a wider variety of mixed models, including those with moderate numbers of clusters or nonparametric smoothing. We propose two approximations to the finite sample null distribution of the RLRT statistic. Both approximations converge weakly to the asymptotic distribution obtained by Stram and Lee (1994) when their assumptions hold. When their assumptions do not hold, we show in extensive simulation studies that both approximations outperform the Stram and Lee approximation and the parametric bootstrap. Application of our approximations is straightforward, and the first has been implemented in an R package.

We also identified and addressed numerical problems associated with standard mixed model software. Our methods have been motivated by and applied to a large longitudinal study on air pollution health effects.

email: sonja.greven@stat.uni-muenchen.de

ONE-SIDED TESTS AND CONFIDENCE BOUNDS FOR THE DIFFERENCE BETWEEN PROBABILITIES FOR MATCHED PAIRS DICHOTOMOUS DATA

Donald J. Schuirmann*, U.S. Food and Drug Administration

We consider the case where data consist of pairs of dichotomous responses (yes/no, success/failure, etc.) obtained under two different conditions. For example, the members of a pair may be responses to two different treatments applied to the same subject. Let p1 and p2 be the probabilities of a 'yes' response for the first and second members of a pair. Our interest here is in hypotheses of the form H0: p1 - p2 > delta0 vs. H1: p1 - p2 <= delta0, where delta0 is not necessarily equal to zero. One way to test these hypotheses at level alpha is to obtain a 1 - alpha upper confidence bound for p1 - p2 and to reject H0 if this confidence bound is <= delta0. We consider normal approximation methods. McNemar (1947) showed that the variance of our estimator of p1 - p2 is a function of p1 - p2 itself and of the sum pB + pC, where pB and pC are the probabilities of discordant responses. Since p1 - p2 and pB + pC are unknown, different test statistics are obtained depending on the quantities that are substituted for them in the variance formula. The test statistic will also depend on whether or not a continuity correction is used. We evaluate the characteristics of several proposals, including those by Fleiss (1981), Lu and Bean (1995) and Nam (1997.) We present numerically obtained values of power (for tests) and coverage probabilities (for confidence bounds).

email: donald.schuirmann@fda.hhs.gov

SPATIAL MODELS WITH APPLICATIONS IN COMPUTER EXPERIMENTS

Ke Wang*, Colorado State University
Wenying Huang, Colorado State University
Frank J. Breidt, Colorado State University
Richard A. Davis, University of Columbia

We consider modeling a deterministic computer response as a realization from a stochastic heteroscedastic process (SHP), a stationary non-Gaussian process with non-stationary covariance function. That is, by conditional on a latent process, the SHP is a non-stationary Gaussian process. As such, the sample paths of this process are more varied and flexible than those produced by a traditional Gaussian process model. We develop an importance sampling method for likelihood computation and use a low-rank kriging approximation to reconstruct the latent process. Both the responses and its predictive error variances at unobserved locations can be predicted using empirical best predictor and empirical best linear unbiased predictor. The predictor error variance can be used to guide the adaptive sampling, using algorithm such as Active-Learning Mackay or Active-Learning Cohn. Using simulated and real computer experiment data, the SHP model is superior to traditional Gaussian process model in prediction and adaptive sampling.

email: kewang@lamar.colostate.edu

GRAPHICAL TOOL FOR BAYESIAN NETWORK RECONSTRUCTION

Peter Salzman*, University of Rochester
Anthony Almudevar, University of Rochester

Bayesian Network is a graphical representation of a multivariate distribution. This representation applied to gene expression data can be useful to understand the direct and indirect interactions between genes or gene products. The estimation/reconstruction of network from data is computationally intensive process as the space of possible models is superexponential in the number of genes. In this talk we will consider the problem of network reconstruction from expression data as a model selection problem. We propose a heuristic algorithm based on dynamic programming that computes a sequence of best scoring graphs.

email: psalzman@bst.rochester.edu

56. CLINICAL TRIALS DESIGN

DESIGN AND INTERIM MONITORING OF A CLINICAL TRIAL BASED ON THE GOMPETZ DISTRIBUTION

Arzu Onar*, St Jude Children's Research Hospital
Robert P. Sanders, St Jude Children's Research Hospital
Amar Gajjar, St Jude Children's Research Hospital
James M. Boyett, St Jude Children's Research Hospital

This talk details the statistical issues encountered in a recent single-arm trial at St Jude Children's Research Hospital, which proposed a risk-adapted therapy for children less than 3 years of age with embryonal brain tumors, choroid plexus carcinoma or ependymoma. The trial design aimed to estimate the event-free survival distribution, which has historically been characterized by a rapid decline during the first 1-2 years followed by an asymptote, representing patients who achieve long-term disease control. This pattern makes the use of conventional exponential distribution-based designs inappropriate. In order to appropriately capture the shape of the EFS distribution, we used the Gompertz distribution which can accommodate increasing, decreasing and constant hazard rates. The distribution's parameters are relatively easy to estimate and have intuitive interpretations, namely the cure-fraction and the median time to failure for those who are not cured. Further the log-rank type procedures can be used for testing. Sample size calculations of the trial were based on estimating the cure rate fraction and a single interim analysis was proposed. The latter however poses some additional challenges which will be detailed and some possible remedies will be presented.

email: arzu.onar@stjude.org

OPTIMAL COST-EFFECTIVE DESIGN STRATEGIES IN LATESTAGE CLINICAL TRIALS

Cong Chen*, Merck & Co., Inc.
Robert A. Beckman, Merck & Co., Inc.

This presentation addresses two important issues in late stage drug development. The first issue is, constrained with financial resource, how to decide number of Phase II proof-of-concept (POC) trials, size for each trial and corresponding Go-No Go decision criterion to Phase III. The second issue is how to appropriately set the futility boundary of interim analysis for a Phase III confirmation trial. Although these decisions have major financial implications and should be natually addressed in a quantitative fashion, they are either made using a heuristic argument that does not take financial factor into full account or made

from a statistical perspective that totally ignores the financial aspect. We will address the two issues using a common strategy, which is to find the optimal cost-effective design parameters by maximizing a benefit-cost ratio function. The numerator of the function is the probability-of-success adjusted benefit and the denominator is the expected total development cost. Design strategies associated with this approach are optimal cost-effective. This approach is most applicable to settings where resources saved from a No Go decision after a POC trial or from early termination of a Phase III confirmation trial can be immediately re-deployed to projects of equal or higher interest.

email: cong_chen@merck.com

RESOLVING A CLINICAL TRIAL WHEN ACCRUAL HAS SLOWED OR STOPPED

David N. Stivers*, Berry Consultants
Scott M. Berry, Berry Consultants
Donald A. Berry, University of Texas, M. D. Anderson Cancer Center

Accrual in a clinical trial may slow or stop well before planned enrollment is reached for many reasons. The sponsor may need to decide whether to close the trial, close and submit to the FDA, or continue the trial. Bayesian predictive methods can help with this decision. We describe an analysis where a medical device trial stalled at approximately half enrollment. Using a two-stage exponential time-to-event process to model 1 year mortality, and data from the current enrollment (including incomplete followup), we used Bayesian methods to predict that the probability of eventual success, according to the endpoints and analysis described in the original protocol, was almost certain. Furthermore, we found that if the trial were closed and followup completed on currently enrolled patients, the probability of eventual success, according to the endpoints and analysis described in the original protocol was also almost certain. Predictive analyses are ideally suited for this situation as they preserve the original sample size and intent of the trial. Results of this analysis might lead to FDA approval while using the analysis from the original protocol. Although this represents a best case, inference about other outcomes is also valuable to a sponsor.

email: david@berryconsultants.com

THE INFLUENCE OF DIFFERENT DESIGNS FOR SURVIVAL TRIALS ON THE TYPE I ERROR RATE OF THE LOGRANK STATISTIC

Jitendra Ganju*, Amgen, Inc.
Julie Ma, Amgen, Inc.

Survival trials are designed by fixing some but not all of the following design parameters: the number of subjects, the number of events, the enrollment time, or the follow-up time. What design parameters are fixed and what are not affect the Type I error rate of the logrank statistic. In particular, when the number of events is fixed and not large, the Type I error rate is erratic. If the number of events is random, the Type I error rate is preserved but for reasons other than that stated in the literature.

email: jganju@yahoo.com

OPTIMAL ENRICHMENT STRATEGIES FOR THE RANDOMIZED DISCONTINUATION DESIGN

Peter Müller, The University of Texas, M.D. Anderson Cancer Center
Gary L. Rosner, The University of Texas, M.D. Anderson Cancer Center
Lorenzo Trippa*, The University of Texas, M.D. Anderson Cancer Center

During the last years the randomized discontinuation design has been successfully applied in many clinical trials. Most applications are to oncology phase II trials for cytostatic agents. The design consists of two stages, a first preliminary open stage and a subsequent phase during which a subgroup of patients are randomly treated with the investigated agent or with a control therapy. The design is characterized by the following tuning parameters: the duration of the preliminary stage, the number of patients in the trial, and the selection criterium for the second stage. We discuss an optimal choice of the tuning parameters based on a Bayesian decision theoretic framework. We define a probability model for putative cytostatic agents and specify a suitable utility function. A computational procedure to select the optimal decision is illustrated and the efficacy of the prosed approach is evaluated through a simulation study.

email: ltrippa@mdanderson.org

CROSSOVER DESIGNS FOR MORTALITY TRIALS

Martha C. Nason*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Dean Follmann, National Institute of Allergy and Infectious Diseases, National Institutes of Health

The crossover is a popular and efficient trial design to assess the effect of treatment where each patient serves as her own control. Conventional wisdom is that crossover designs are not appropriate for absorbing binary endpoints, such as death or HIV infection. We explore the use of crossover designs in the context of a mortality endpoint and show that they can be more efficient than the standard parallel group design when there is heterogeneity in individuals' risks. We also introduce a new two-period design where first period survivors are re-randomized for the second period. This design combines the cross-over design with the parallel design and achieves some of the efficiency advantages of the cross-over design while ensuring that the second period groups are comparable by randomization. We discuss the validity of the new designs and propose both a mixture model and a modified Mantel-Haenszel test for inference. We consider extensions where patients repeatedly switch treatments, and where exact death times are used in the analysis. Simulations are used to compare the different designs and an example is provided to explore practical issues in implementation.

email: mnason@niaid.nih.gov

A COMPARISON OF TEST STATISTICS FOR THE SEQUENTIAL PARALLEL DESIGN

Xiaohong Huang, Sanofi-Aventis Corporation
Roy N. Tamura*, Eli Lilly and Company

The sequential parallel design has been proposed to improve the efficiency of clinical trials in psychiatry. The design consists of two

phases, an initial phase in which patients are randomized to placebo or drug, and a subsequent phase in which placebo non-responders are again randomized to placebo or drug. A crucial element in the choice of test statistics is the underlying treatment effect in the two phases of the design. We develop various likelihood based statistics under different assumptions of the treatment effect in the two phases. The statistics are compared in simulation studies to determine their underlying null and non-null behavior.

email: tamura_roy_n@lilly.com

57. STATISTICAL METHODS FOR GENOMICS

HARNESSING NATURALLY RANDOMIZED TRANSCRIPTION TO INFER CAUSAL REGULATORY RELATIONSHIPS AMONG GENES

Lin S. Chen*, University of Washington
Frank Emmert-Streib, University of Washington
John D. Storey, University of Washington

We develop an approach utilizing randomized genotypes to rigorously infer causal regulatory relationships among genes at the transcriptional level. Based on experiments where large-scale genotyping and expression profiling are performed, we calculate the probability that variation in expression level of a given gene has a regulatory effect on each other gene, for all pairs of genes. These probabilities can be used to build transcriptional regulatory networks and to identify putative regulators of genes, providing direct detection of the genes inducing variation in these expression traits. Since the method is based on randomized variables (genotypes), it avoids the usual pitfalls of correlation and model-selection based construction of networks. We apply the method to an experiment in yeast, where genes known to be in the same processes and functions are recovered in the resulting transcriptional regulatory network. We estimate a lower bound on the total number of regulatory relationships, yielding new insights into the topology of the yeast transcriptional regulatory network.

email: chenlin@u.washington.edu

INCORPORATING GENE NETWORKS INTO STATISTICAL TESTS FOR GENOMIC DATA

Peng Wei*, University of Minnesota
Wei Pan, University of Minnesota

It is a common task in genomic studies to identify a subset of the genes satisfying certain conditions, such as differentially expressed genes or regulatory target genes of a transcription factor (TF). Most existing approaches treat the genes as having an identical and independent distribution a priori, testing each gene independently or testing some subsets of the genes one by one. On the other hand, it is known that the genes work coordinately as dictated by gene networks. We propose incorporating gene network information into statistical analysis of genomic data. Specifically, rather than treating the genes equally and independently a priori in a standard mixture model, we assume that gene-specific prior probabilities are correlated as induced by a gene network: while the genes are allowed to have different prior probabilities, those neighboring ones in the network have similar prior probabilities, reflecting their shared biological functions and co-regulation. We applied the two approaches to a real ChIP-chip dataset (and simulated data) to identify the transcriptional target genes of TF GCN4. The new method was found to be more powerful in discovering the target genes.

email: weixx035@umn.edu

PREDICTION OF SURVIVAL TIME USING GENE EXPRESSION PROFILES OF MULTIPLE MYELOMA PATIENTS

Jimin Choi*, H.Lee Moffitt Cancer Center & Research Institute
Jimmy Fulp, H.Lee Moffitt Cancer Center & Research Institute
Dan Sullivan, H.Lee Moffitt Cancer Center & Research Institute
Choongrak Kim, Pusan National University

It is of great interest to predict survival time for cancer patients. In this paper, we estimate patient survival based on gene expression profiles. Initially, we use Significance Analysis of Microarrays (SAM) to select statistically significant genes related to survival time. We ultimately will fit a well-known Cox proportional hazard (PH) regression model, but first will use Partial Least Squares (PLS), in order to reduce the number of covariates, which is called PLSPH regression model. For each patient, we obtain median survival time predicted via the PLSPH regression model with leave-one-out cross validation. We compare models from the different number of PLS components with root mean Predicted Residual Sum of Squares (PRESS). This paper suggests that gene expression analysis may predict the each patient's survival time. Results are illustrated with Multiple Myeloma cancer data.

email: jimin.choi@moffitt.org

ANALYSIS OF GENE SETS BASED ON THE UNDERLYING REGULATORY NETWORK

Ali Shojaie*, University of Michigan
George Michailidis, University of Michigan

Development of high throughput technologies including DNA microarrays has facilitated the study of cells and living organisms. The challenge is no longer to identify the genes or proteins that are differentially expressed, but rather entire sub-systems that interact with each other in response to a given environmental condition. Study of these interacting sub-systems has provided an invaluable source of additional information that can be used to better understand the complex mechanisms of life. In this paper, we propose a latent variable model that directly incorporates the external information about the underlying network. We then use the theory of mixed linear models to present a general inference framework for the problem of testing the significance of subnetworks. Performance of the proposed model is evaluated on simulated and real data examples.

email: shojaie@umich.edu

A MODEL-BASED APPROACH FOR COMBINING HETEROGENEOUS HIGH THROUGHPUT GENOMIC DATA TO IMPROVE DETECTION OF REGULATORY MOTIFS

Heejung Shim*, University of Wisconsin-Madison
Adam Hinz, University of Wisconsin-Madison
Sunduz Keles, University of Wisconsin-Madison

Identification of transcription factor binding sites is one of the central challenges in genome research and it is central to understanding transcriptional regulatory processes in the cell. A recent innovation in this area is to integrate various sources of genomic data to improve detection of these regulatory sites. Heterogeneous sources of high throughput genomic data are typically obtained by ChIP-chip experiments measuring transcription factor localization or nucleosome occupancy and multi-species sequencing. Currently available approaches utilize prior information inferred from such data either by utilizing them one at a time or in a sequential manner. We propose a unified

conditional mixture model named SUCCESS that can capitalize on multiple sources of genomic data simultaneously and efficiently. This model allows the use of multiple sources of genomic data including sequence data and microarray ChIP-chip data from arrays with different resolutions. We use this model to analyze more than >80 yeast transcription factor datasets and illustrate how various sources of genomic data contribute to the detection of yeast transcription factor binding sites.

email: shim@stat.wisc.edu

NONPARAMETRIC META-ANALYSIS FOR IDENTIFYING SIGNATURE GENES IN THE INTEGRATION OF MULTIPLE GENOMIC STUDIES

Jia Li*, University of Pittsburgh
George C. Tseng, University of Pittsburgh

With the availability of tons of expression profiles, the need for meta-analyses to integrate different types of microarray data are obvious. For detection of differentially expressed genes, most of the current efforts are focused on comparing and evaluating gene lists obtained from each individual dataset. The statistical framework is often not rigorously formulated and a real sense of information integration is rarely performed. In this paper, we tackle two often-asked biological questions: "What are the signature genes significant in one or more data sets?" and "What are the signature genes significant in all data sets?". We illustrated two statistical hypothesis settings and proposed a best weighted statistic and a maximum p-value statistic for the two questions, respectively. Permutation analysis is then applied to control the false discovery rate. The proposed test statistic is shown to be admissible. And we further show the advantage of our proposed test procedures over existing methods by power comparison, simulation study and real data analyses of a multiple-tissue energy metabolism mouse model data and prostate cancer data sets.

email: jil53@pitt.edu

A GENOME-WIDE STUDY OF NUCLEOSOME FREE REGION IN YEAST

Wei Sun*, University of North Carolina
Wei Xie, University of North Carolina
Feng Xu, University of North Carolina
Michael Grunstein, University of North Carolina
Ker-Chau Li, University of North Carolina

Chromatin immunoprecipitation coupled with microarray (ChIPchip) has been widely used for detecting the binding of transcription factors (TFs) and epigenetic information such as nucleosome occupancy, histone posttranslational modification, DNA methylation etc. The array signal patterns generated from these two types of studies are however distinct: those from TF binding are typically short and located sparsely across the genome, while those from the occurring of epigenetic events are generally more abundant and cover considerably longer regions. Currently, most signal detection algorithms for ChIP-chip data are designed with the primary goal of identifying TF binding. To characterize epigenetic information from ChIP-chip data, we developed an efficient algorithm based on a segmental semi-Markov model, which outputs not only the locations but also the shape parameters of the recognized patterns. By applying our algorithm to genomewide nucleosome occupancy data in yeast generated by 4-bp highresolution tiling arrays, we identified both the locations of nucleosome free regions (NFRs) and their degrees of nucleosome depletion (DoND). Furthermore, our results enable a comparative study on two driving forces of evicting histones from the detected NFRs: the transcriptional machinery and the DNA affinity for histones.

email: wsun@bios.unc.edu

58. SURVIVAL ANALYSIS - NONPARAMETRIC METHODS

BEZIER CURVE SMOOTHING OF THE KAPLAN-MEIER ESTIMATOR

Choongrak Kim*, Pusan National University-Korea
Eunyoung Yun, Pusan National University-Korea
Mina Baek, Pusan National University-Korea

Estimation of a survival function from randomly censored data is very important in survival analysis. The Kaplan-Meier estimator (Kaplan and Meier, 1958) is a very popular choice, and kernel smoothing is a simple way of obtaining a smooth estimator. In this paper, we propose a new smooth version of the Kaplan-Meier estimator using a Bezier curve. We show that the proposed estimator is strongly consistent. Numerical results reveal that the proposed estimator outperforms the Kaplan-Meier estimator and its kernel weighted smooth version in the sense of mean integrated square error. Also, in estimating quantile of survival function, the proposed estimator gives better numerical results than others.

email: crkim@pusan.ac.kr

NONPARAMETRIC ESTIMATION OF STATE OCCUPATION, ENTRY AND EXIT TIMES WITH MULTISTATE CURRENT STATUS DATA

Ling Lan*, University of Louisville
Somnath Datta, University of Louisville

As a type of multivariate survival data, multistate models have a wide range of applications, notably in cancer and infectious disease progression studies. In this paper, we revisit the problem of estimation of state occupation, entry and exit times in a multistate model where various estimators have been proposed in the past under a variety of parametric and nonparametric assumptions. We focus on two nonparametric approaches, one using a product limit formula as recently proposed in Datta and Sundaram (Biometrics, 2006, 829-837) and a novel approach using a fractional risk set calculation followed by a subtraction formula to calculate the state occupation probability of a transient state. A numerical comparison between the two methods is presented using simulation studies. We illustrate the two methods using a pubertal development data set obtained from the NHANES III study.

email: l0lan001@louisville.edu

USING THE SEMINONPARAMETRIC DENSITY TO ESTIMATE SURVIVAL FUNCTIONS IN THE PRESENCE OF CENSORED DATA

Kirsten Doehler*, University of North Carolina at Greensboro
Marie Davidian, North Carolina State University

We propose a new procedure for estimating the survival function of a time-to-event random variable under arbitrary patterns of censoring. With only mild smoothness assumptions, this method allows a unified approach to handling different kinds of censoring, while in many cases increasing efficiency. Our approach uses a seminonparametric (SNP) density to represent the density of failure times. The SNP density has a flexible 'parametric' representation that results in a convenient expression for the likelihood and allows it to capture arbitrary shapes through choice of a tuning parameter. We have tested the performance of our procedure through simulation studies in which we have compared our SNP density approach to parametric and nonparametric methods, and to a semiparametric technique which uses smoothing splines. We also apply our method to data sets from biomedical studies.

email: kadoehle@uncg.edu

ESTIMATION FOR CENSORED HETEROSCEDASTIC LINEAR REGRESSION

Xuewen Lu*, University of Calgary
Zhulin He, University of Calgary

We consider censored heteroscedastic linear regression models. We propose an estimation procedure based on the method of iterative weighted least squares to incorporate both censoring and heteroscedasticity. The weights used in the estimation are the product of the inverse probability of censoring and the inverse of the error variance, where the probability of censoring distribution is estimated by the Kaplan-Meier estimator, while the error variance is estimated consistently and nonparametrically using local polynomial smoothing. The new estimator for the regression parameter is proved to be consistent and asymptotic normal. We conduct simulation studies to assess the finite sample performance of the proposed estimator and find that it is more efficient than the estimator that does not use the inverse variance weighting. The well known PBC data set is used for illustration.

email: lux@math.ucalgary.ca

LOCAL LINEAR ESTIMATION OF CONDITIONAL HAZARD FUNCTION

Jinmi Kim*, Pusan National University-Korea
Hyunmi Hwang, Pusan National University-Korea
Choongrak Kim, Pusan National University-Korea

The existing parametric or semiparametric estimation methods often fail to capture the shape of the conditional hazard function (hazard function with covariates) when the data are randomly right censored. In this paper we propose a nonparametric estimator for conditional hazard function using the double kernel approach of Fan, Yao and Tong (1996). Also, the proposed estimator is an extension of the unconditional hazard function estimator suggested by Kim et al. (2005). The proposed estimator is defined as the ratio of a kernel-weighted local linear regression for the conditional density and survival function. The resulting estimator is shown to be asymptotically normally distributed under appropriate assumptions.

email: banny2248@hanmail.net

EXAMINING MODEL FIT FOR PENALIZED SPLINES: A SIMULATION STUDY

Elizabeth J. Malloy*, American University
Donna Spiegelman, Harvard School of Public Health
Ellen A. Eisen, Harvard School of Public Health and School of Public Health, University of California-Berkeley

We present a simulation study of model fit criteria for selecting the optimal degree of smoothness for penalized splines in Cox models. The criteria considered are the Akaike information criterion (AIC), the corrected AIC (AICc), two formulations of the Bayesian information criterion (BIC), and a generalized cross-validation (GCV) method. The fit of nonlinear Cox regression models in six scenarios, which reflect classes of nonlinear exposure-response curves that are typically considered in practice, are estimated using the five model fit criteria. The methods are compared based on a mean squared-error score and the power and size of hypothesis tests for any effect and for detecting nonlinearity. No single criterion was clearly superior in all six exposure-response scenarios. BIC criteria were conservative and tended to choose degrees of freedom (df) close to 1 with little variability. GCV performed considerably worse than all other criteria when there was no effect. The AIC methods incorrectly rejected the null hypothesis of no association too often, but did correctly reject linearity for nonlinear relationships as often as the other criteria. The AIC and AICc methods had variable df and error scores, with both better, as well as worse fits, than other methods.

email: malloy@american.edu

INFERENCE FOR A CHANGE-POINT COX MODEL WITH CURRENT STATUS DATA

Rui Song*, University of North Carolina
Michael R. Kosorok, University of North Carolina
Shuangge Ma, Yale University

In this work, we investigate the inference of the change-point Cox model with an unknown covariate threshold for current status data. We show that the threshold parameter, the finite-dimensional regression parameter, and the infinite-dimensional baseline hazard function possess three different convergence rates and implement bootstrap likelihood ratio test and exponential average tests to detect the existence of a change-point; and apply the proposed method to a real dataset.

email: rsong@bios.unc.edu

59. FUNCTIONAL DATA ANALYSIS

FUNCTIONAL MIXED REGISTRATION MODELS

Donatello Telesca*, University of Texas, M.D. Anderson Cancer Center
Lurdes YT Inoue, University of Washington

Functional data arise in a number of scientific fields as the main response of interest. When the goal of our analysis coincides with understanding how much of the functional variation is explained by a set of covariates, standard linear and mixed effect models fail to provide a realistic scheme for the data generating process. For this reason a number of functional mixed effect models have been proposed and applied with success to the analysis of a sample of random functions. However, such models often assume that the function argument has been normalized to define a coherent response scale across functions (usually via curve registration). We introduce the class of Functional Mixed Registration Models, which allow for functional regression and curve registration in a joint fashion, fully accounting for random variability in phase and amplitude. Our model is developed in the general framework of Reproducing Kernel Hilbert Spaces and a few illustrations are provided in the case of Penalized Spline regression.

email: donatello.telesca@gmail.com

PRINCIPAL DIFFERENTIAL ANALYSIS: ESTIMATING COEFFICIENTS AND LINEARLY INDEPENDENT SOLUTIONS OF A LINEAR DIFFERENTIAL OPERATOR WITH COVARIATES FOR FUNCTIONAL DATA

Seoweon Jin*, Southern Methodist University
Joan G. Staniswalis, University of Texas at El Paso
Anu Sharma, University of Colorado at Boulder

Functional data, more commonly referred to as data curves, arise in many fields. It is often believed that the data curves correspond to a physical process described well by a differential equation, that is, each data curve is a sum of a random error term and a smooth curve in the null space of a linear differential operator. J.O. Ramsay first proposed the method of regularized principal differential analysis (PDA) for fitting a differential equation to a collection of noisy data curves. A smooth estimate of the coefficients of the linear differential operator is obtained by minimizing a penalized sum of the squared norm of the residuals. Once the coefficients of the linear differential operator are estimated, a basis for the null space is computed using iterative methods from numerical analysis for solving differential equations. Ultimately, a smooth low dimensional approximation to the data curves is obtained by regressing the data curves on the null space basis. This paper extends PDA to allow for the coefficients in the linear differential equation to smoothly depend upon covariates. The penalty of Paul H. C. Eilers and Brian D. Marx is used to impose smoothness. The estimating equations for coefficients in the PDA with covariates are derived; these are implemented in Splus and used to analyze evoked brain potential curves of subjects with normal hearing. The results of a small computer simulation study investigating the bias and variance properties of the estimator are reported.

email: sjin@smu.edu

IDENTIFYING TEMPORALLY DIFFERENTIALLY EXPRESSED GENES THROUGH FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS

Xueli Liu*, University of Florida
Mark C.K. Yang, University of Florida

Time course gene microarray is an important tool to identify genes with differential expressions over time under different conditions. The traditional analysis of variance type of longitudinal investigation may not be applicable, however, because the experimental designs for time course gene expression data cannot always be consistent across subjects, e.g., sampling rates and time intervals among different subjects may not be the same. Moreover, missing data is very common due to contamination in microarray experiments. Here we use functional principal components in covariance analysis to test hypotheses in the change of mean curves. A permutation test is used to make the method more robust. Analysis of a C. elegans microarray experiment identified many temporally differentially expressed genes under two different experimental conditions. Some genes were not identified by other methods.

email: xueli@stat.ufl.edu

FUNCTIONAL SLICED INVERSE REGRESSION IN LIPOPROTEIN PROFILE DATA

Yehua Li*, University of Georgia
Tailen Hsing, University of Michigan

Lipoprotein concentration in human serum is an important risk factor for cardiovascular heart disease. Different species of lipoprotein in a serum sample can be separated by a centrifugation treatment, according to their density differences. A lipoprotein profile for a patient is obtained by taking the image of his centrifuged serum sample, after an application of a lipophilic stain. In this work, we investigate functional regression models to predict the cholesterol levels from the lipoprotein profile curves. Our model is an extension of the multivariate dimension reduction theory to the functional data in the sense that the response depends on the functional predictor only through a few functional projections. The coefficient functions defining such projections are called Effective Dimension Reduction (EDR) direction, and the corresponding functional space is called the EDR space. We use a sliced inverse regression method to estimate the EDR space, and we generalize Li's (1991) χ^2 test to decide the dimension of the EDR space. When the functional predictor is non-Gaussion, an adjusted test is proposed. The procedures are examined through simulation studies and applied to the lipoprotein profile data.

email: yehuali@uga.edu

A GENERALIZED AKAIKE'S INFORMATION CRITERION FOR SELECTING PENALTY FUNCTION IN SPATIALLY ADAPTIVE SMOOTHING SPLINES

Ziyue Liu*, University of Pennsylvania
Wensheng Guo, University of Pennsylvania

Pintore, Speckman and Holmes (2006) proposed a novel spatially adaptive function estimation method by allowing the smoothing parameter to change over the domain. However, the question on how to use a data driven method to model the change of the smoothing parameter is left open. In this paper we propose a generalized Akaike's information criterion for modelling the penalty function in spatially adaptive smoothing splines. This is based on the equivalent Bayesian model that leads to a marginal likelihood for selecting the smoothing parameters. We develop a state space representation for efficient computation. The proposed criterion is purely data driven, computationally efficient and works well in simulation. We apply the proposed procedure to a real electroencephalograms data.

email: zliu5@mail.med.upenn.edu

BOOTSTRAPPING SUMS OF INDEPENDENT BUT NOT IDENTICALLY DISTRIBUTED CONTINUOUS PROCESSES WITH APPLICATIONS TO FUNCTIONAL DATA

Chung Chang*, Columbia University
Robert T. Ogden, Columbia University

In many areas of application, the data are of functional nature, such as (1-dimensional) spectral data and 2- or 3-dimensional imaging data. It is often of interest to test for the significance of some set of factors in the functional observations (e.g., test for the mean differences between two groups). Testing hypotheses point-by-point (voxel-by-voxel in neuroimaging studies) results in a severe multiple comparisons problem as the number of measurements made per observation is typically much larger than the number of observations ('large p, small n'). Thus solutions to this problem should take into account the spatial correlation structure inherent in the data. Popular approaches in such a setting include the general Statistical Parametric Mapping (SPM) approach and the permutation test, but these rely on strong parametric and exchangeability assumptions. In situations in which these assumptions are not satisfied, a nonparametric multiplier bootstrap approach may be used. Motivated by this problem, we present general results for multiplier bootstraps for sums of independent but not identically distributed processes. We also consider application of these results to an imaging setting and provide sufficient conditions that will ensure asymptotic control of the family-wise error rate.

email: cc2240@columbia.edu

DIRECTED SPATIO-TEMPORAL MONITORING OF DISEASE INCIDENCE RATES

Dan J. Spitzner*, University of Virginia
Brooke Marshall, Virginia Tech University

The talk describes a methodology for monitoring disease incidence-rates over time across a large spatial region. For instance, the methodology would be suitable for situations involving statewide county-by-county counts of a chronic disease reported on a monthly basis. The approach fits with existing approaches that model an incidence-rate surface using nonparametric estimation methods, but here estimation is weighted so as to emphasize structures that would naturally be characterized as clusters. Spatial estimation is then coupled with classical sequential process-control methods to produce a spatio-temporal monitoring scheme. The main advantage of the weighted-estimation approach is to direct statistical power away from structures of little interest in applications, thereby decreasing the signaling time for developing clusters. Other advantages include the flexibility to configure the procedure for custom monitoring objectives, such as the detection of departures from baseline levels, changes in the incidence surface, or space-time interactions. The approach also admits a follow-up mapping procedure, which provides information about the character of the cluster after its detection.

email: spitzner@virginia.edu

60. ADAPTIVE DESIGNS: PERSPECTIVES FROM ACADEMIA, INDUSTRY AND REGULATORY

ADAPTIVE DESIGNS: WHY, HOW AND WHEN?

Christopher Jennison*, University of Bath-U.K.

The potential applications of adaptive design methodology are diverse and one should not expect a simple answer to the question of when it may be beneficial to use an adaptive design. We shall consider the range of application areas and offer conclusions for each area. There are other, conventional approaches to consider in the domain of sample size re-assessment, in particular group sequential tests with pre-specified boundaries or error spending rules. Adaptive designs do offer the flexibility to respond to unexpected changes of intent or to rescue an inadequately planned trial. More significantly, combination tests at the heart of many adaptive designs play a key role in two-stage studies set up to include treatment selection, change of endpoint, or change of focus to a sub-population. Where two of the traditional phases of clinical testing are combined in this way, a 'seamless transition' may be achieved with benefits of continuity and, ultimately, more rapid recognition of an effective treatment --- as long as all the relevant practical and logistical issues can be addressed.

email: C.Jennison@bath.ac.uk

DESIGN AND IMPLEMENTATION OF ADAPTIVE TRIALS: EXPERIENCES OF AN INDUSTRY CONSULTANT

Cyrus R. Mehta*, Cytel Inc.

Sound statistical principles combined with careful planning of the logistical details are essential for successful implementation of an adaptive clinical trial. In this presentation we will share our experience as consultants involved in several late stage adaptive designs. Topics that we will cover include preservation of type-1 error, power, parameter estimation, multiple testing, population enrichment, the crucial role of simulation and the creation of an adaptive interim analysis charter to protect the study from biases. Actual cases will be used to illustrate all these topics.

email: mehta@cytel.com

ADAPTABILITY OF CLINICAL TRIAL DESIGN: A LARGER CONTEXT

H.M. James Hung*, U.S. Food and DrugAdministration

The recent advances in adaptive design methodology for evaluation of an experimental treatment range widely from a new look of sample size re-estimation to a mid-term change of statistical decision tree, such as alpha allocation. Regulatory experience with such methodology remains very much limited. This talk will give a brief overview of the interesting advances and present the scenarios where some types of adaptation may be worthy of and needs further exploration. Implementation issues related to practical concerns will be discussed.

email: hsienming.hung@fda.hhs.gov

61. PANEL DISCUSSION: THE EDUCATION OF CLINICAL TRIAL STATISTICIANS: DO WE NEED TO CHANGE WITH THE TIMES?

THE EDUCATION OF CLINICAL TRIAL STATISTICIANS: DO WE NEED TO CHANGE WITH THE TIMES?

Naitee Ting, Pfizer Inc., Global R&D

There are many aspects about potential need of changes in how we train clinical trial statisticians technical skills, how to work in a team, consulting, computing, communication skills, and other aspects. In this discussion, I hope to focus on the communication skills. There are at least 3 areas of communications that are important for clinical trial statisticians: oral communication, writing skills, and data presentation. To improve students ability in these areas, there are many things we can do: Allowing more consulting opportunities for students to participate;

Assign homework so that students have to turn in reports; Pay more attention to descriptive statistics and data interpretation. Changes can be made in lectures, in courses, in a particular program, or across the entire department.

email: naitee.ting@pfizer.com

62. CRITICAL DESIGN AND STATISTICAL CONSIDERATIONS WITH USE OF (GENOMIC) BIOMARKER FOR PREDICTION OR OPTIMIZING THERAPY

PREDICTION OF CLINICAL OUTCOMES WITH HIGH DIMENSIONAL MARKERS

Tianxi Cai*, Harvard University
Lu Tian, Northwestern University
Jie Huang, Northwestern University
Samuel McDaniel, Harvard University
Rebecca Betensky, Harvard University

Continuing technological advancements allow researchers and clinicians to measure an increasingly vast diversity of clinical and biological markers, rapidly increasing our understanding of disease processes. The wide range of newly available markers holds great potential for the personalization of medical care through accurate prediction of outcomes in individual patients. In the presence of high dimensional markers, it is challenging to develop reliable regression models that can be used to accurately predict future outcomes. We develop various robust approaches to the construction of classification and prognosis rules for predicting clinical outcomes.

email: tcai@hsph.harvard.edu

USE OF CLINICAL RECLASSIFICATION TO ASSESS RISK PREDICTION MODELS

Nancy R. Cook*, Brigham & Women's Hospital-Harvard University

The clinical literature, particularly in cardiology, has relied on the c-statistic, or area under the ROC curve, for assessment of models for risk stratification. This measure, since it is a function of sensitivity and specificity, is more appropriate for classification,or diagnostic testing, than in predicting future risk. In the latter setting, predictive values and calibration of estimated risk are as important as discrimination as assessed by the c-statistic. In risk prediction, perfectly calibrated models can usually only achieve values for the c-statistic well below 1. Comparing models through reclassification, or cross-stratification into clinical risk categories, can provide useful insight into the clinical value of new models, and can be used to assess the potential impact on treatment decisions for individual patients. Examples with respect to risk factors for cardiovascular disease will be presented.

email: ncook@rics.bwh.harvard.edu

UNCERTAINTIES IN THE MULTIPLE-BIOMARKER CLASSIFIER PROBLEM AND A PROPOSED STRATEGY FOR STUDY DESIGN AND ANALYSIS

Robert F. Wagner*, U.S. Food and Drug Administration
Weijie Chen, U.S. Food and Drug Administration
Waleed A. Yousef, Helwan University-Cairo, Egypt

Uncertainties of estimates of classifier performance in the classic work of Fukunaga were derived for problems in the neighborhood of the Bayes classifier. The curious results have led to neglect of the finite-training-sample contribution to the total variance, which provides a measure of the stability versus fragility of the classifier. We review these fundamentals of the field of statistical pattern recognition. They lead to Cover's theorem, the point of departure for defining classifier capacity and complexity, and their effects on uncertainties in classifier performance estimates. In practice, we need a coherent measurement strategy for estimating the total (i.e., finite-training plus finite-testing) contributions to uncertainty. We review the limitations of conventional cross-validation for this task and offer a more general approach, which makes use of the statistical influence function within a Pilot Study. The resulting total uncertainty is then used to size a Pivotal Study. This technical material will be given in a pedestrian (or 'Metro-Rider') tutorial style. We show how this assessment framework works on summary measures of the receiver operating characteristic (ROC) curve in binary classification tasks, as well as its potential on more general figures of merit beyond the binary classification task.

email: robert.wagner@fda.hhs.gov

PREDICT-1: THE FIRST POWERED, PROSPECTIVE, RANDOMISED TRIAL OF PHARMACOGENETIC SCREENING TO REDUCE DRUG ADVERSE EVENTS

Sara H. Hughes*, GlaxoSmithKline

Pharmacogenetics (PGx) is becoming an increasingly important research tool as physicians, patients, regulatory authorities and payers look for innovative ways to improve the risk:benefit ratio of medicines. One area where significant progress has been made is in the identification of PGx markers associated with variable response to antiretroviral medicines. For example, the major histocompatability complex HLA-B*5701 allele has been associated with hypersensitivity to abacavir by several independent researchers. However, this association between HLA-B*5701 and abacavir hypersensitivity has been identified largely through retrospective examination and in order to assure the medical and scientific communities of the clinical utility of this marker in patient care, a prospective, randomized study was required. This talk will outline the design of the study to evaluate the utility of prospective screening for HLA-B*5701 to reduce the incidence of abacavir hypersensitivity. The talk will also discuss statistical design issues considered during the development of the protocol and will present statistical analysis results. As the first fully powered, randomized, blinded, prospective study to determine the clinical utility of PGx screening to reduce drug-associated adverse events, the talk will further discuss methodological lessons learnt for future PGx researchers.

email: sara.h.hughes@gsk.com

63. ADVANCED SEMIPARAMETRIC MODELING OF GENETICS/GENOMIC DATA

SEMIPARAMETRIC METHODS FOR CASE-CONTROL ASSOCIATION STUDIES

Danyu Lin*, University of North Carolina

Proper analysis of case-control association studies requires the useof retrospective likelihood, which pertains to the conditional distribution of the genetic and environmental factors given the disease status. This distribution is infinite-dimensional in the presence of continuous environmental factors, so that the maximization of retrospective likelihood is in general a non-trivial task. For the logistic regression analysis of case-control data, the familiar prospective likelihood yields the same maximum likelihood estimators of odds ratios and their variances as the retrospective likelihood. This remarkable result does not hold when there are missing data in the genetic or environmental factors or when the scientific interest goes beyond the logistic regression on case-control status. In this talk, we describe a profile-likelihood approach to eliminating the infinite-dimensional parameter in retrospective likelihood and discuss applications to several important problems with real examples.

email: lin@bios.unc.edu

BIOMARKER DISCOVERY FOR GENOMICS AND PROTEOMICS DATA USING FUNCTIONAL MIXED MODELS

Jeffrey S. Morris*, The University of Texas-M.D. Anderson Cancer Center

Various genomic/proteomic assays yield high dimensional, irregular functional data. For example, MALDI-MS yields one-dimensional spectra with peaks, 2D gel electrophoresis and LC-MS yield two-dimensional images with spots that correspond to peptides present in the sample, and array CGH or SNP chip arrays yield one-dimensional functions of copy number information along the genome. I will discuss how to identify candidate biomarkers for various types of proteomic and genomic data using Bayesian wavelet-based functional mixed models. This approach models the functions in their entirety, so avoids reliance on peak or spot detection methods. The flexibility of this framework in modeling nonparametric fixed and random effect functions enables it to model the effects of multiple factors simultaneously, allowing one to perform inference on multiple factors of interest using the same model fit, while adjusting for clinical for experimental covariates that may affect both the intensities and locations of the peaks and spots in the data. I will demonstrate how to identify differentially expressed functional regions in a way that takes both statistical and clinical significance into account and controls the Bayesian false discovery rate to a pre-specified level. These methods are applied to proteomic and genomic data sets from cancer-related studies.

email: jeffmo@mdanderson.org

TESTING THE SIGNIFICANCE OF CELL-CYCLE PATTERNS IN TIME-COURSE MICROARRAY DATA

Guei-Feng Tsai, Center for Drug Evaluation-Taipei,Taiwan
Annie Qu*, Oregon State University

We develop an approach to analyze time-course microarray data which are obtained from a single sample at multiple time points and to identify which genes are cell-cycle regulated. Since some genes have similar gene expression patterns, to reduce the amount of hypothesistesting, we first perform a clustering analysis to group genes into classes with similar cellcycle patterns, including a class with no cell-cycle phenomena at all. Then we build a statistical model and an inference function assuming that genes within a cluster share the samemean model. A varying coefficient nonparametric approach is employed to be more flexible to fit the time-course data. In order to incorporate the correlation of longitudinal measurements, the quadratic inference function method is applied to obtain more efficient estimators and more powerful tests. Furthermore, this method allows us to perform chi-squared tests to determine whether certain genes are cell-cycle regulated. A data example on cell-cycle microarray data as well as simulations are illustrated."

email: qu@science.oregonstate.edu

A FOREST-BASED APPROACH TO IDENTIFYING GENE AND GENE-GENE INTERACTIONS

Xiang Chen, Yale University
Ching-Ti Liu, Yale University
Meizhuo Zhang, Yale University
Heping Zhang*, Yale University

Multiple genes, gene-by-gene interactions, and gene-by-environment interactions are believed to underlie most complex diseases. However, such interactions are difficult to identify. While there have been recent successes in identifying genetic variants for complex diseases, it still remains difficult to identify gene-gene and gene-environment interactions. To overcome this difficulty, we propose a forest-based approach and a concept of variable importance. The proposed approach is demonstrated by simulation study for its validity and illustrated by a real data analysis for its use. Analyses of both real data and simulated data based on published genetic models show the effectiveness of our approach. For example, our analysis of a published data set on age-related macular degeneration (AMD) not only confirmed a known genetic variant (p-value =2e-6) for AMD, but also revealed an un-reported haplotype surrounding single nucleotide polymorphism (SNP) rs10272438 on chromosome 7 that was significantly associated with AMD (p-value = 0.0024). These significance levels are obtained after the consideration for a large number of SNPs. Thus, the importance of this work is two-fold: a powerful and flexible method to identify high-risk haplotypes and their interactions, and the revelation of a potentially protective variant for AMD.

email: heping.zhang@yale.edu

64. THE USE OF SURVEY SAMPLES IN EPIDEMIOLOGICAL FOLLOW-UP STUDIES

SESSION OVERVIEW

Barry I. Graubard*, National Cancer Institute

The panel will address the use of random probability sample designs for identifying a cohort for recruitment into an epidemiological study that is intended to be followed for a number of years. Random sample designs are popular for their ability to select the study cohort with a minimum of bias and to support estimation of baseline prevalence of various risk factors to some target population of inference. Disadvantages are that the cohort selected through random sampling, for example as opposed to volunteer samples, may not be ideal for long-term retention, thereby limiting the ability to relate baseline risk to disease outcomes that occur during follow-up. Also, when using random samples, there are cost considerations related to contacting and recruiting study subjects residing in wide geographic areas as well as statistical efficiency considerations because of incorporation of sample weighting and other aspects of the sample design into analyses. Experiences from the planning phase of two actual studies, the National Children's Study and the Hispanic Community Health Study, incorporating random sample designs will be presented, along with the rationale for choosing these designs. Discussants will then address the advantages and disadvantages of random sample study designs versus other ways of selecting samples.

email: graubarb@mail.nih.gov

THE NCS: ESTABLISHMENT AND PROTECTION OF THE INFERENTIAL BASE

Jonas H. Ellenberg*, University of Pennsylvania

After many years of public debate involving experts in surveys,epidemiology, survey statistics, and biostatistics as well as the clinical disciplines involved, the decision was made to design the NCS as a national probability sample of 100,000 pregnancies and follow-up of the resulting offspring for 21 years. One of the major benefits that accompanies this decision is the potential of freedom from selection bias in determining the cohort. This benefit can be diminished if: the choice to join the NCS brings about its own selection bias (e.g. low SEI factors into potential participants' decisions to join the study); if, after joining, participants drop out of the study in a fashion related to demographics or potential outcomes (e.g. predilection for participants with at risk pregnancies to maintain their membership); if after joining, participants quit the study for non random reasons (e.g. related to environmental factors); or if, after joining the NCS, participants miss scheduled data collection appointments, or do not complete a visit schedule (e.g. participants with smooth obstetric or pediatric course may not feel it important to follow the protocol specified visit schedule to the letter or may not participate in tests or interviews that are annoying to them). Examples will be presented to justify the need for a non selected sample and the need to protect the inferential base throughout the duration of the study.

email: jellenbe@mail.med.upenn.edu

65. ANALYSIS OF RECURRENT MARKER DATA

FORWARD AND BACKWARD RECURRENT MARKER PROCESSES

Mei-Cheng Wang*, Johns Hopkins University

In biomedical follow-up studies, marker measurements are frequently collected or observed conditioning on the occurrence of recurrent events. In some situations, marker measurement does not even exist unless a recurrent event took place. A recurrent marker process is defined using both recurrent event times and markers where these two kinds of measurements are possibly correlated with each other. The collection of recurrent marker data is typically terminated by administrative censoring or occurrence of a terminal event such as death. This talk will consider forward and backward recurrent marker processes, and discuss definitions, models and estimation related to these two kinds of processes. Data examples will be presented to illustrate the proposed models and methods.

email: mcwang@jhsph.edu

ANALYSIS OF LONGITUDINAL DATA IN THE PRESENCE OF INFORMATIVE OBSERVATIONAL TIMES

Lei Liu*, University of Virginia
Xuelin Huang, University of Texas, M.D.-Anderson Cancer Center
John O' Quigley, University of Virginia

In longitudinal observational studies, repeated measures are often taken at informative observation times. Also, there may exist a dependent terminal event such as death that stops the follow up. For example, patients in poorer health are more likely to seek medical treatment and their medical cost for each visit tends to be higher. They are also subject to a higher mortality rate. In this paper we propose a random effects model of repeated measures in the presence of both informative observation times and a dependent terminal event. Three submodels are used, respectively, for (1) the intensity of recurrent observation times, (2) the amount of repeated measure at each observation time, and (3) the hazard of death. Correlated random effects are incorporated to join the three submodels. The estimation can be conveniently accomplished by Gaussian quadrature techniques, e.g., SAS Proc NLMIXED. An analysis of the cost accrual process of chronic heart failure patients from the clinical data repository (CDR) at the University of Virginia Health System is presented to illustrate the proposed method.

email: liulei@virginia.edu

REGRESSION ANALYSIS OF LONGITUDINAL DATA WITH DEPENDENT OBSERVATION PROCESS

(Tony) Jianguo Sun*, University of Missouri
Liuquan Sun, University of Missouri
Dandan Lu, University of Mchigan

Longitudinal data frequently occur in many studies such s longitudinal follow-up studies. To develop statistical methods and theory for the analysis of them, independent or noninformative observation process is typically assumed, which naturally leads to inference procedures conditional on the observation process. In many situations, however, this may not be true or realistic. This talk considers situations where the assumption does not hold and a joint modeling approach that uses some latent variables to characterize the correlations is presented.

email: sunj@missouri.edu

66. JOINT MODELING

PREDICTING OUTCOMES FROM LATER-STAGE LONGITUDINAL TRIALS WITH POTENTIAL WITHDRAWALS USING FINDINGS FROM EARLY STAGE TRIALS: A CASE HISTORY

Lawrence Gould*, Merck Research Laboratories
Kenneth Liu, Merck Research Laboratories

Simulating the outcome of clinical trials using models based on pharmacologic principles and information obtained in the early stages of drug development is anticipated to be useful for evaluating alternative trial design strategies. Although this principle is straightforward, its application in practice may not be. In the application considered here, information from pre-phase 2b longitudinal trials (from which patients withdrew) was to be used to inform the design of phase 2b and phase 3 trials. This presentation describes the issues that were addressed, such as appropriate models for the trajectory of longitudinal observations, the pattern of withdrawals, and the error structure that could occur. A Bayesian approach was used to develop the predictive models, which were based on thin-plate cubic splines and did not assume normality for the errors. The analyses applied to the training studies yielded reasonable predictions for the validation study, and some useful lessons for future practice.

email: goulda@merck.com

A BAYESIAN APPROACH TO CLUSTERING TRAJECTORIES IN THE PRESENCE OF MULTIPLE CHANGEPOINTS

Pulak Ghosh, Georgia State University
Kaushik Ghosh*, University of Nevada-Las Vegas
Ram C. Tiwari, National Cancer Institute

In longitudinal studies of patients with the Human Immunodeficiency Virus (HIV), objectives of interest often include modeling of individual-level trajectories of HIV Ribonucleic Acid (RNA) as a function of time. Empirical evidence suggests that individual trajectories often possess multiple points of rapid change, which may vary from subject to subject, both in number and in location. Presence of such changepoints make the modeling of individual viral RNA levels difficult and usual methods become unsuitable. In this article, we develop a new robust multiple-changepoint model which satisfactorily addresses the above issues. The proposed method uses a joint model to incorporate information from the longitudinal data as well as those from informative dropouts, which are common in such studies. A Dirichlet process prior is used to model the distribution of the changepoints of individual subject trajectories. The inherent clustering property of Dirichlet processes leads to natural clustering of subjects with similar changepoint experience. The possibility of unequal number changepoints is handled by using a prior distribution on the number of changepoints. A fully Bayesian approach is used for model fitting and prediction and is implemented using the Gibbs sampler. The proposed method is illustrated using data from the ACTG 398 clinical trial.

email: kaushik.ghosh@unlv.edu

HIERARCHICAL MODELS FOR SPATIALLY-REFERENCED LONGITUDINAL DATA WITH APPLICATIONS TO CROP GROWTH

Haowen Cai*, Sudipto Banerjee
Ryan T. Thelemann, University of Minnesota
Gregg A. Johnson, University of Minnesota

Modelling of longitudinal data from agricultural experiments using growth curves helps understand crop growth, in particular the impact of environmental conditions and spatial variation. Designs for carrying out such field experiments involve spatially arranged plots and replicates whose GIS coordinates are also available. The objectives of such studies include determining differences in plant growth and development between a diverse set of species as a function of landscape position, time and environmental conditions to better understand the relationship between plant growth and the environment. We explore different spatiotemporal models to capture the effect of time, space and enviromental conditions. We adopt a flexible hierarchical Bayesian framework and discuss two broad classes: the first uses spatiotemporal splines for modelling growth, while the second models growth curves as realizations of a spatial process. We also explore how these frameworks accommodate joint modelling of different crop-species that recognize dependence between these species. We compare the different approaches and report on the relative merits of each. We apply our models to plant growth data collected from Waseca, Minnesota.

email: caihaowen@gmail.com

INCORPORATING LIBIDO INTO HUMAN FECUNDABILITY MODELS

Kirsten Lum*, National Institutes of Health and American University
Rajeshwari Sundaram, National Institutes of Health
Germaine Buck Louis, National Institutes of Health

Time to pregnancy, typically defined as the number of menstrual cycles required to achieve a clinical pregnancy, is widely used as a measure of couple fecundability in epidemiologic studies. Until recently, these studies did not incorporate information on the timing and frequency of intercourse and the timing of ovulation in the fertile window, the period in each menstrual cycle during which sexual intercourse can result in conception. Dunson (2003, 2007) has shown that the frequency and timing of intercourse can have a large impact on probability of conception in a menstrual cycle, and consequently on time to pregnancy. It is also known that intercourse behavior in couples varies greatly, part of which is influenced by the libido of the couple (Wilcox, 2004). We incorporate this by jointly modeling the libido and intercourse behavior using association models for binary correlated data, where the dependency structure is also modeled. Consequently, we incorporate the libido effect and jointly model the intercourse behavior and the biologic fecundability, measured through probability of conception in a menstrual cycle. This is achieved using a likelihood based approach. The proposed methods are applied to the NY State Angler Cohort study in which couples are attempting to get pregnant.

email: lumkirst@mail.nih.gov

JOINT MODELING OF CLUSTER SIZE AND CLUSTERED FAILURE TIMES

Hanna Yoo*, Korea University
Yang-Jin Kim, Korea University
Jae Won Lee, Korea University
Shin-Jae Lee, Korea University

In this article, joint models of clustered failure times and cluster size are considered. In clustered failure times, the number of subjects within cluster may be correlated with their failure times, termed as informative cluster size. Dunson et al. (2003, Biometrics) suggested a bayesian method for joint model of cluster size and clustered response variables without censoring. We extended their approach to the clustered failure time data by incorporating common latent variables. More concretely, marginal regression model and ordinal regression model are applied and a random effect is adopted to connect two models. Simulation studies show that our methods produce unbiased parameter estimation in the presence of informative cluster size. For the analysis of real data, the clustered failure time from a lymphatic filariasis study is analyzed.

email: pinkcan78@korea.ac.kr

SEMIPARAMETRIC ANALYSIS OF PANEL COUNT DATA WITH CORRELATED OBSERVATION AND FOLLOW-UP TIMES

Xin He*, The Ohio State University
Xingwei Tong, Beijing Normal University-Peoples Republic of China
Jianguo Sun, University of Missouri

This paper discusses regression analysis of panel count data that often arise in longitudinal studies concerning occurrence rates of certain recurrent events. Panel count data mean that each study subject is observed only at discrete time points rather than under continuous observation. Furthermore, both observation and follow-up times can vary from subject to subject and may be correlated with the recurrent events. For inference, we propose some shared frailty models and estimating equations are developed for estimation of regression parameters. The proposed estimates are consistent and have asymptotically a normal distribution. The finite sample properties of the proposed estimates are investigated through simulation and an illustrative example from a cancer study is provided.

email: xhe@cph.osu.edu

LATENT CLASS TRANSITION MODELS FOR CHRONIC DISABILITY SURVEY DATA WITH CENSORING AND STAGGERED ENTRY

Toby A. White*, University of Washington

Latent class transition models are used to partition a population into a small number of relatively homogeneous subgroups so that the movement of individuals among these subgroups can be followed through time. One context for these models involves the U.S. elderly chronically disabled, who may be grouped into one of 4-5 disability classes which differ by both type and severity of disability. Such data appear in longitudinal surveys, which can have large assessment intervals, considerable right and left censoring, and staggered entry and exit. Thus, methodology is needed to account for all the possible time sequences at which individuals can be observed, since traditional latent class transition models assume a complete set of observations for each individual. I develop a group-based modeling approach that encompasses various time sequences of observation, and use the E-M algorithm with adjustments to estimate model parameters and parameter standard errors. I also extend basic latent class transition models to incorporate age, period, and cohort effects, while satisfying identifiability constraints. I illustrate this methodology using ADL and IADL data from the National Long-Term Care Survey (1982-2004), and discuss transition probability estimates among classes of varying disability level and death.

email: toby@stat.washington.edu

67. APPLICATIONS IN CAUSAL INFERENCE

IDENTIFICATION OF CAUSAL TREATMENT EFFECTS USING REFERENCE STRATIFICATION

Booil Jo*, Stanford University

This paper introduces the method of reference stratification (RS), which is a practical way of identifying causal treatment effects in randomized trials conditional on posttreatment (intermediate or final) outcome variables. Reference stratification refers to classification of individuals according to their potential posttreatment outcome values under only one treatment condition (i.e., reference condition), whereas principal stratification (PS: Frangakis & Rubin, 2002) refers to classification according to all compared conditions. Each set of RS results in one latent class variable, which consists of reference strata. This process may end after one round of stratification, or may repeat to formulate multiple latent class variables that correspond to multiple treatment conditions. On the basis of a combination of membership in multiple latent class variables, each individual can be categorized into finer strata such as principal strata. The resulting causal effects require less identifying assumptions (compared to causal effects based on PS), and therefore it is relatively easy to identify mixing proportions and to estimate causal treatment effects with reasonable precision. Reference stratification can be used as a stand-alone method or as a way of facilitating identification/estimation of models based on PS.

email: booil@stanford.edu

ESTIMATION AND INFERENCE FOR THE CAUSAL EFFECT OF RECEIVING TREATMENT ON A MULTINOMIAL OUTCOME

Jing Cheng*, University of Florida College of Medicine

This paper considers the analysis of two-arm randomized trials with noncompliance which have multinomial outcomes. We first define the causal effect in these trials as some function of outcome distributions of compliers with and without treatment, e.g., the Complier Average Causal Effect (CACE) and the measure of stochastic superiority of treatment over control for compliers, then estimate the causal effect with the likelihood method. Next we test those functions of or the equality of the outcome distributions of compliers with and without treatment based on the likelihood ratio statistic. Although the corresponding likelihood ratio statistic follows a chi-squared distribution asymptotically when the true values of parameters are in the interior of the parameter space under the null, its asymptotic distribution is not chi-squared when the true values of parameters are on the boundary of the parameter space under the null. Therefore, we propose a bootstrap/double bootstrap version of a likelihood ratio test for the causal effect in these trials. The methods are illustrated by an analysis of data from a randomized trial of an encouragement intervention to improve adherence to prescribed depression treatments among depressed elderly patients in primary care practices.

email: jcheng@biostat.ufl.edu

ESTIMATION OF TREATMENT EFFICACY IN RANDOMIZED CLINICAL TRIALS WHICH INVOLVE NON-TRIAL DEPARTURES

Tingting Ge*, University of Southern California
Stanley P. Azen, University of Southern California

Motivated by a real clinical trial, we consider the problem of estimating the causal treatment effect of a two-arm randomized controlled trial in which some of the participants went for a third treatment outside of the protocol. Following Rubin's potential outcome model approach, we classified the study sample into nine subgroups according to their potential compliance behaviors under each assignment. Under two alternative sets of assumptions outlined, a relative risk estimator for binary outcomes is proposed and estimated for the subgroups of participants identified as providing information to the causal estimation. Asymptotic performance of the proposed estimator is evaluated both theoretically and through simulation. Our approach is compared with traditional approaches including intent-to-treat, per-protocol, as-treated and instrumental variable analyses. Results show our proposed estimator is asymptotically unbiased and thus gives better estimate of the true effect of a treatment for the subpopulation than other estimators. Illustration of application to a real dataset is also presented.

email: tge@usc.edu

BOUNDS ON ATE AND UNMEASURED CONFOUNDING BIAS IN OBSERVATIONAL STUDIES

Tao liu*, Brown University
Joseph W. Hogan, Brown University
Allison K. DeLong, Brown University

We investigate two issues that are commonly encountered in observational studies: 1) The estimation of the average causal effect (ACE) of an exposure/treatment, and 2) quantification of the magnitude of unmeasured confounding. We use the method of instrumental variables (IV), combined with contextually appropriate (but sometimes untestable) assumptions, to place and tighten bounds on ACE. We account for measured confounding via inverse probability weighting (IPW) and incorporate a sensitivity parameter into the IPW estimating equation to account for unmeasured confounding. The causal bounds of ACE from IV allow us to infer the plausible range of the sensitivity parameter, and hence the magnitude of unmeasured confounding. We applied our methods to the HIV Epidemiology Research Study, where one of the study's primary interests is the initial stage ACE of highly active antiretroviral therapy on CD4 count among HIV positive women, and the existence of unmeasured confounding is of concern.

email: tliu@stat.brown.edu

DIRECT AND INDIRECT EFFECTS FOR CLUSTERED AND LONGITUDINAL DATA

Tyler J. VanderWeele*, University of Chicago

Definitions of direct and indirect effects are given for settings in which individuals are clustered in groups or neighborhoods and in which treatments are administered at the group level. A particular intervention may affect individual outcomes both through its effect on the individual and by changing the group or neighborhood itself. Identification conditions are given for controlled direct effects and for natural direct and indirect effects. The interpretation of these identification conditions are discussed within the context of neighborhood research and multilevel modeling. Interventions at a single point in time and time-varying interventions are both considered. The definition of direct and indirect effects requires certain stability conditions; some discussion is given as to how these stability conditions can be relaxed.

email: vanderweele@uchicago.edu

CORRECTING FOR SURVIVOR TREATMENT-SELECTION BIAS WITH A STRUTURAL FAILURE TIME MODEL:SURVIVAL IN OSCAR AWARD WINNING PERFORMERS

Xu Han*, University of Pennsylvania
Dylan Small, University of Pennsylvania
Dean Foster, University of Pennsylvania
Vishal Patel, University of Pennsylvania

We study the causal effect of winning an Oscar Award on an actor or actresses' survival. Does the increase in social rank from a performer winning an Oscar increase the performer's life expectancy? Previous studies of this issue have sufferec from survivor treatment selection bias, that is, candidates will have more chance to win Oscar Awards if they live longer, and winning Oscar Awards at a certain age is also an indicator of health status. To correct this bias, we adapt Robins' rank perserving structural failure time model and G-estimation method. We show in simulation studies that our approach corrects the survivor treatment selection bias contained in other approaches and apply our model to the Oscar data.

email: hanxu3@wharton.upenn.edu

A SYSTEMATIC REVIEW OF PROPENSITY-SCORE MATCHING IN THE MEDICAL LITERATURE FROM 1996 TO 2003

Peter C. Austin*, Institute for Clinical Evaluative Sciences

Propensity score methods are increasingly being used in observational studies to estimate the effects of treatments and interventions on health outcomes. Propensity-score matching is frequently used in the medical literature, particularly in the cardiovascular literature. However, propensity-score matching requires that appropriate statistical methods be used, both for assessing balance in baseline covariates between treated and untreated subjects in the matched sample, and for assessing the statistical significance of the treatment effect in the matched sample. I evaluated the appropriateness of statistical methods employed in 47 studies that employed propensity-score matching and that were published between 1996 and 2003 in the medical literature. Of the 47 studies evaluated, only 2 conducted all statistical analyses appropriately. I conclude with suggestions for the design and analysis of studies that employ propensity-score matching.

email: peter.austin@ices.on.ca

68. PROTEOMICS AND GENOMICS

FINDING THE OPTIMUM NUMBER OF PROBES TO INTERROGATE TRANSCRIPTS USING AFFYMETRIX GENECHIP TECHNOLOGY

Fenghai Duan*, College of Public Health University of Nebraska Medical Center

Affymetrix GeneChip is one of the most frequently used technologies in biomedical research. It is characterized by representing each transcript with multiple 25-mer probes. The number of probes per transcript varies from 4 to 16 according to the type of Affymetrix array. To our best knowledge, there were no studies that systematically investigated the optimum number of probes to interrogate transcripts using Affymetrix GeneChip technology. To explore this issue, we took advantage of similarities between rhesus macaque transcripts and human probes from the Affymetrix U133 GeneChip and extracted a larger amount of probes with various numbers per transcript. After summarizing the data using the probe dependent nearest neighbors approach (PDNN), we compared the probeset signals to the ones that were obtained when rhesus samples from the same sources were hybridized with the rhesus GeneChip. We hope our study can provide valuable information as to the optimization of probe selection and array design for Affymetrix GeneChip technology.

email: fduan@unmc.edu

COMPARISON OF RESAMPLING METHODS ON VALIDATION OF THE PROCEDURE FOR CONSTRUCTING A SURVIVAL PREDICTION MODEL FOR PROTEOMIC STUDIES

Heidi Chen*, Vanderbilt University
Ming Li, Vanderbilt University
Dean Billheimer, University of Utah
Yu Shyr, Vanderbilt University

MALDI-TOF mass spectrometry is one of the leading techniques inproteomics research. It allows direct measurement of the protein signatures of tissue, blood and other biological samples. One goal of studies using mass spectometey data is to build a prediction model for future outcomes based on the features extracted from such data. In general, there are three steps to building a prediction model: feature selection, model building, and model validation. While both external and internal validations are common methods for model validation, in this research we focus on internal validation to assess the predictive power of a prediction model for survival outcomes. Measurement of predictive accuracy can be difficult for survival data in the presence of censoring. To counter this, the C-index measures the probability of concordance (agreement) between the predicted and observed outcomes in terms of lengths of survival for any two subjects; it is a generalization of the area under the Receiver Operating Characteristics (ROC) curve. We use the C-index to measure predictive accuracy. With a focus on prediction assessment, we conducted Monte Carlo simulation studies to compare several resampling methods such as split sample, Bootstrap and k-fold cross validation of the estimation of true predictive accuracy.

email: heidi.chen@vanderbilt.edu

QUALITY ASSESSMENT AND REPRODUCIBILITY ANALYSIS OF MASS SPECTROMETRY DATA

Shuo Chen*, Emory University

As we are dealing with more and more high-throughput biomedical data,we notice that any inference, such as biomarker discovery and pattern prediction should be built on the raw data with valid quality and high reproducibility. In our recent research we are trying to find metrics to measure the quality and evaluate the reproducibility of the mass spectrum data for proteomic study. In practice, not only the variation resources from experiment procedure, but also the data quantification methods will affect the data quality and reproducibility. We used several MALDI-TOF MS data sets to demonstrate the performance metrics to measure QA and reproducibility as well as how different quantification methods could affect the results. The QA and reproducibility of MALDI-TOF MS data could evaluate validity of findings from such data sets; furthermore with high reproducibility, we would assess the quality of biological samples that will also benefit the more complex and time-consuming "tandem mass" proteomic experiment.

email: hitcx@hotmail.com

A NOVEL WAVELET-BASED APPROACH FOR THE PREPROCESSING OF MASS SPECTROMETRY DATA

Deukwoo Kwon*, National Cancer Institute
Joon Jin Song, University of Arkansas
Jaesik Jeong, Texas A&M University
Ruth M. Pfeiffer, National Cancer Institute
Marina Vannucci, Rice University

In recent years there has been an increased interest in using protein mass spectroscopy to discriminate diseased from healthy individuals with the aim of discovering molecular markers for disease. A crucial step before any statistical analysis is pre-processing of the data. Statistical results are typically strongly affected by the specific pre-processing technique used, and no firm guidelines exist. One important pre-processing step is the removal of noise from the mass spectral data. Wavelet denoising techniques are a standard method for denoising. Existing techniques, however, do not accommodate errors that vary across the mass spectrum but rather assume homogeneous error structure. In this paper we propose a novel wavelet denoising method that deals with heterogeneous errors by incorporating a variance change point detection method in the thresholding procedure. We use real and simulated MS data and show that how our local wavelet thresholding improves on performances of peak detection methods.

email: kwonde@mail.nih.gov

TESTING FOR TREATMENT EFFECTS ON GENE ONTOLOGY

Taewon Lee*, National Center for Toxicological Research
Varsha G. Desai, National Center for Toxicological Research
Robert R. Delongchamp, University of Arkansas
Cruz Velasco, Louisiana State University

In studies that use DNA arrays to assess changes in gene expression, it is preferable to measure the significance of treatment effects on group of genes under various pathways and functional categories such as gene ontology terms (GO terms, http://www.geneontology.org) for its benefits in the interpretation of data. Most statistical methods proposed on this purpose assume the independence between the univariate statistical tests for treatment effects on each gene or are applicable to only simple design of studies. A modified meta-analysis method to combine p-values was developed to measure the significance of an overall treatment effect on the group of genes. It takes into the correlation structure among genes. The method was used to distinguish altered pathways in a microarray study that uses a mitochondrial-specific oligonucleotide microarray, Mitochip. An application to Mitochip data shows that the developed method has more specificity in choosing altered pathways by incorporating the correlation among genes in a pathway. The test can be applied to the experimental data under a fixed effect linear model which is very common design to measure treatment effects. It can be applied to the data from the design which has further complicated structures to manage technical variation associated probes, dyes, and batches by blocking treatments within these sources of variation.

email: taewon.lee@fda.hhs.gov

A MODEL FOR THE ANALYSIS OF PROTEOLYTIC 18O STABLE-ISOTOPE LABELED PEPTIDES IN MALDI-TOF MASS-SPECTRA

Dirk Valkenborg, Hasselt University-Belgium
Tomasz Burzykowski*, Hasselt University-Belgium

Random noise introduced by chromatography and mass spectrometry makes it difficult to compare proteomic profiles between mass spectra. Therefore, stable-isotope coding (cfr. ICAT, iTRAQ, etc.) is often used, so that peptides from distinct samples are pooled together and analyzed simultaneously in a single mass spectrum, making a direct comparison possible (in terms of, e.g, the ratio of measured total intensities). A relatively new technique for stable-isotope coding is the proteolytic 18O labeling. In this setting, oxygen atoms from the carboxyl-terminus (COO-) of peptides are replaced with oxygens from heavy-oxygen-water during proteolysis. However, variable 18O incorporation rate, back-exchange, and impurities (presence of 16O and 17O atoms) in the heavy-oxygen-water result in multiple isotopic distributions from the labeled peptide, which superimpose with the unlabeled peptide. This complicates determination of the intensity ratio from the observed spectrum. We propose a model, similar in spirit to that developed by Eckel-Passow et al. (2007), which estimates the isotopic distribution, incorporation rate, and the intensity ratio directly from the observed mixture of peptide peaks. The model requires the knowledge of the composition of the heavy-oxygen water. The performance of the method is illustrated using real-life datasets.

email: tomasz.burzykowski@uhasselt.be

BACKGROUND CORRECTION, NORMALIZATION AND SUMMARIES OF BEAD LEVEL EXPRESSION DATA FROM ILLUMINA BEADARRAY

Yang Xie*, Simmons Cancer Center, University of Texas Southwestern Medical Center
Guanghua Xiao, University of Texas Southwestern Medical Center
Lianghao Ding, Simmons Cancer Center, University of Texas Southwestern Medical Center
Jeff Allen, Southern Methodist University
Michael D. Story, Simmons Cancer Center, University of Texas Southwestern Medical Center

The purpose of preprocessing microarray data is to control the effects of systematic error while retaining full biological variation. This is a critical step to obtain valid results. Due to different noise sources, the data preprocess step is platform specific. Illumina Beadarray became a very popular microarray platform due to its high quality and relatively low cost. However, no statistical methods have been formally developed to preprocess the beadarray data. In this talk, I propose statistical methods for background correction, normalization and measurement summarization for Illumina beadarray. Both the simulation and real data examples show the good performance of the proposed methods.

email: yang.xie@utsouthwestern.edu

69. SAMPLE SIZE

SAMPLE SIZES FOR PILOT STUDIES

Rickey E. Carter*, Medical University of South Carolina
Robert F. Woolson, Medical University of South Carolina

Pilot studies have objectives centered on feasibility, refinement of study procedures, estimation of primary endpoint variation, and on generating preliminary data for subsequent larger scale studies. Recently, the way some investigators use pilot data for planning purposes (and in grant proposals) was called into question (Kraemer et al., 2006). Their caution focused on the unreliability of pilot data for effect size estimation. The strength of their arguments against relying on pilot studies for effect size estimation stems from the concept of a clinically relevant difference; this quantity can not be obtained from a pilot study, as it is a clinical, not a statistical, notion. Nonetheless, one of the major objectives of pilots remains the generation of initial estimates of key parameters including means, standard deviations, and proportions. There are common features of the precision of estimators that can be displayed to offer guidance for the pilot study sample size. Useful expressions are developed and displayed for a variety of study designs including single arm studies, two arm studies, and study designs that implement repeated measures. Our recommendations rest on fundamental confidence interval principles and they fundamentally mirror an idea described by Van Belle (2002).

email: carterre@musc.edu

EVALUATION OF TEN EVENTS PER COVARIATE RECOMMENDATION FOR COX PROPORTIONAL HAZARDS REGRESSION MODELS

Mehmet Kocak*, St. Jude Children's Research Hospital
Arzu Onar, St. Jude Children's Research Hospital
Seok P. Wong, University of Memphis

Cox proportional hazards (PH) models are commonly utilized in medical research to investigate the associations between covariates and survival-type outcomes. Several authors have recommended that for less than ten events per covariate, these models produce spurious results, and therefore, not recommended. In the context of Phase-I and Phase-II clinical trials, as well as many of retrospective studies, the number of events is generally too small to allow a multivariable Cox PH model based on this recommendation. Our investigations aimed at confirming or challenging the 'minimum of ten events per covariate' recommendation for Cox PH models in the case of a single covariate. We conducted an extensive simulation study with various scenarios, where number of events and sample size were varied for a continuous and a dichotomous covariate. Empirical powers from those scenarios were then compared with Schoenfeld's (6) and Hsieh's (7) formulae. Our simulations indicated that in univariable models ten events may not be adequate to capture the desired effect. Furthermore, we illustrate that the number of events suggested by Schoenfeld's (6) and Hsieh's (7) formulae are often too small in these small-sample cases. Imbalance of the number of events at each level of a dichotomous covariate was also found to affect the empirical power substantially.

email: mehmet.kocak@stjude.org

DESIGN AND SAMPLE SIZE FOR EVALUATING COMBINATIONS OF DRUGS OF LINEAR AND LOGLINEAR DOSE RESPONSE CURVES

Hongbin Fang*, University of Maryland-Greenebaum Cancer Center
Guo-Liang Tian, University of Maryland-Greenebaum Cancer Center
Wei Li, University of Tennessee at Memphis
Ming Tan, University of Maryland-Greenebaum Cancer Center

The study of drug combinations has become important in drug development due to its potential for efficacy at lower less toxic doses and the need to move new therapies rapidly into clinical trials. Motivated by two anticancer combination studies that we are involved, this article proposes dose-finding and sample size method for detecting departures from additivity of two drugs with linear and log-linear single dose-response curves. The first study involves combination of two drugs where one single drug dose-response curve is linear and the other is log-linear. The second study involves combination of drugs whose single drug dose-response curves are linear. The experiment had been planned with the common fixed ratio design before we were consulted but the resulting data missed the synergistic combinations. In contrast, experiment based on the proposed design was able to identify the synergistic combinations as anticipated. Thus we shall summarize the analysis of the data collected according to the proposed design and discuss why the commonly used fixed ratio method failed and the implications of the method for other combination studies.

email: hfang@som.umaryland.edu

SAMPLE SIZE FORMULA FOR TIME-TO-EVENT DATA WITH INCOMPLETE EVENT ADJUDICATION USING GENERALIZED LOGRANK STATISTIC

Shanhong Guan*, Merck & Co.
Michael Kosorok, University of North Carolina-Chapel Hill
Tom Cook, University of Wisconsin-Madison

Event classification committees (ECC) are used routinely to adjudicate suspected end-points in cardiology clinical trials. When interim analysis are performed in such trials, the final classification for many reported events will not be known, and hence using ECC may introduce delay in adjudication. Cook and Kosorok (2004) proposed a generalized logrank statistic for analyzing time-to-event data with incomplete adjudication. Their results suggest that it may be unnecessary to adjudicate all of the events in a trial. In this research, under the assumption that only a fraction, q, of randomly selected subjects whose events will be adjudicated, we derive the theoretical formula for the variance of the weighted logrank statistic under a general class of local alternatives in which the conditional distribution of time to failure are assumed to vary in the two treatment arms. Using this result, a sample size formula that takes into account adjudication process is derived. Numerical illustrations are provided to evaluate the variance formula via simulating time-to-event data with incomplete adjudication. Numerical studies are conducted to evaluate the power of the logrank test for the time-to-event data with the number of subjects determined on the basis of sample size formula.

email: shanhong_guan@merck.com

INVERSE-PROBABILITY-WEIGHTING BASED SAMPLE SIZE FORMULA FOR TWO-STAGE ADAPTIVE TREATMENT STRATEGIES

Abdus S. Wahed*, University of Pittsburgh
Wentao Feng, University of Pittsburgh

In adaptive treatment strategies, patients are treated with various treatments in different stages of disease progression according to their response to the treatment in previous stages. Clinical trials are often designed to compare two or more adaptive treatment strategies. This article presents a sample size formula to compare the survival probabilities for different treatment strategies using Wald's test, when there are only two stages of treatments. The formula is based on the large sample properties of inverse-probability-weighted estimator. Simulation study provides strong evidence that the proposed sample size formula guarantees desired power, regardless of the true distributions of survival times.

email: wahed@pitt.edu

SAMPLE SIZE CALCULATION FOR A MIXTURE OF DISCRETE AND CONTINUOUS ENDPOINTS:AN APPLICATION OF THE POISSON-BINOMIAL DISTRIBUTION

Yolanda Munoz Maldonado*, University of Texas-Houston, School of Public Health
Sarah M. Baraniuk, University of Texas-Houston, School of Public Health
Lemuel A. Moye, University of Texas-Houston, School of Public Health

In this paper we address the problem of performing sample size calculations in studies where the question of interest involves multiple endpoints. The statistic of interest combines clinical events, which often have discrete distributions, and continuous outcomes. A Poisson-Binomial distribution is used in combination with the multivariate normal distribution to construct a composite endpoint and perform sample size and power analyses. We provide an example using clinical data and carry out a Monte Carlo simulation to investigate the critical values and power of the procedure and the sensitivity to some of the assumptions involved.

email: Yolanda.M.Munoz@uth.tmc.edu

SAMPLE SIZE CALCULATION FOR THOROUGH QT/QTC STUDY

Zhaoling Meng*, Sanofi-Aventis
Xun Chen, Sanofi-Aventis
Robert Kringle, Sanofi-Aventis
Peng-Liang Zhao, Sanofi-Aventis

Thorough QT/QTc study is routinely required in a new non-antiarrhythmic drug application to evaluate the drug's clinical electrocardiographic (ECG) effect on QT prolongation. The requirement for demonstration of no safety concern by ICH E14 is to achieve an upper bound of the 95% one-sided confidence interval for the largest time-matched mean effect of the drug on the QTc interval that is less than 10 ms. The corresponding study design is usually to measure QTc over a series of time points post-treatment, in either a parallel or crossover study, and then estimate the treatment difference from placebo at each time point, and conclude based on the maximum difference and its 95% upper CI. Sample size calculation following such a comparison routine is not trivial when considering the impacts of correlations of multiple time points and impacts of various treatment mean structure across time. In this paper, we will present a sample size calculation procedure for thorough QT/QTc study. The impacts of various treatment mean structure across time, correlation among time points and number of time points on power/sample size estimation are illustrated. Finally, the sample size calculation method is extended to when a bootstrap correction methodology (Boos et. al. 2007) is applied in the analysis.

email: zhaoling.meng@sanofi-aventis.com

70. IMAGING ANALYSIS

CORRELATION COEFFICIENT FROM DATA WITH MISSING COVARIATES: METHOD COMPARISON

Dana L. Tudorascu*, University of Pittsburgh
Lisa Weissfeld, University of Pittsburgh

The correlation between any two random variables can be estimated using a variety of techniques including parametric methods based on the Pearson correlation coefficient, nonparametric methods, and regression analysis. While these estimators have been widely used, the computation of these estimates in the presence of missing data has not been as widely studied. There has been some work addressing the estimation of parameters in the presence of missing data for regression analysis; including imputation, inverse probability weighted regression and weighted estimating equations. However, there has been little work focused on the estimation of the correlation coefficient. To assess the usefulness of these methods in a practical setting, we present a simulation study comparing imputation, inverse probability regression, weighted estimating equations and complete cases and provide recommendations on the basis of these results. We apply these results in a positron emission tomography data set consisting of five different brain regions as response variables and blood samples as covariates of interest, where the correlation matrix is used as input into a partial least squares analysis.

email: danatud@gmail.com

ADJUSTED EXPONENTIALLY TILTED EMPIRICAL LIKELIHOOD WITH APPLICATIONS TO NEUROIMAGING DATA

Xiaoyan Shi*, University of North Carolina at Chapel Hill
Hongtu Zhu, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Martin Styner, University of North Carolina at Chapel Hill

Anatomical and functional magnetic resonance images from cross-sectional and longitudinal studies have been widely used to understand normal brain development and the neural basis of neuropsychiatric disorders. The main objective of this paper is to develop an adjusted exponentially tilted empirical likelihood. We propose a nonparametric likelihood ratio statistic to test linear hypotheses of unknown parameters, such as associations between brain structure and function and covariates of interest. We also construct goodness-of-fit statistics for testing possible model misspecifications and apply them to the classification of time-dependent covariate types. The exponentially tilted empirical likelihood (ETEL) method is a nonparametric method, and thus it avoids standard parametric assumptions in general linear regression including that imaging data follow a Gaussian distribution. Our adjustment to ETEL method can dramatically improve its finite performance of the original ETEL. Our goodness-of-fit statistics overcome two important problems of many existing test statistics caused. Simulation studies are used to examine the finite performance of the adjusted ETEL ratio statistic and goodness-of-fit statistics. We demonstrate the application to a longitudinal schizophrenia study.

email: xyshi@email.unc.edu

FDR THRESHOLDING IN A NEUROIMAGING CONTEXT

Lynn E. Eberly*, University of Minnesota
Brian S. Caffo, Johns Hopkins University-Bloomberg School of Public Health

In functional neuroimaging (eg, fMRI, PET), voxel values in the brain map may represent clinically-based contrasts, such as activation in a disease group versus healthy controls. Scientific conclusions are based on identifying 'significant' voxels by thresholding: large voxel values are declared significant. Such thresholding is also done in gene expression (selecting genes with significantly elevated expression), discrete wavelet transforms (selecting significant wavelet coefficients), and many other areas. Numerous thresholding procedures are based on choosing a threshold so that the False Discovery Rate (FDR) is controlled. Because of the inherent spatial nature of neuroimage data, it it not clear whether procedures developed in other contexts will perform well in this setting. We compared the Benjamini-Hochberg and Storey procedures with other parametric and semi-parametric thresholding procedures developed within other contexts, and with general (not FDR specific) AIC and BIC thresholding procedures. We examined each procedure's estimated FDR, actual FDR, actual false non-discovery rate (FNR), power, and mean square error (MSE) for the contrast of interest. We found surprisingly good performance by the general BIC threshold, relatively poor performance by the parametric mixture methods, and widely varying MSE.

email: lynn@biostat.umn.edu

A BAYESIAN IMAGE ANALYSIS OF THE CHANGE IN TUMOR/BRAIN CONTRAST UPTAKE INDUCED BY RADIATION

Xiaoxi Zhang*, University of Michigan
Timothy D. Johnson, University of Michigan
Roderick J. A. Little, University of Michigan

This work is motivated by a quantitative Magnetic Resonance Imaging study of the change in tumor/brain contrast uptake induced by radiation. The results suggest the optimal timing of administering chemotherapy during the course of radiotherapy. A notable feature of the data is spatial heterogeneity. Since the tumor is physiologically and pathologically distinct from surrounding healthy tissue. Also, the tumor itself is typically highly heterogeneous. We employ a Gaussian hidden Markov random field model that respects the above features. The model introduces a latent layer of discrete labels from a Markov random field (MRF) governed by a spatial regularization parameter. Conditional on the hidden labels, the observed data are assumed independent and normally distributed, we treat the regularization parameter of the MRF, as well as the number of states of the MRF, as parameters, and estimate them via a reversible jump Markov chain Monte Carlo algorithm with a novel and nontrivial implementation. The performance of the method is examined in both simulation studies and a real data analysis. We report the pixel-wise posterior mean and standard deviation of the change in contrast uptake marginalized over the number of states and hidden labels. We also compare the performance with a parallel Expectation Maximization algorithm for maximum likelihood estimation.

e-mail: xiaoxi@umich.edu

BAYESIAN SPATIAL MODELING OF FMRI DATA: A MULTIPLE-SUBJECT ANALYSIS

Lei Xu*, University of Michigan
Timothy D. Johnson, University of Michigan
Thomas E. Nichols, University of Michigan

The aim of this work is to develop a spatial model for multi-subject fMRI data. While there has been much work on univariate modeling of each voxel for single-subject and multi-subject data, and some work on spatial modeling for single-subject data, there has been no work on spatial models that explicitly account for intersubject variability in activation location. We use a Bayesian hierarchical spatial model to fit the data. At the first level we model "population centers" that mark the centers of regions of activation. For a given population center each subject may have zero or more associated "individual components". While most previous work uses Gaussian mixtures for the activation shape, we instead use Gaussian mixtures for the probability that a voxel belongs to an activated region. Our approach incorporates the unknown number of mixture components into the model as a parameter whose posterior distribution is estimated by reversible jump Markov Chain Monte Carlo. We demonstrate our method with a fMRI study of visual working memory and show dramatically better precision of localization with our method relative to the standard mass-univariate method. Although we are motivated by fMRI data, this model could easily be modified to handle other types of imaging data.

e-mail: leix@umich.edu

A NONPARAMETRIC MIXTURE MODEL FOR THE FMRI VISUAL FIELD MAP

Raymond G. Hoffmann*, Medical College of Wisconsin
Nicholas M. Pajewski, Medical College of Wisconsin
Edward A. Deyoe, Medical College of Wisconsin

The fMRI visual field map (VFM) is obtained by using a rotating-expanding visual target to identify the area of the retina that corresponds to activated visual cortex. The VFM provides a methodology to identify and locate problems in the visual system, such as those induced by surgery for epilepsy. Since fMRI data is (1) relative rather than absolute and (2) has a degree of noise that may mask the activation, identifying differences in fMRI VFMs requires a stochastic model that will differentiate changes in the underlying structure from differences due to imaging variability. The VFM produces a non-homogenous set of points on a disk that correspond to areas of the visual cortex. This map, like the retina, has more points in the center of the field, is non-isotropic and includes features like the blind spot (where the optic nerve masks the retina). A non-parametric mixture model, using a Dirichlet prior on a space of 2D density functions, will be used to model the VFM under experimental conditions where part of the visual field is masked by a circular wedge. The posterior probability of the difference in the models, will be used to quantify the probable location of the wedge.

e-mail: hoffmann@mcw.edu

USING SCIENTIFIC AND STATISTICAL SUFFICIENCY TO LIFT THE CURSE OF DIMENSIONALITY IN IMAGING

Yueh-Yun Chi*, University of Florida
Keith E. Muller, University of Florida

Medical image segmentation, which outlines anatomic structures, plays a key role in clinical diagnosis and radiation therapy. An experiment compared computerized automatic segmentation method to manual segmentation due to the promise of greatly reduced cost and time. A total of 12 pairs of kidney images were segmented by both methods, and the distance between each pair of segmentations of the same image was measured on a set of 2,562 surface points. We overcame the High Dimension, Low Sample Size (HDLSS) data structure in two steps: 1. We tabulated the empirical histogram of distances for all points separately for each of the 24 kidney images. 2. We used standard nonparametric and parametric density fitting and estimation methods to find an excellent fit for cube root of distance with a possibly truncated Gaussian. The example illustrates a strategy we recommend for lifting the curse of dimensionality. The success came from replacing the HDLSS data structure with a scientifically and statistically sufficient structure amenable to traditional statistical tools. Thinking outside the HDLSS box has appeal for other types of HDLSS data as well, such as microarray and chemical spectrography data.

e-mail: yychi@biostat.ufl.edu

71. SURVIVAL ANALYSIS - DEPENDENT CENSORING

‘SMOOTH’ INFERENCE FOR BIVARIATE SURVIVAL FUNCTIONS WITH ARBITRARILY CENSORED DATA

Lihua Tang*, North Carolina State University
Marie Davidian, North Carolina State University

In many situations, such as twin studies, matched pair studies, and studies of organ such as the eyes and kidneys, correlated, bivariate failure times are recorded. Based on a sample of possibly censored such failure times, an objective of analysis is to estimate the joint survival distribution. We propose an approach to estimation of a bivariate survival distribution under the assumption that the distribution has a density satisfying mild ‘smoothness’ conditions, and we approximate the smooth density by the ‘seminonparametric’ (SNP) representation of Gallant and Nychka (1987), which admits a ‘parametric’ form for the density depending on a known ‘kernel’ density and a tuning parameter that determines the degree of flexibility for capturing the true density. This representation facilitates likelihood-based inference on the survival distribution, and we choose the tuning parameter and ‘kernel’ using standard information criteria. Moreover, arbitrary censoring patterns may be accommodated straightforwardly. We illustrate the approach via simulations and by application to data from the Diabetic Retinopathy Study.

email: ltang@ncsu.edu

A CLASS OF SEMIPARAMETRIC MIXTURE CURE SURVIVAL MODELS WITH DEPENDENT CENSORING

Megan Othus*, Harvard University
Yi Li, Harvard University
Ram Tiwari, National Cancer Institute

Modern cancer treatments have substantially improved cure rates and have generated a great interest in and need for proper statistical tools to analyze survival data with non-negligible cure fractions. Data with cure fractions are often complicated by dependent censoring, and the analysis of this type of data typically involves untestable assumptions on the dependence of the censoring mechanism and survival times. Motivated by the analysis of the NCI SEER prostate cancer data, we propose a class of general semiparametric transformation cure models that allows for dependent censoring without making parametric assumptions on the dependence relationship. An inverse-censoring-probability re-weighting scheme is used to derive unbiased estimating equations that account for dependent censoring and do not require untestable assumptions about the dependence structure between the survival and censoring times. The proposed model is general and encompasses a number of common models for the latency survival function, including the proportional hazards model and the proportional odds model. To properly evaluate the variability of the parameter estimates, we employ a novel application of the weighted bootstrap and verify its utility via extensive simulations. We apply the proposed methods to the NCI SEER prostate cancer data set to investigate potential racial disparities in prostate cancer cures.

email: mothus@fas.harvard.edu

WEIGHTED LIKELIHOOD METHOD FOR GROUPED SURVIVAL DATA IN CASE-COHORT STUDIES, WITH APPLICATION TO HIV VACCINE TRIALS

Zhiguo Li*, University of Michigan
Peter Gilbert, Cancer Research Center

Fred Hutchinson, Cancer Research Center
Bin Nan, University of Michigan

Grouped failure time data arise often in biomedical studies. In a recent preventive HIV vaccine efficacy trial, immune responses generated by the vaccine were measured from a case-cohort sample of vaccine recipients, who were subsequently evaluated for the study endpoint of HIV infection at pre-specified follow-up visits. Gilbert et al. (2005) and Forthal et al. (2007) analyzed the association between the immune responses and HIV incidence with a Cox proportional hazards model, treating the HIV infection diagnosis time as a right censored random variable. The data, however, are of the form of grouped failure time data with case-cohort covariate sampling, and we propose an inverse selection probability weighted likelihood method for fitting the Cox model to these data. The method allows covariates to be time-dependent, and uses multiple imputation to accommodate covariate data that are missing at random. We establish asymptotic properties of the proposed estimators, and present simulation results showing their good finite sample performance. We apply the method to the HIV vaccine trial data, showing that higher antibody levels are associated with a lower hazard of HIV infection.

email: zhiguo@umich.edu

ANALYSIS OF TIME-TO-EVENT OUTCOMES IN NEONATAL CLINICAL TRIALS WITH TWIN BIRTHS

Michele L. Shaffer*, Penn State College of Medicine
Sasiprapa Hiriote, The Pennsylvania State University

In neonatal trials of pre-term and/or low-birth-weight infants twins may represent 10-20% of the study sample. Frailty models and proportional hazards regression with a robust sandwich variance estimate are common approaches for handling correlated time-to-event outcomes. However, the operating characteristics of these methods for mixes of correlated and independent data are not well established. Simulation studies were conducted to compare frailty models and proportional hazards regression models with a robust sandwich variance estimate to standard proportional hazards regression models to estimate the treatment effect in two-armed clinical trials. While overall frailty models showed the best performance, caution must be exercised as the interpretation of the parameters differs from the marginal models, unless the correlation within twin births is zero. Data from the National Institute of Child Health & Human Development Neonatal Research Network are used for illustration.

email: shaffer.michele@psu.edu

SEMIPARAMETRIC COPULA REGRESSION MODEL FOR CENSORED LIFETIME MEDICAL COST

Jing Qian*, Rollins School of Public Health-Emory University
Yijian Huang, Rollins School of Public Health-Emory University

The analysis of lifetime medical costs with censored data confronts several statistical challenges. One of the major concerns is that the censoring pattern on the cost scale is typically induced to be dependent. Moreover, the lifetime medical cost distribution is potentially nowhere identifiable. Currently available approaches either bypass this issue by estimating time-restricted medical costs or address the joint distribution of costs and survival times instead. However, lifetime medical cost distribution for defined group is a quantity of interest in health outcome research, and thus desirable. To this end, the article proposes a semiparametric copula regression model where the marginal lifetime medical cost distribution for given covariates becomes identifiable. An inference procedure is proposed to estimate the regression coefficients on lifetime medical cost and the normal copula association parameter simultaneously. The resulting estimators are shown to be consistent and asymptotically normal. The estimators for regression coefficients on cost scale are robust to the misspecification of the copula structure. The performance of the proposed method is evaluated through simulation studies, and illustrated with application to a lung cancer clinical trial.

email: jqian@emory.edu

RESOURCE-USE ANALYSES WITH DEATH AS A CONFOUNDING FACTOR

Benjamin L. Trzaskoma*, Genentech
Amy C. Rundle, Genentech
David R. Nelson, Eli Lilly & Company
Fang Xie, Novartis Vaccines-Genentech

Analyses of the impact of interventions on resource utilization (e.g., use of mechanical ventilators, vasopressors, renal replacement therapy) are essential in evaluating overall therapeutic impact, particularly in the ICU. Resource measures traditionally serve as important secondary endpoints, but are now surfacing as primary endpoints. However, in the ICU setting, mortality during the study period confounds the resource-use comparisons. Days on a ventilator, over a window of time, are shortened in treatment groups with a greater mortality risk, and vice versa. Measures such as Ventilator Free Days (VFDs), which are the total number of days both alive and off of the ventilator, attempts to avoid penalizing treatments that improve mortality. However, if a treatment prolongs life but has no impact on shortening the time of mechanical ventilation, VFDs will often be higher, even though there is no true effect on resource use. Another approach is to censor deaths in an analysis of time-to-resource discontinuation. This is adequate as long as censoring is non-informative. But this is rarely the case, as risk factors for mortality tend to be associated with continued ventilator use as well. We will compare these approaches to alternatives that include evaluation of competing risks and gatekeeping strategies under various scenarios.

email: nelsondr@lilly.com

A MIXTURE-MODEL APPROACH TO BIVARIATE HYBRID CENSORED SURVIVAL DATA

Suhong Zhang*, University of Iowa
Ying J. Zhang, University of Iowa
Kathryn Chaloner, University of Iowa
Jack T. Stapleton, University of Iowa

A mixture model approach for bivariate survival data is proposed to study the association between survival time of individuals infected with HIV and the persistence time of infection with an additional virus. Survival with HIV is right censored. Persistence time of the virus is interval censored and is modeled using a mixture distribution. The proposed model employs a copula structure and applies a pseudo likelihood approach to estimate the association between the two survival times. The asymptotic consistency and normality of the estimator are established, and simulation studies are conducted to examine its finite sample performance. The method is applied to a motivating example.

email: suhong-zhang@uiowa.edu

72. PRESIDENTIAL INVITED ADDRESS

BIOSTATISTICS, SCIENCE, AND THE PUBLIC EYE

Donald A. Berry, University of Texas M.D.-Anderson Cancer Center

PRESIDENTIAL ADDRESS: BIOSTATISTICS, SCIENCE, AND THE PUBLIC EYE

Donald A. Berry, Ph.D.

Statisticians hold the scepter of science. Too often we wield it in a supporting role. The public sees us as number people: we can multiply and divide and we know things like the altitude and volume of the world's highest lake. Many of our medical colleagues see us as reservoirs of sample size calculations--the statistician as technician. I fight this attitude, in defense of myself and in defense of our profession. I frequently use my scepter as a weapon. It leaves a mark. And I don't always escape unscathed. But the controversies that arise make life interesting ... and fun! For example, I've been quoted in The New York Times scores of times about such diverse issues as the risks and benefits of cancer screening, mad cow disease, death counts in Iraq, and whether dogs can sniff out cancer. I'll relate some of these stories, addressing statistics, science and politics along the way. While I don't aim to convert all statisticians to be controversy seekers, my goal is not only to educate, but also to encourage each of you to find your own unique way to lift your scepter.

email: dberry@mdanderson.org

73. JOINT MODELS FOR SURROGATE AND FINAL OUTCOMES

ASSESSING SURROGACY IN CLINICAL TRIALS USING COUNTERFACTUAL MODELS

Yun Li*, University of Michigan
Jeremy MG Taylor, University of Michigan
Michael R. Elliott, University of Michigan

A surrogate marker (S) is a variable that can often be measured earlier than the true endpoint (T) in a clinical trial. Previous research has been devoted to developing surrogacy measures to quantify how well S can replace T or examining the use of S in predicting the treatment (Z) effect. However, the research often requires one to fit models for the distribution of T given S and Z. It is known that they do not have causal interpretations because the models condition on a post-randomization variable S. In this paper, we directly model the relationship among T, S and Z in a causal framework, specifically using a potential outcomes framework introduced by Frangakis and Rubin (2002). We propose a Bayesian estimation method to evaluate the causal probabilities associated with the cross-classification of the potential outcomes of S and T when they are both binary. We use a log-linear model to model the associations of the potential outcomes. This causal model is not identifiable without additional assumptions. To reduce the non-identifiability problem and increase precision for the statistical inferences, we incorporate assumptions that are plausible in the surrogate context by using prior distributions. The methods are applied to glaucoma data.

email: yunlisph@umich.edu

THE META-ANALYTIC FRAMEWORK FOR THE EVALUATION OF SURROGATE ENDPOINTS IN CLINICAL TRIALS

Geert Molenberghs*, Hasselt University-Belgium

The validation of surrogate endpoints has been initially studied by Prentice and Freedman. Noting operational difficulties, Buyse and Molenberghs proposed instead to use jointly the within-treatment partial association of true and surrogate responses, and the treatment effect on the surrogate relative to that on the true outcome. In a multi-center setting, these quantities can be generalized to individual-level and trial-level measures of surrogacy. Buyse and colleagues have proposed a meta-analytic framework to study surrogacy at the trial and individual-patient levels. Variations for various endpoints have been developed. Efforts have been made to converge to a common framework. This includes a so-called variance reducation factor and an information-theoretic approach. Work has been done regarding sample size assessment, leading to the surrogate threshold effect.

email: geert.molenberghs@uhasselt.be

A UNIFIED FRAMEWORK FOR SURROGATE OUTCOMES AND MARKERS

Marshall M. Joffe*, University of Pennsylvania
Tom Greene, University of Utah

Four major frameworks have been developed for evaluating surrogate markers in randomized trials: one based on conditional independence of observable variables, a second based on direct and indirect effects, a third based on a meta-analysis, and a fourth based on principal stratification. We provide a unifying framework for all of the approaches. The meta-analytic approach may be reinterpreted in a way that allows its use with individual-level data within a single study. In this reinterpretation, the main criteria for a good surrogate are that the effect of treatment on the putative surrogate within subgroups of the study defined by baseline covariates should predict well the effect of the treatment on the clinical outcome of interest, and that the effect of treatment on the outcome in the subgroups in which there is no effect on the surrogate is zero. This definition is consistent with both the direct/indirect effects approach and with principal surrogacy, each of which can provide valuable motivation and interpretation, but is more generally applicable. The approach based on conditional independence provides misleading conclusions when applied to randomized trial data, but, when properly reinterpreted, leads to a generally useful measure of the degree of surrogacy of a variable.

email: mjoffe@mail.med.upenn.edu

74. STATISTICAL ANALYSIS OF LARGE-SCALE ENVIRONMENTAL DATASETS

CLOUD HEIGHT ESTIMATION BASED ON MULTI-ANGLE SATELLITE (MISR) IMAGES

Ethan B. Anderes*, University of California at Berkeley

Clouds play a major role in determining the Earth's energy budget. As a result, monitoring and characterizing the distribution of clouds becomes important in global studies of climate. In this talk, we will report our efforts on collaborating with the MISR team at JPL to develop more advanced techniques for cloud height estimation. By viewing the multi-angle cloud images as discrete sub-samples of a continuous random field, one can view the cloud-top height estimation problem as statistical parameter estimation. We apply this methodology to the problem of height recovery for a two layer could ensemble where both layers are relatively planer, the bottom layer is optically thick and textured, and the top layer is optically thin.

email: anderes@stat.berkeley.edu

NONSTATIONARY COVARIANCE MODELS FOR GLOBAL DATA

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

With the widespread availability of satellite-based instruments, many geophysical processes are measured on a global scale and they often show strong nonstationarity in the covariance structure. In this paper, we present a flexible class of parametric covariance models that can capture the nonstationarity in global data, especially strong dependency of covariance structure on latitudes. We apply the Discrete Fourier Transform to data on regular grids, which enables us to calculate exact likelihood for large data sets. Our covariance model is applied to global total column ozone level data on a given day. We discuss how our covariance model compares with some of the existing models.

email: mjun@stat.tamu.edu

INTEGRATING SATELLITE AND MONITORING DATA TO RETROSPECTIVELY ESTIMATE MONTHLY PM2.5 CONCENTRATIONS IN THE EASTERN UNITED STATES

Christopher J. Paciorek*, Harvard School of Public Health
Yang Liu, Harvard School of Public Health

Advances in spatial modeling and GIS technology, combined with the availability and demonstrated utility of satellite proxy data for particulate matter (PM) estimation, present the opportunity for integrated estimation of PM2.5 for use in health analyses of the chronic effects of PM. Bayesian statistical techniques provide a natural framework for the integration. I present a hierarchical Bayesian model that attempts to capture the key features of the available data through multiple likelihood terms, one for each AOD proxy and one for ground-level monitoring data, while accounting for the complicated spatial and temporal misalignment of the data sources. I focus on two key questions. First, how can we model the possibility of spatially-varying bias in AOD as a proxy for PM. Evidence suggests that the bias does vary spatially, which causes identifiability problems inherent in the structure of the data. I show that predictions of PM are very sensitive to the flexibility of the model term that represents spatially-varying bias. The second key question I address is whether including the AOD proxy materially improve predictions of ground-level PM beyond what can be achieved based on the PM data and various covariates.

email: paciorek@hsph.harvard.edu

DIMENSION REDUCTION APPROACHES FOR ANALYZING LARGE SPATIAL DATASETS

Alan E. Gelfand*, Duke University

Fitting hierarchical spatial models often involves expensive matrix decompositions whose computational complexity increases in cubic order with the number of spatial locations, rendering such models infeasible for large spatial data sets. This computational burden is exacerbated in multivariate settings with several spatially dependent response variables. It is also aggravated when data is collected at frequent time points and spatiotemporal process models are used. Dimension reduction approaches provide one strategy for addressing this problem. Here, we argue for the use of predictive process models for spatial and spatiotemporal data. Every spatial (or spatiotemporal) process induces a predictive process model (in fact, arbitrarily many of them). The latter models project process realizations of the former to a lower-dimensional subspace; we achieve the flexibility to accommodate nonstationary, non-Gaussian, possibly multivariate, possibly spatiotemporal processes in the context of large datasets. We illustrate with spatial modelling of forest biomass using a spatially varying coefficient model, resulting in 28,500 spatial random effects.

email: alan@stat.duke.edu

75. VARIABLE SELECTION AND DIMENSION REDUCTION IN GENOMICS

CLUSTERING WITH MULTIPLE DISTANCE METRICS - MIXTURE MODELS WITH PROFILE TRANSFORMATIONS

Rebecka J. Jornsten*, Rutgers University

Clustering methods often require the selection of a distance metric;how do we define data objects as ‘close’ enough to be grouped together, or ‘far’ enough apart to be separated? Choosing an appropriate distance metric is not always easy. We consider high-dimensional gene expression data as an example. The shape of a gene’s expression profile across experimental conditions is often considered to be the most informative, which translates to choosing correlation as a similarity metric. However, when genes with a similar expression profile exhibit expression differences on a scale of two-fold to ten-fold, correlation comparisons do not suffice, implying that a Euclidean distance metric is more appropriate. We propose a model-based clustering approach, MIX-T (MIXture modeling with profile Transformations), which incorporates multiple distance metrics simultaneously. The modeling framework constitutes a between-cluster parameterization, allowing for direct and objective cluster comparisons. With this more efficient parameterization, we detect clusters that a standard model-based clustering approach may miss. We demonstrate the utility of the MIX-T model via the analysis of a time-course gene expression data set, with two experimental factors, and discuss the biological relevance of the gene clusters identified.

email: rebecka@stat.rutgers.edu

SPARSE PARTIAL LEAST SQUARES REGRESSION WITH AN APPLICATION TO GENOME SCALE TRANSCRIPTION FACTOR ANALYSIS

Hyonho Chun*, University of Wisconsin-Madison
Sunduz Keles, University of Wisconsin-Madison

Analysis of modern biological data often involves ill-posed problems due to high dimensionality and multicollinearity. Partial Least Squares (PLS) regression has been an alternative to ordinary least squares for handling multicollinearity in several areas of scientific research since 1960s. At the core of the PLS methodology lies a dimension reduction technique coupled with a regression model. Although PLS regression has been shown to achieve good predictive performance, it is not particularly tailored for variable/feature selection and therefore often produces linear combinations of the original predictors that are hard

to interpret due to high dimensionality. In this paper, we investigate the known asymptotic properties of the PLS estimator under special case of normality and show that its consistency breaks down with the very large p and small n paradigm. We, then, propose a sparse partial least squares (SPLS) formulation which aims to simultaneously achieve good predictive performance and variable selection by producing sparse linear combinations of the original predictors. We provide an efficient implementation of SPLS based on the LARS algorithm. An additional advantage of the SPLS algorithm is that it naturally handles multivariate responses. We illustrate the methodology in a joint analysis of gene expression and genome-wide binding data.

email: chun@stat.wisc.edu

VARIABLE SELECTION FOR VARYING-COEFFICIENTS MODELS WITH APPLICATIONS TO ANALYSIS OF GENOMIC DATA

Hongzhe Li*, University of Pennsylvania
Lifeng Wang, University of Pennsylvania

We will present several examples from genomic data analysis where varying-coefficients models are more appropriate than the standard regression models with parametric linear forms, including the problem of identifying the transcription factors that are related to a given biological processes based on microarray time course gene expression data and the problem of studying the effects of gene expression levels on time to cancer recurrence. We will present a general regularized estimation procedure for variable selection for varying-coefficients models based on basis function expansions and a group version of the SCAD penalty. When the tuning parameter is appropriately selected, we can show that this procedure has the desirable oracle property for variable selection. In addition, we have also obtained the rates of convergence of the estimates of the nonparametric coefficients functions. We demonstrate the methods by simulations and analysis of two microarray gene expression data sets.

email: hongzhe@mail.med.upenn.edu

76. TARGETING TREATMENT EFFICACY: FLEXIBLE MODELING TO CLINICAL TRIAL DESIGN

A NOVEL DATA DRIVEN APPROACH TO STAGE GROUPING OF ESOPHAGEAL CANCER

Hemant Ishwaran*, Cleveland Clinic

A novel three-level random forest analysis involving random survival forests for survival data, multiclass forests for ordinal data, and regression forests for continuous data is developed and used to define a data driven TNM stage grouping for esophageal cancer. The analysis is applied to 4627 patients collected from a world-wide consortium effort focusing on esophageal cancer.

email: hemant.ishwaran@gmail.com

CAUSAL SUBGROUP MODELING IN CANCER CLINICAL TRIALS

Mary W. Redman*, Fred Hutchinson Cancer Research Center
Michael L. LeBlanc, Fred Hutchinson Cancer Research Center

In cancer clinical trials it is often of interest to assess the prognostic and possibly predictive value of biomarkers measured pre-and post-treatment. Conventional approaches to the estimation of the effects of variables affected by prior treatment may be confounded because the variable is intermediate (in the causal pathway) between the treatment and outcome of interest. Brumback, Greenland, Redman et al. (2003) introduced the intensity-score approach which can yield unbiased estimates in such a setting, employing a set of often plausible assumptions. Van der Laan (2006) recently proposed an approach to screen a large number of factors to determine key factors measured at a single point in time without having to fully parameterize a regression model to avoid over-fitting and the specification of function forms of associations between factors and outcomes that are not well understood. They briefly mention the use of structural nested models (such as the intensity-score approach) in this setting, but do not develop their use with time-dependent covariates. In this talk we present an approach which builds on the aforementioned approaches to deal with time-varying factors and non-binary factors. We then apply our methods to a lung cancer clinical trial from the Southwest Oncology group.

email: mredman@fhcrc.org

MOVING FROM CORRELATIVE CLINICAL SCIENCE TO PREDICTIVE MEDICINE IN CLINICAL TRIAL DESIGNS

Richard M. Simon*, National Cancer Institute

The genomics and biotechnology revolutions sweeping biology will influence clinical trials in important ways. Many of the entities currently treated in clinical trials will be recognized as distinct molecularly and unlikely to be responsive to the same treatments. The drugs that target the key oncogenic mutations can be expected to be effective against subsets of our current heterogeneous cancer categories. If we continue to treat broad patient populations with the new generations of drugs, we may fail to recognize effective drugs because the overall effects will be so diluted. It also leads to over-treatment of patients. This presentation will review the results of Simon and Maitournam on the efficiency of targeted enrichment trials in which eligibility is limited based on a baseline measurement of a predictive biomarker. We will also review randomized designs in which entry is not restricted to classifier positive patients, but in which a specific analysis plan involving the classifier is defined and the experiment-wise error is preserved. Adaptive designs described by Freidlin and Simon and Jiang, Freidlin and Simon will also be described.

email: rsimon@mail.nih.gov

77. STATISTICAL LEARNING IN BIOLOGICAL SIGNALS AND IMAGES

DETECTING COGNITIVE FATIGUE VIA MIXTURES OF AUTOREGRESSIVE MODELS

Raquel Prado*, University of California-Santa Cruz

Mental fatigue is one of the main causes of human performance failures that can lead to accidents in vehicle operation, air traffic control and space missions. Automatic detection of early signs of mental fatigue that can then trigger appropriate countermeasures is therefore key for increasing safety and human performance. EEG signals recorded in subjects who performed continuous mental arithmetic for a period of 90-180 minutes are studied. Tests confirmed that individuals were alert prior to the experiment and showed signs of severe fatigue after the experiment ended, however, there was no assessment of cognitive fatigue during the course of the experiment. We analyze portions of the EEGs from the first and last 15 minutes of the experiment. These analyses, based on autoregressive and time-varying autoregressive models for single channel data, suggest that changes in frontal theta and parietal alpha EEG rhythms over time may be associated with changes in the level of mental alertness. Based on these results we propose the use of mixtures of autoregressive processes with structured prior distributions for describing variation in mental alertness over time. Structured prior distributions are used to incorporate relevant scientific information whenever available. Algorithms for on-line posterior estimation are developed and illustrated.

email: raquel@ams.ucsc.edu

CLASSIFICATION OF BIOMEDICAL SIGNALS VIA MULTISCALE LOCAL STATIONARITY

Piotr Z. Fryzlewicz*, University of Bristol
Hernando Ombao, Brown University

Accurate classification of biomedical time series is of paramount importance in modern medical practice. Biomedical signals, such as those arising from measurements of brain activity, are often measured in an evolving environment and are thus best modelled as nonstationary. Amongst statistically rigorous nonstationary time series models, the Locally Stationary Wavelet (LSW) model differs from most other proposals in that it employs wavelets, rather than Fourier exponentials, as building blocks. This makes for excellent localisation properties of the resulting estimators of the time-varying spectral structure, their good theoretical and empirical performance, easy interpretability and rapid computability. We use the LSW model and its estimation machinery to develop a classification procedure whereby a time series is classified to either of two groups A or B of previously observed time series according to the distance of its evolutionary wavelet spectrum from the evolutionary spectra of A and B. We discuss various technical issues that arise, and demonstrate the excellent practical performance of our procedure.

email: p.z.fryzlewicz@bristol.ac.uk

SEMIPARAMETRIC LOGISTIC REGRESSION FOR GENETIC PATHWAY DATA: KERNEL MACHINES AND GENERALIZED LINEAR MIXED MODELS

Dawei Liu*, Brown University
Xihong Lin, Harvard University
Debashis Ghosh, The Pennsylvania State University

Growing interest on biological pathways has called for new statistical methods for modeling and testing the multi-dimensional pathway effect. In this talk, we propose a semiparametric logistic regression model for binary outcomes, where the clinical effects are modeled parametrically and the genetic pathway effect is modeled nonparametrically using kernel machines. The nonparametric function of a genetic pathway allows for the possibility that genes within the same pathway are likely to interact with each other and relate to the clinical outcome in a complicated way. We show that the kernel machine estimate can be formulated using a generalized linear mixed model. Estimation hence can proceed within the generalized linear mixed model framework using standard mixed model software. A score test based on a nonstationary Gaussian process is developed to test for the genetic pathway effect. The methods are illustrated using a prostate cancer data set and evaluated using simulations.

email: Dawei_Liu@brown.edu

PFCLUSTER AND FCA, AND THEIR APPLICATIONS TO PROFILE DATA ANALYSIS

Jiayang Sun*, Case Western Reserve University
Yaomin Xu, Case Western Reserve University and Cleveland Clinic Foundation

Properly clustering gene profiles collected under different conditions or from a longitudinal study is important in building regulatory networks and for finding pathway information for certain diseases. In this talk, we first present a new clustering technique called PfCluster for profile cluster analysis. The PfCluster is efficient and flexible. It can uncover clusters determined by any distance from a class of biologically meaningful ones. In addition, “Within” and “Between” coherence indices are developed to measure how coherent resulting clusters are. The null distributions of coherence indices and their approximate critical values are given. These indices provide some measure of the integrity of resulting groups, and an inferential procedure for deciding where to cut a dendrogram, often not at the same height for all branches. After clusters are found, we show how to adapt Formal Concept Analysis (FCA) to find associations of selected influential genes or important clusters. Applications to real data and simulation results will also be presented.

email: jsun@case.edu

78. NONPARAMETRIC REGRESSION FOR SURVIVAL ANALYSIS

PARTIALLY LINEAR HAZARD REGRESSION WITH VARYINGCOEFFICIENTS

Jianwen Cai*, University of North Carolina at Chapel Hill
Jianqing Fan, Princeton University
Jiancheng Jiang, University of North Carolina at Charlotte
Haibo Zhou, University of North Carolina at Chapel Hill

This talk will present estimation of partially linear hazard regression models with varying coefficients for multivariate survival data. A profile pseudo-partial likelihood estimation method is proposed. The estimation of the parameters of the linear part is accomplished via maximization of the profile pseudo-partial likelihood, while the varying-coefficient functions are considered as nuisance parameters profiled out of the likelihood. The estimators of the parameters are shown to be root-n consistent and the estimators of the nonparametric coefficient functions achieve optimal convergence rates. Asymptotic normality is obtained for the estimators of the finite parameters and varying-coefficient functions. Simulations are conducted to demonstrate the performance of the proposed estimators in finite samples. A real dataset is analyzed to illustrate the proposed methodology.

email: cai@bios.unc.edu

A PARTIAL LINEAR SEMIPARAMETRIC ADDITIVE RISKS MODEL FOR TWO-STAGE DESIGN SURVIVAL STUDIES

Gang Li, University of California-Los Angeles
Tong Tong Wu*, University of Maryland, College Park

In this talk we consider a semiparametric additive risks model (McKeague and Sasieni 1994, Biometrika pp. 501-514) for two-stage design survival data. The first stage data could be biased or missing and the second stage data is assumed to be accurate. We develop two-stage estimators by combining data from both stages. Large sample properties of the two-stage estimators are established and asymptotic inferences are developed. As a byproduct, we also obtain asymptotic properties of the single stage estimators of McKeague and Sasieni (1994) when the semiparametric additive risks model is misspecified. The proposed two-stage estimators, compared with the second stage estimators, are asymptotically more efficient. They also demonstrate smaller bias and variance for finite samples in a simulation study. The developed methods are illustrated using the small intestine cancer data from the SEER (Surveillance, Epidemiology, and End Results) Program supported by NCI.

email: ttwu@umd.edu

THRESHOLD REGRESSION MODELS WITH SEMIPARAMETRIC COVARIATE FUNCTION FOR SURVIVAL DATA ANALYSIS

Zhangsheng Yu*, The Ohio State University
Mei-Ling Ting Lee, The Ohio State University

We study a threshold regression model with a semiparametric covariate function for censored survival data. Both regression spline and local kernel method are considered for the nonparametric covariate function estimation. A cross validation method is proposed for the tuning parameter selection. Simulation result shows that the proposed estimates of nonparametric function and parametric coe±cient are close to true values. The proposed variance estimates also perform very well. We then apply the proposed regression spline based estimate to the western Kenya parasitemia study.

email: zyu@cph.osu.edu

79. CGH ARRAY AND COPY NUMBER

DNA COPY NUMBER ABNORMAL DETECTION BASED ON DECOMPOSITION OF INTENSITY MULTIMODAL DISTRIBUTION

Aixiang Jiang*, Vanderbilt University
Jirong Long, Vanderbilt University
Wei Zheng, Vanderbilt University
Yu Shyr, Vanderbilt University

The very first question for DNA copy number analysis is how to define copy number abnormal. Most of current approaches are based on log2 ratio data, which is the base 2 log transformation data of the ratio of testing sample over reference. This method is firstly used in CGH two channel arrays, and then extended to single channel arrays such as SNP when reference sample is well defined. It is simple and straightforward, but it is difficult to define a reasonable cutoff, and sometimes reference sample is not available or there are multiple reference samples mixed with possible abnormal samples. Since distribution of DNA intensity data is multimodal instead of unimodal, our current research is trying to define copy number abnormal based on decomposition of DNA intensity multimodal distribution. The decomposition methods used in our current research are EM, smooth spline, kernel density, and wavelet. Both simulation and real SNP data are tested.

email: aixiang.jiang@vanderbilt.edu

A BAYESIAN CHANGE-POINT ALGORITHM FOR THE ANALYSIS OF SNP-DATA

Fridtjof Thomas*, University of Texas, Health Science Center
Stanley Pounds, St. Jude Children's Research Hospital

Recent technical developments have made it possible to collect high-resolution genomics data using single nucleotide polymorphism (SNP) arrays. These arrays can be used in a paired data context to compare cancer tissue to normal samples in an effort to identify regions of genomic amplification or deletion. Such regions potentially contain oncogenes or tumor suppressor genes and are therefore of particular interest. However, using SNP array signals to identifying regions of copy number alteration is a challenging task due to the properties of the derived measurements. We apply here a Bayesian change-point algorithm to pre-normalized signals from SNP microarrays obtained from a set of leukemia samples in an effort to infer regions of copy number alteration. This algorithm detects multiple change-points where a change can be in the mean of the subsequent measurements, in their variance, in their autocorrelation structure, or in a combination of two or all of these aspects.

email: fthomas4@utmem.edu

TOLERANCE INTERVALS FROM PROBE-SPECIFIC MIXEDMODELS TO DETECT GAINS AND LOSES USING MULTIPLEX LIGATION-DEPENDENT PROBE AMPLIFICATION (MLPA)

Juan R. Gonzalez*, Center for Research in Environmental Epidemiology (Creal)-Barcelona, Spain
Josep L. Carrasco, University of Barcelona-Spain
Lluis Armengol, Center for Genomic Regulation,-Barcelona, Spain
Yutaka Yasui, University of Alberta-Canada

Copy number variations play an important role in genes and other regulatory elements that may have phenotypical consequences. Several techniques and platforms have been developed for genome-wide analysis of DNA copy number, such as array-based comparative genomic hybridization (aCGH). However, the ability of aCGH to discern between different number of copies is very limited. MLPA is a recent method that aims to detect copy number alterations at the genomic level (gains or loses) in a test DNA with respect to a reference. In this work, we propose a method for the normalization procedure based on a non-linear mixed-model, as well as a new approach for determining the statistical significance of altered probes based on linear mixed-model. This method establishes a threshold by using different tolerance intervals that accommodates the specific random error variability observed in each test sample. Through simulation studies we have shown that our proposed method outperforms the existing ones based on threshold rule or iterative regression. We illustrate the method using a controlled MLPA assay in which probes interrogate regions that are variable in copy number in individuals suffering from different genomic disorders such as Prader-Willi, DiGeorge or Autism.

email: jrgonzalez@imim.es

PATHWAY SCREENING ANALYSIS OF GENOMIC COPY NUMBER CHANGE DATA

Stanley B. Pounds*, St. Jude Children's Research Hospital

SNP and CGH microarrays are used to identify regions of genomic copy number alteration in tumor cell samples. Once these regions have been identified, it is of biological interest to determine whether these alterations focus on genes in predefined pathways of interest. A method is proposed to perform an analysis to address this particular question. This method models the number of affected genes of a particular pathway on each chromosome as an independent observation from a hypergeometric distribution. The model is used to derive analytical expressions for null distributions of the number of affected pathway genes in each subject and the number of subjects with at least one pathway gene affected. The method was used to screen 111 cases of acute myeloid leukemia for focused alterations of the genes in the KEGG-database pathways. The results provide proof-of-principle of the method's practical utility: the most significant p-value was obtained for the biologically most obvious match in the database, the pathway for chronic myeloid leukemia (a similar but less serious disease).

email: stanley.pounds@stjude.org

A LATENT CLASS MODEL WITH HIDDEN MARKOV DEPENDENCE FOR ARRAY CGH DATA

Stacia M. DeSantis*, Medical University of South Carolina
E.A. Houseman, University of Massachusetts-Lowell
Brent A. Coull, Harvard School of Public Health
David N. Louis, Massachusetts General Hospital
Gayatry Mohapatra, Massachusetts General Hospital
Rebecca A. Betensky, Harvard School of Public Health

Array CGH is a high-throughput technique designed to detect genomic alterations linked to the development and progression of cancer. The technique yields fluorescence ratios that characterize DNA copy number change in tumor versus healthy cells. Classification of tumors based on aCGH profiles is of scientific interest but the analysis of these data is complicated by the large number of highly correlated measures. In this paper, we develop a Bayesian latent class approach for classification that relies on a hidden Markov model to account for the dependence in the intensity ratios. Posterior inferences are made about class-specific copy number gains and losses. We demonstrate our technique on a study of brain tumors, for which our approach is capable of identifying subsets of tumors with different genomic profiles. The method can be used for supervised classification as well, when there is a clinical endpoint available for guidance.

email: desantis@musc.edu

A HYBRID METHOD FOR GENOMIC ALTERATION DETECTION IN CGH MICROARRAY DATA ANALYSIS

Ao Yuan, National Human Genome Center-Howard University
Wenqing He*, University of Western Ontario
Juan Xiong, University of Western Ontario

The development and progression of cancer are linked to genomic alterations. Comparative Genomic Hybridization (CGH) is a useful technique to explore the whole genome for possible genomic alterations such as amplification or deletion for DNA copy numbers. There are many methods proposed to detect the genomic alterations in the literature, from either Frequentist's or Bayesian points of view. We propose to use a hybrid approach to detect the genomic alterations: a subset of the model parameters is inferred through Bayesian method assuming we have good prior information, and the other parameters are examined by likelihood method. The spatial dependence is also addressed in the proposed method.

email: whe@stats.uwo.ca

80. BIOPHARMACEUTICAL STUDIES

DESIGN AND ANALYSIS OF ALCOHOL/BENZO INTERACTION STUDIES

Daowen Zhang*, North Carolina State University
Hui Quan, Sanofi-Aventis
Zhaoling Meng, Sanofi-Aventis

One aspect of the safety profile of a compound may concern how the combined use of the compound and alcohol or benzo will affect patients' psychomotor and cognitive functions, which are often characterized by many pharmacodynamic (PD) endpoints. In this talk, we discuss the use of cross-over and parallel designs to address the interaction effect between a compound and alcohol. Since there is currently no universally accepted approach for the design and analysis, we present three different procedures for making inference on the interaction effect and the corresponding sample size calculation strategies, all of which are based on a multivariate linear mixed model for the multiple endpoints of interest. The first procedure takes a non-inferiority viewpoint to investigate the existence of a worsening interaction effect. The second procedure takes a superiority viewpoint. We adapt Hochberg's approach in the third procedure. Sample size calculation is then followed for each procedure.

email: dzhang2@stat.ncsu.edu

ASSESSING DRUG INTERACTION WHEN DATA COLLECTED AT FIXED RAYS

Maiying Kong*, University of Louisville
J. Jack Lee, M.D., Anderson Cancer Center, University of Texas at Houston

Studies of interactions among biologically active agents have become increasingly important in many branches of biomedical research. We believe that the Loewe additivity model is one of the best general reference models for defining drug interactions. Based on the Loewe additivity model, synergy occurs when the interaction index is less than one, and antagonism occurs when interaction index is greater than one. Based on the Loewe additivity model and following Chou and Talalay's method for assessing drug interaction based on the plot of interaction indices versus effects for combination doses at a fixed ray, we construct a pointwise $(1-\alpha) \times 100 \%$ confidence bound for the curve of interaction indices versus effects. We found that this method works better on the logarithm transformed scale than on the untransformed scale of the interaction index. Simulations and case studies are given to illustrate the performances of this procedure, and S-PLUS/R code is provided to facilitate the implementation of this
procedure.

email: maiying.kong@louisville.edu

QUANTIFYING COMBINATION DRUG SYNERGY USING NONLINEAR BLENDING

John J. Peterson*, GlaxoSmithKline Pharmaceuticals, R&D

Many classical synergy measures were derived under somewhat idealized pharmacological situations. These measures are rather limited relative to the wide variety of response surfaces that occur in practice. The statistical area of 'mixture experiments', however, makes use of a concept called nonlinear blending to quantify synergy. Nonlinear blending by its simple nature is well defined for any shaped dose response surface. Drugs with different relative potencies, different effect maxima, or situations of potentiation or coalism pose no problem for nonlinear blending as a way to assess the increased response benefit to be gained by combining two drugs. This talk introduces the concept dichotomy of weak and strong nonlinear blending, and shows how strong nonlinear blending can be used for determining whether or not to blend compounds for enhanced efficacy.

email: john.peterson@gsk.com

ASSESSING ASSOCIATION IN A STRATIFIED EXPERIMENT

Jason Liao*, Merck Research Laboratories
Daniel Holder, Merck Research Laboratories

Pearson's correlation coefficient has been widely used to measure the association of two variables in an un-stratified experiment. However, in practice, experiments are often stratified based on factors. In this paper, a common correlation coefficient is derived to estimate the common association when the association at each level of a stratified experiment is the same. Our estimator of the common correlation coefficient has a better small sample performance than the commonly used sample size weighted estimator in terms of bias and mean squared error. The proposed statistics are demonstrated using data from a vaccine potency assay in mice.

email: jason_liao@merck.com

AN ALGORITHM WITH GENETIC INFORMATION FOR WARFARIN DOSING

Kerrie Nelson*, University of South Carolina
David A. Schoenfeld, Harvard University

The anticoagulant drug, warfarin, is commonly prescribed for the prevention of blood clotting in many different medical situations. Since its introduction in the 1950's, it has proved to be a challenge to determine exact doses for any particular patient due to wide intra- and inter-variability in its effects, and since over- or underdosing of warfarin can have serious consequences such as severe bleeding. Many studies have been carried out to derive starting doses and maintenance dose algorithms that can be used for new patients in order to achieve stability of the drug's effects in an individual in the shortest possible time. In this talk we propose a statistical model for describing the association between warfarin dosing and a patient's international normalized ratio (INR) and algorithm for determining a starting dose and maintenance dosing to achieve stability, while taking into account important genetic and clinical factors. Results are based upon a study currently being undertaken at Massachussetts General Hospital.

email: kerrie@stat.sc.edu

SEMIPARAMETRIC BAYESIAN INFERENCE FOR PHAGE DISPLAY EXPERIMENTS

Luis G. Leon-Novelo*, Rice University and University of Texas, M.D. Anderson Cancer Center
Peter Mueller, University of Texas, M.D. Anderson Cancer Center
Kim-Anh Do, University of Texas, M.D. Anderson Cancer Center
Renata Pasqualini, University of Texas, M.D. Anderson Cancer Center
Wadih Arap, University of Texas, M.D. Anderson Cancer Center
Mikhail Kolonin, University of Texas, M.D. Anderson Cancer Center
Jessica Sun, University of Texas, M.D. Anderson Cancer Center

We discuss inference for a mouse phage display experiment with three stages. The data are tri-peptide counts by organ and stage. The primary aim of the experiment is to identify ligands that bind with high affinity to a given organ. We formalize the research question as inference about the monotonicity of mean counts over stages. The inference goal is then to identify a list of peptide and organ pairs with significant increase over stages. We develop a semi-parametric model as a mixture of Poisson distributions with a Dirichlet process prior on the mixing measure. The posterior distribution under this model allows the desired inference about the monotonicity of mean counts. However, the desired inference summary as a list peptide and organ pairs with significant increase involves a massive multiplicity problem. We consider two alternative approaches to address this multiplicity issue. First we propose an approach based on the control of the posterior expected false discovery rate. We notice that the implied solution ignores the relative size of the increase. This motivates a second approach based on a utility function that includes explicit weights for the size of the increase.

email: lgl1@rice.edu

COMPARING RANDOMLY RIGHT CENSORED SAMPLES BASED ON OVERLAP OF DISTRIBUTIONS

Michael J. Dallas, Merck & Co., Inc.

By measuring the overlap of distributions of two treatment effects from a clinical trial, comparisons of the treatment effects can be made on an individual trial participant basis rather than on a mean response basis. This is useful in many clinical settings. For uncensored data, overlap methodology has been studied previously, but it has apparently not been studied in the setting of randomly right censored data. Using an extension of a suggested measure proposed for the uncensored case, I present an overlap method to use for randomly right censored data. Properties of the method are illustrated.

email: michael_dallas@merck.com

81. POWER ANALYSIS

DETECTING QUALITATIVE INTERACTION AMONG SUBGROUPS OF A CLINICAL TRIAL: A BAYESIAN APPROACH

Emine O. Bayman*, The University of Iowa and Uludag University-Turkey
Kathryn Chaloner, The University of Iowa

Sometimes it is important in design to consider not only the overall power, but also the subgroup power. In general the analysis by centeris a subgroup analysis. Differences between treatment effects between centers in a multi-center trial represent interaction. Peto (1982) defines quantitative and qualitative interaction. Quantitative interaction occurs when the simple subgroup treatment effects are different and the signs of the treatment effects are identical across all subgroups: quantitative interaction is common and is not usually important. Qualitative interaction occurs when the simple treatment effect in at least one subgroup is in a different direction than in other subgroups: this interaction is important and is important to detect. A hierarchical model for binary responses is assumed, with exchangeable subgroup treatment effects. A Bayesian test of qualitative interaction is developed by calculating the posterior probability of qualitative interaction. The frequentist power of the Bayesian test is examined and compared to other approaches such as the method of Gail and Simon (1985). The impact on power of imbalance between the sample sizes in each subgroup is examined. The method is implemented using WinBUGS and R, and the R2WinBUGS interface.

email: emine-unlu@uiowa.edu

SELECTION OF THE CLINICALLY RELEVANT DIFFERENCE FOR SAMPLE SIZE DETERMINATION IN A CLINICAL TRIAL BY MATCHING BAYESIAN AND FREQUENTIST PROCEDURES

Maria M. Ciarleglio*, V.A. Cooperative Studies Program

This paper proposes a method for specifying an estimate of the clinically relevant difference (deltaF) to be detected in a clinical trial for use in frequentist sample size determination. The sample size is chosen so that pre-specified significance levels and power levels are achieved. The matching Bayesian test assumes that the difference has a prior distribution, and the sample size is chosen so that Bayesian significance level and the Bayesian power match the pre-specified frequentist significance and power levels. Each Bayesian prior on the unknown difference will generate a corresponding frequentist estimate of the clinically relevant difference. Iterative solutions are presented for a range of the prior hyperparameters for the Normal two-sample model with known variance assuming a Normal prior is placed on the difference in means. We found that to achieve an equivalent level of Bayesian power, the clinically relevant difference for a given study should be down-weighted from the prior mean to account for the increased variability in the pre-trial estimates. The functional relationship deltaF = (prior mean) -0.5*(prior standard deviation) may be used as a practical guide for determining the value of the clinically relevant difference that approximately matches Bayesian power to frequentist power for a reasonable range of actual effect sizes.

email: maria.ciarleglio@yale.edu

TWO-STAGE CLINICAL TRIAL DESIGNS FOR MULTIPLE ENDPOINTS WITH CORRELATED OBSERVATIONS

Fei Ye*, Vanderbilt University
Yu Shyr, Vanderbilt University

Many clinical trials have multiple endpoints. In addition, some trials may have participants with multiple observations on the same endpoint of interest. Two different decision rules can be applied in the case of multiple endpoints: terminating a trial when either endpoint exceeds a boundary, or terminating a trial when all endpoints exceed a boundary. In the latter situation, the correlation between the outcomes must be taken into consideration when determining the endpoints to satisfy the type I error constraint. We propose sample size calculation methods and the corresponding stopping criteria for two-stage phase II trials when both of the correlation between endpoints and the correlation between observations on the same endpoint need to be considered. A simulation study is carried out to conduct designs for different parameter settings and to evaluate the relationship between degree of correlations versus sample sizes and statistical power.

email: fei.ye@vanderbilt.edu

COMPARISON OF TWO-PHASE ANALYSES FOR CASE-CONTROL GENETIC

Gang Zheng, National Heart, Lung and Blood Institute
Mark J. Meyer*, American University
Wentian Li, The Robert S. Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, Manhasset
Yaning Yang*, University of Science and Technology of China
Hefei Anhui, P.R. China

Genetic association tests often employ Cochran-Armitage trend tests (CATTs) and Pearson's chi-square test. Both tests are genotype-based. Song and Elston (2006, Statist Med) introduced the Hardy-Weinberg disequilibrium trend test (HWDTT) and combined it with CATT. Recently, several two phase analysis procedures based on CATTs and HWDTT have been proposed. Compared to using a single statistic to test for case-control genetic association, two-phase analyses employ two test statistics in one analysis framework, each statistic using all available case-control data. Either these two test statistics are asymptotically statistically independent under the null hypothesis of no association or the first phase can be used to determine the underlying genetic model which is then used to test for association in the second phase. The two proposed two-phase analyses have been compared to single-phase analyses. However, no comparison between them has been conducted. Hardy-Weinberg equilibrium (HWE) in the population is a key assumption for the existing two phase analyses. We propose a new two-phase analysis procedure, referred to as two-phase beta-test (TBT), which does not require HWE in the population. We compare three two-phase analyses strategies by extensive simulation studies, application to case-control studies, and providing guidelines for practical use of two-phase techniques.

email: mark.john.meyer@gmail.com

BORROWING STRENGTH IN A SIMPLE HIERARCHICAL MODEL

Xuefeng Li*, U.S. Food and Drug Administration

The work studied how much strength is borrowed from historical data for a current trial in a simple Bayesian hierarchical model. Specifically, the study is to find out how type I error rate, power and sample size change when the historical control data is deviating from the current control data in a randomized superiority trial. Simulation studies were performed with normal distributions. It was found that when the historical data is close to the current data, strength is borrowed in terms of power gaining. When the historical data is far from the current data, negative strength is borrowed in terms of power losing. And the borrowing is not symmetric around the target value.

email: xuefeng.li@fda.hhs.gov

AN EFFICIENT GROUP SIZE RATIO IN TRIALS WITH MULTIPLE DOSE GROUPS VERSUS A COMMON CONTROL

Jianliang Zhang*, MedImmune, Inc.

It is common in POC trials to include multiple doses with a common control to detect drug activities in early phase development without multiplicity adjustment or assumptions for dose response relationships. The dose groups are usually equally sized and the control group is either the same size or arbitrarily inflated, e.g. 1.5 times the size of a dose group. In such a design, the typical 1:1 ratio of dose group versus control group sizes may not be an efficient ratio with respect to usage of statistical information. The efficiency of a design with unequal group size relative to a design with equal group size is a function of the number of dose groups (K) and the ratio (r) of the dose versus control group sizes for a fixed study size. Thus, a value of r can be determined to maximize the relative efficiency and to add statistical power for individual comparisons in such POC trials. Such a design may be particularly useful for trials comparing a single test agent to multiple controls with a desire for over-proportional exposure to the agent to provide adequate safety data of the agent. The design with unequal group size may randomize more patients to the test agent group while maintaining a high efficiency.

email: zhangj@medimmune.com

STATISTICAL METHODS FOR ACTIVE EXTENSION TRIALS

Zonghui Hu*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Dean Follmann, National Institute of Allergy and Infectious Diseases, National Institutes of Health

This paper develops methods of analysis for active extension clinical trials. Under this design, patients are randomized to treatment or placebo for a period of time (period 1), and then all patients receive treatment for an additional period of time (period 2). We assume a continuous outcome is measured at baseline and at the end of the two consecutive periods. If only period 1 data is available, classic estimators of the treatment effect include the change score, ANCOVA, and maximum likelihood (ML). We show how to extend these estimators by incorporating period 2 data which we refer to as the period 2 estimators. Under the assumption that the mean responses for treatment and placebo arms are the same at the end of period 2, the new estimators are unbiased and more efficient than estimators that ignore period 2 data. If this assumption is not met, the period 2 estimators can be biased downward (upward) if the treatment effect during period 2 is larger (smaller) in treatment arm than the placebo arm. In general, the proposed period 2 procedure can provide an efficient way to supplement but not supplant the usual period 1 analysis.

email: huzo@niaid.nih.gov

82. SURVIVAL ANALYSIS – CURE RATE MODELS AND RECURRENT EVENTS

CONDITIONAL MEAN GAP TIME ESTIMATION WITH RECURRENT EVENTS

Adin-Cristian Andrei*, University of Wisconsin-Madison

In numerous clinical trials, information is available on a series of successive landmark events. Such sequences may include: randomization time, date of first hospitalization and death date. The time elapsed between two successive events is called a gap time. In the presence of censoring, one may not observe the gap times of interest in their entirety. Based on inverse probability-of-censoring weighting techniques, we construct consistent and asymptotically normal estimators of the conditional mean of the most recent gap time, given all prior gap times. Simulations are performed in a variety of scenarios and an example from the International Breast Cancer Study Group Trial V is used to illustrate these methodological developments.

email: andrei@biostat.wisc.edu

SEMIPARAMETRIC CURE RATE MODELS WITH RANDOM EFFECTS

Guoqing Diao*, George Mason University
Guosheng Yin, University of Texas-M. D. Anderson Cancer Center

We propose a novel class of cure rate models for multivariate failure time data with a survival fraction. The class is formulated through a transformation on the unknown population survival function. It incorporates random effects to account for the underlying correlation, and includes the mixture cure model structure and the proportional hazards cure model structure as two special cases. We propose a general form of the covariate structure which automatically satisfies an inherent parameter constraint. Moreover, it accommodates the corresponding binomial and exponential covariate structures in the two main formulations of cure models. The proposed class provides a natural link between the mixture and proportional hazards cure models, and it offers a wide variety of new modeling structures as well. We show that the nonparametric maximum likelihood estimators (NPMLE) for the parameters of these models are consistent and asymptotically normal. The limiting variances achieve the semiparametric efficiency bounds and can be consistently estimated. Simulation studies demonstrate that the proposed methods perform well in practical situations. This class of models is illustrated with a real example.

email: gdiao@gmu.edu

OPTIMAL GOODNESS-OF-FIT TESTS FOR RECURRENT EVENT DATA

Russell S. Stocker*, Mississippi State University
Edsel A. Pena, University of South Carolina

To analyze recurrent event data researchers often use models that incorporate an effective age process. We consider a class of models that utilize a perfect repair process. For this class of models, we propose a class of goodness-of-fit tests for the hypothesis that the hazard rate function belongs to a parametric family. The asymptotic properties of the test statistic are given. Finite sample properties are examined using computer simulation studies. A real data set is used to illustrate the proposed tests.

email: rstocker@math.msstate.edu

AN ACCELERATED FAILURE TIME CURE MODEL FOR TIME-TO-EVENT DATA WITH MASKED CAUSE OF FAILURE

Jing J. Zhang*, Harvard University
Molin Wang, Harvard University and Dana-Farber Cancer Institute

We consider the analysis of time-to-event data that is subject to a cure rate with masked cause of failure. Assuming an accelerated failure time (AFT) model with unspecified error distribution for the time to event of interest, we propose a rank-based estimating equation for the model parameters and use a generalization of the EM algorithm for parameter estimation. The motivation comes from an International Breast Cancer Study Group (IBCSG) clinical trial; some breast cancer adjuvant therapies for premenopausal women have been shown to interrupt menses, or cause amenorrhea. We characterize the process of treatment-induced amenorrhea (TIA), which is complicated by the fact that natural menopause may also occur and is indistinguishable from TIA unless a recovery of menses is observed after treatment end. Moreover, there is a cured proportion associated with TIA, i.e. not all patients will experience TIA. A small simulation study is conducted to evaluate the performance of the proposed method, and an application to data from the IBCSG clinical trial is undertaken.

email: jjzhang@fas.harvard.edu

A BAYESIAN ANALYSIS OF RECURRENT EVENTS DATA WITH DEPENDENT TERMINATION: AN APPLICATION TO HEART TRANSPLANT PROBLEM

Bichun Ouyang*, Medical University of South Carolina
Debajyoti Sinha, Florida State University
Joseph G. Ibrahim, University of North Carolina

In this article,we demonstrate the usefulness of Bayesian method for analyzing recurrent events data with risk of termination dependent on the history of the recurrent events through an analysis of a data set from a heart transplant study. We investigate the practical consequences of key modeling assumptions of several fully specified stochastic models for such data. We focus on a class of models which allows both negative and positive association between the risk of termination and the rate of recurrent events via a frailty variable. We also discuss the relationship as well as the major differences between these models in terms of their motivations and physical interpretations. We discuss associated Bayesian methods based on Markov chain Monte Carlo tools, and novel diagnostic tools to perform inference based on fully specified models.

email: ouyang@musc.edu

FLEXIBLE MODELING OF ADDITIVE TREATMENT EFFECTS ON THE RECURRENT EVENT MEAN IN THE PRESENCE OF A TERMINATING EVENT

Qing Pan*, George Washington University
Douglas E. Schaubel, University of Michigan

We consider recurrent events (e.g., hospitalization) ended by a terminating event (e.g., death). Often, the difference between treatment-specific recurrent event means are not constant over time, particularly when treatment-specific differences in survival exist. Thus treatment effect based on the cumulative differences in the recurrent event mean as opposed to the instantaneous difference in the rate is of interest. We propose two methods that compare treatments by separately estimating the survival probabilities and recurrent event rate given survival, then integrating to get the mean number of events. Both methods combine an additive model for the conditional recurrent event rate and a proportional hazards model for the terminating event hazard. The first proposed method features treatment-specific survival distributions and generates an estimated difference in treatment-specific means. The second method factors out treatment-specific differences in survival by employing a common survival function estimator (intended to serve as a standard) for both treatment groups. The motivating example is the repeated occurrence of hospitalization among kidney transplant recipients, where the effect of Expanded Criteria Donor (ECD) compared to non-ECD kidney transplantation on the mean number of hospitalizations is of interest.

email: qpan@gwu.edu

NONPARAMETRIC METHOD OF MIXTURE MODEL WITH PREVALENT SAMPLING

Yu-Jen Cheng*, Johns Hopkins University
Mei-Cheng Wang, Johns Hopkins University
Noya Galai, Johns Hopkins University

The objective of this article is to make inference on the survival function and the cure probability subject to left truncation. The problem is especially complex because death and cure are contrast events and both events could be truncated before data recruitment. Mixture model is considered in this article. We addressed the connection between mixture model and competing risk model under prevalent sampling scheme and developed a nonparametric approach to estimate the survival function based on a weaker assumption, conditional independence. Nonparametric MLE of the survival function and the probability of cure are derived in this article. We also show that the model under conditional independence assumption is nonidentifiable subject to right censoring. Our methodology was motivated by and applied to the intensive care unit (ICU) study in Israel.

email: ycheng3@jhsph.edu

83. EXPERIMENTAL DESIGN

OPTIMAL AND EFFICIENT DESIGNS FOR GOMPERTZ REGRESSION MODELS

Gang Li*, GlaxoSmithKline
Dibyen Majumdar, University of Illinois at Chicago

Gompertz functions have been widely used in characterizing the biological growth curve in both research and applications, especially in cancer research. In this paper we consider the designs for two Gompertz regression models. It is theoretically shown that the locally D-optimal designs for these models are minimally supported. Using the D-optimal designs as the reference, we propose several alternative efficient designs and efficiencies for these design were investigated.

email: gangli_stat@yahoo.com

IS THE CONTINUAL REASSESSMENT METHOD A BETTER PHASE I DESIGN? A COMPREHENSIVE COMPARISON WITH THE STANDARD 3 + 3 DOSE ESCALATION SCHEME

Alexia Iasonos*, Memorial Sloan Kettering Cancer Center
Elyn Riedel, Memorial Sloan Kettering Cancer Center

An extensive literature has covered the statistical properties of the Continual Reassessment Method (CRM) and the modifications of this method. While there are some applications of CRM designs in Phase I trials, the standard method of escalating doses after 3 patients with an option for an additional 3 patients (SM) remains very popular, mainly due to its simplicity. This paper compares CRM-based methods with the SM in terms of the number of patients needed to reach the MTD and the total sample size required. The comparisons are performed under two alternative schemes: a fixed or a varying sample approach with the implementation of a stopping rule. Several CRM-based methods were evaluated under different scenarios by varying the number of dose levels and the location of the true MTD. CRM and SM are comparable in terms of how fast they reach the MTD and the total sample size required when testing a limited number of dose levels (d 5), but as the number of dose levels increases, CRM reaches the MTD faster when used with a fixed sample of 20 patients. However, a sample size of 20-25 patients is not sufficient to achieve a narrow precision around the estimated toxicity rate at the MTD.

email: iasonosa@mskcc.org

THREE-DOSE-COHORT DESIGNS IN CANCER PHASE I TRIALS

Bo Huang*, University of Wisconsin-Madison
Rick Chappell, University of Wisconsin-Madison

Traditional designs for phase I clinical trials assign the same dose to patients in the same cohort. In this paper, we present a new class of designs for cancer Phase I trials which initially rapidly escalate by allowing multiple doses (usually 3) to be assigned to each cohort of patients. The class of designs, called the LMH-CRM (an extension of the continual reassessment method (CRM) by administering different percentiles of the MTD, denoted 'low', 'medium', 'high'), is proven to be consistent and coherent (a common-sense property of phase I trials for dose escalation and de-escalation). Three designs (slow, moderate and fast) are derived based on different dose-escalation restrictions. Simulation results show that moderate and fast LMH-CRM combine the advantages of the CRM with one patient per cohort and three patients per cohort: it accurately estimates the MTD; controls overall toxicity rates; and is time efficient.

email: huang@stat.wisc.edu

INTERNAL PILOT DESIGNS FOR OBSERVATIONAL STUDIES

Matthew J. Gurka*, University of Virginia
Christopher S. Coffey, University of Alabama at Birmingham

A proper determination of the sample size required to detect clinical effects of interest is an integral part of study design. A power analysis requires specifying a desired treatment effect as well as estimates of any 'nuisance' parameters (the variance or overall success rate for continuous or binary outcomes, respectively). Internal pilot (IP) designs allow for revising these parameter estimates at an interim stage to adjust the final sample size to achieve the appropriate power. Such designs are widely used for clinical trials but have not been studied extensively in observational settings. In an observational study, the random allocation of the observations among the groups introduces an additional parameter to estimate when computing the required sample size. We extend the use of IP studies to observational designs, allowing for sample size adjustment based on re-estimates of both the usual nuisance parameters and the allocation among the groups from the IP sample. We demonstrate the benefits and consequences of this extension via simulations. Finally, the benefits of the approach are demonstrated through the planning of a genetic epidemiology study focusing on the association between outcome after brain injury and having a particular allele (whose prevalence in this population was unknown).

email: mgurka@virginia.edu

BAYESIAN ADAPTIVELY RANDOMIZED CLINICAL TRIALS IN THE PRESENCE OF PATIENT HETEROGENEITY

J. Kyle Wathen*, University of Texas-M.D. Anderson Cancer Center
Mark F. Munsell, University of Texas-M.D. Anderson Cancer Center
Marcos J. De Lima, University of Texas-M.D. Anderson Cancer Center

Randomized clinical trials are the gold standard for obtaining scientifically valid treatment comparisons. However, if a physician favors one treatment over another, it may be more ethical for the physician to used the favored treatment rather than randomizing. As a compromise, many adaptively randomized procedures exist to unbalance the randomization in favor of the treatment that, on average, has superior results. Most of these methods assume that patients are homogeneous, and thus are likely to reach incorrect conclusions in the presence of patient heterogeneity. We propose a Bayesian adaptively randomized design that may reach different conclusions in two or more prognostic subgroups and we allow for both binary and time-to-event outcomes. The design utilizes subgroup specific randomization probabilities so that it is superior to designs that ignore treatment-subgroup interactions. The design is motivated by results from a recent study of experimental treatment versus Placebo where it was found that the experimental treatment appeared to benefit only one subgroup. A simulation study is presented and the method is illustrated by a phase III study for prevention of acute graft-versus-host disease.

email: jkwathen@mdanderson.org

RANK-BASED ANALYSIS OF CROSSOVER TRIALS FOR CLASSIC AND NOVEL DESIGNS

Mary Putt*, University of Pennsylvania

Most crossover trials in clinical settings use the classic two-treatment, two-period AB:BA design. Compared to the AB:BA design, studies with additional treatment sequences and/or periods generally have greater power to detect treatment effects of interest. In crossover designs, a rank-based analysis is complicated by the presence of nuisance parameters such as carryover and period effects. The Wilcoxon rank sum test (WSR) can be manipulated to test relevant hypotheses. However the WSR does not explicitly use all of the information in the crossover design's repeated measures. Here we use a permutation-based aligned rank test that can be used to test for treatment effects in two-treatment designs where subjects receive the same treatment across multiple periods. We illustrate the approach using a placebo-controlled trial examining the effects of donepezil on cognitive function in Parkinson's disease. A simulation study indicates that both the Wilcoxon rank-sum test and the new approach have higher power than analysis of variance when the data have a skewed or symmetric multivariate T distribution. The proposed approach is valid under the assumption that carryover effects occur only in the presence of treatment effects, and consistently has higher power than the WSR when the dispersion matrix deviates from compound symmetry.

email: mputt@mail.med.upenn.edu

IMPLEMENTING OPTIMAL ALLOCATION FOR SEQUENTIAL CONTINUOUS RESPONSES WITH MULTIPLE TREATMENTS

Hongjian Zhu*, University of Virginia
Feifang, Hu, University of Virginia

In practice, it is important to find optimal allocation strategies for continuous response with multiple treatments under some optimization criteria. In this talk, we focus on the exponential responses. For a multivariate test of homogeneity, we obtain the optimal allocation strategies to maximize power while (1) fixing sample size with K treatments and (2) fixing expected total responses with K=3 treatments. Then the doubly adaptive biased coin design (Hu and Zhang, 2004) is used to implement the optimal allocation strategies. Simulation results show that the proposed method increases the power over complete randomization for equal allocation.

email: zhuhongjian83@hotmail.com

84. ROC/DIAGNOSTIC METHODS

GUESSING FREE-RESPONSE PROCESS AND EMPIRICAL FROC CURVE

Andriy Bandos*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh
David Gur, University of Pittsburgh
Tao Song, University of Pittsburgh

A free-response paradigm is an approach to assess the ability of a diagnostic system to search, locate and classify abnormalities by allowing the system to mark and rate multiple locations within a subject. Unlike the conventional ROC, free-response ROC (FROC) analysis considers both the accuracy and the number of the rated locations as important characteristics of the diagnostic system. The central tool in FROC analysis is the FROC curve. Unlike in the ROC analysis, there is no reference FROC curve (such as diagonal or ‘guessing’ ROC) and empirical FROC curve does not extend to the same trivial point (such as (1,1) in ROC) for all studies. These features substantially complicate the overall assessment of the FROC curves. Similar to the ROC analysis we propose an artificial ‘guessing’ FROC process and use it to generate a ‘guessing’ FROC curve as a reference, and to augment the empirical FROC curve to the trivial point where there are no ‘negative’ findings. The area between the guessing and augmented FROC curves is a direct analog of the area between the diagonal line and empirical ROC curve. We have shown that this index is equivalent to a Wilcoxon-type of performance index with good structural and statistical properties.

email: anb61@pitt.edu

ROC BASED UTILITY FUNCTION MAXIMIZATION FOR FEATURE SELECTION AND CLASSIFICATION

Zhenqiu Liu*, University of Maryland Medicine

In medical diagnosis, the diseased and non-diseased classes are usually unbalanced and one class may be more important than the others depending on the diagnosis purpose. Most standard classification methods, however, are designed to maximize the overall accuracy and cannot incorporate different costs to different classes explicitly. In this paper, we propose a novel algorithm to directly maximize the weighted specificity and sensitivity of the Receiver Operating Characteristic (ROC) curve. Combining advances in machine learning and statistics, the proposed algorithm has excellent generalization property and assigns different error costs to different classes explicitly. We present experiments that compare the proposed algorithms with support vector machines (SVM) and regularized logistic regression using data from HIV-1 protease study as well as six other public available data sets. Our main conclusion is that the performance of proposed algorithm is significantly better in most cases than the other classifiers tested.

email: zliu@umm.edu

ASSESSING AGREEMENT AMONG ACUPUNCTURISTS IN TEAMSI TRIAL

Anna TR Legedza*, Vertex Pharmaceuticals, Inc.
Roger B. Davis, Beth Israel Deaconess Medical Center and Harvard University

We present statistical methodology and results from a recently completed NIH-funded clinical trial (TEAMSI) whose objective was to further develop a reliable and valid East Asian medicine (EAM) diagnostic instrument. Such an instrument could provide a better understanding of the clinical significance and mechanisms underlying the specificity of acupuncture techniques and development of a research methodology that integrates EAM principles and biomedical science in clinical trials of acupuncture and EAM. We evaluated the inter-rater reliability and specificity of acupuncture experts among themselves and versus a gold standard. Major aims included assessment of reliability of the tool (test-retest), inter-rater reliability, and validity. We examined the advantages and disadvantages of two frequently used agreement statistics (Cohen’s kappa; tetrachoric correlation) to assess tool, training, and experience effects vs. a model-based approach.

email: anna_legedza@vrtx.com

ESTIMATE BENEFITS DUE TO FECAL OCCULT BLOOD TEST FOR COLORECTAL CANCER SCREENING

Dongfeng Wu*, University of Louisville
Diane Erwin, Information Management Services, Inc.
Gary L. Rosner, University of Texas-M.D. Anderson Cancer Center
Lyle D. Broemeling, University of Texas-M.D. Anderson Cancer Center

We applied the statistical methods developed by Wu, Rosner and Broemeling 2005, 2007 using the Minnesota colorectal cancer study group data, to make Bayesian inference for the age-dependent screening sensitivity, the age-dependent transition probability from disease-free to preclinical state, the sojourn time distribution, and the lead time distribution, for both male and female participants in a periodic screening program. The sensitivity appears to increase with age for both genders. However, the posterior mean sensitivity is not monotonic with age for males; it has a peak around age 74. The standard errors of the sensitivity are not monotone either; there is a minimum at age 69 for males and at age 78 for females. The age-dependent transition probability is not a monotone function of age; it has a single maximum at age 72 for males and a single maximum at age 75 for females. The posterior mean sojourn time is 4.08 years for males and 2.41 years for females. Based on the lead time distribution, to guarantee a 90% chance of early detection, it seems necessary for the males to take the fecal occult blood test every 9 months, while the females only need to take it annually.

email: dongfeng.wu@louisville.edu

IMPROVING THE EFFICIENCY OF TESTING PREDICTIVE VALUES USING AUXILIARY COVARIATE

Yoonjin Cho*, North Carolina State University
Kosinski Andrezj, Duke University

Diagnostic tests in health care are important. It is necessary to evaluate the performance of a diagnostic test to its gold standard. There are several measures to evaluate the performance of a diagnostictest. We focus on measuring positive and negative predictive values in this article. When more than one diagnostic test are available, the interest is to determine the diagnostic test which performs the best. Comparing the predictive values of the diagnostic test is a possible method and there have been several papers on comparing diagnostic tests when every individual has the outcome of the gold standard. However, it is often the case that not every patient undergoes the gold standard. If we only use patients with gold standard outcomes in evaluating the accuracy of the diagnostic test, the estimate of the predictive values of the diagnostic test is likely to be biased. In this article, we focus on comparing two diagnostic tests with missing gold standard on some individual. In a typical clinical study, both diagnostic tests are given to all patients, while the invasive or expensive gold standard may not be available to some patients. We also assume that subjects in this clinical study are randomly selected from the population since the predictive values depend on the disease prevalence.

email: einsheart@gmail.com

A TRANSFORMATION-INVARIANT MONOTONE SMOOTHING OF RECEIVER OPERATING CHARACTERISTIC CURVES

Liansheng Tang, George Mason University
Pang Du*, Virginia Tech University

When a new diagnostic test is developed, it is of interest to evaluate its accuracy in distinguishing diseased subjects from non-diseased subjects. The accuracy of the test is often evaluated by using Receiver operating characteristic (ROC) curves. In this presentation, we propose a monotone spline approach for obtaining a smooth estimate of an ROC curve. Unlike the current ROC smoothing methods, our method ensures important inherent properties of underlying ROC curves which include monotonicity and transformation invariance. We compare the finite sample performance of the proposed ROC method with other ROC smoothing methods in large-scale simulation studies. We illustrate our method through a real life example.

email: pangdu@vt.edu

A GENERALIZED NONPARAMETRIC APPROACH OF COMPARING FROC SYSTEMS

Tao Song*, University of Pittsburgh
Andriy Bandos, University of Pittsburgh
Howard Rockette, University of Pittsburgh
David Gur, University of Pittsburgh

ROC curve analysis is a widely used method of comparing diagnostic imaging systems. One formulation of the area under the ROC curve is based on the probability of selecting the abnormal subject from a random pair of normal - abnormal subjects. In a Free Response ROC (FROC) process, which requires searching and marking the locations of all suspected abnormalities with a level of suspicion (rating), normal subjects may have multiple false positives and abnormal subjects may have multiple true positives and false positives. We consider a general approach that uses as a summary index the area under an ROC curve derived from an FROC process. The method entails specifying a function that is used to select the abnormal subject from the normal-abnormal pair. A previously proposed index based on the highest rating on a subject can be viewed as a special case of this method. We consider various discriminating functions including average score and stochastic dominance. Simulation studies are conducted to compare the statistical power of these methods to distinguish between two FROC processes.

email: tas67@pitt.edu

85. IMS MEDALLION LECTURE

STATISTICAL CHALLENGES IN GENETIC ASSOCIATION STUDIES

Mary Sara McPeek, University of Chicago
Medallion Speaker

Common diseases such as asthma, diabetes, and hypertension, which currently account for a large portion of the health care burden, are complex in the sense that they are influenced by many factors, both environmental and genetic. One fundamental problem of interest is to understand what the genetic risk factors are that predispose some people to get a particular complex disease. Technological advances have made it feasible to perform case-control association studies on a genome-wide basis. The observations in these studies can have several sources of dependence, including correlation in the genotypes of nearby markers on a chromosome as well as relatedness of the individuals in the study. How to model the effects of this dependence and how to appropriately take it into account in the analysis of genome-wide association studies present interesting statistical challenges, which will be discussed in this talk along with proposed solutions.

email: mcpeek@galton.uchicago.edu

86. ON FUTILITY CALCULATIONS IN RANDOMIZED CLINICAL TRIALS

PRACTICAL ASPECTS OF SOME COMMON FUTILITY RULES

Boris Freidlin*, National Cancer Institute

The ultimate goal of a randomized phase III clinical trial is to provide data on the benefit-to-risk profile of the new therapy that is sufficiently compelling to change medical practice. Accordingly, the purpose of futility monitoring is to allow termination of the trial if early data provide convincing evidence that the benefit-to-risk profile is not improved. What makes the data sufficiently convincing for early futility stopping varies across clinical settings. Therefore, futility boundary needs to be carefully calibrated to match the study circumstances. This talk will discuss several practical aspects of commonly used futility rules. In particular, it will describe a number of popular futility approaches with aggressive stopping boundaries that suggest stopping a trial for futility when the experimental arm looks better than the control arm.

email: freidlinb@ctep.nci.nih.gov

FORMULATING AND SELECTING FUTILITY RULES IN A LONG-TERM TRIAL

Christy Chuang-Stein*, Pfizer Inc.
Michael Brown, Pfizer Inc.
Wayne Ewy, Pfizer Inc.
Cathie Spino, University of Michigan

Consider a long-term trial where the endpoint is measured at the end of two years with an interim measurement at one year. The sponsor and investigators have a strong desire to conduct at least one futility analysis because there is little evidence that the development will be successful. A futility analysis, however, is hampered by a fast enrollment and by the risk associated with predicting the two-year outcome from the one-year interim outcome without a good scientific basis for the prediction model. The presentation will describe a team's deliberations on how to formulate and select among various options. One fundamental question is whether a futility analysis is even justified in this case.

email: christy.j.chuang-stein@pfizer.com

CONDITIONAL POWER CONSIDERATIONS IN THE DESIGN OF A PHASE MICROBICIDE TRIAL IN AFRICA

Jennifer Schumi*, Statistics Collaborative, Inc.
Zeda Rosenberg, International Partnership for Microbicides
Stephanie Dickinson, Indiana Statistical Consulting Center
Janet Wittes, Statistics Collaborative, Inc.

Microbicides have the potential to slow the spread of HIV by reducing transmission. Designing a placebo-controlled Phase 3 trial to show efficacy of these novel products presents many challenges, including accurate estimation of HIV incidence and effectiveness of the microbicide. Researchers and funders recognize the need for study designs that balance safety and efficacy while making efficient use of scarce financial and human resources. We propose a design incorporating futility analysis that calculates incidence rates from data observed at interim looks and projects those rates forward on the basis of conditional power. Such a trial design allows researchers to select the minimum efficacy that is likely to be clinically relevant and to discontinue the trial early for products that have unacceptably low efficacy or that may even be harmful.

email: jennifer@statcollab.com

87. STATISTICAL METHODS FOR DETECTING COPY NUMBER VARIATION

DETECTION OF COPY NUMBER VARIATIONS FROM HIGH-DENSITY SNP ARRAYS: AN INTEGRATED BAYESIAN HIDDEN MARKOV MODEL APPROACH INCORPORATING PEDIGREE INFORMATION

Zhen Chen*, University of Pennsylvania School of Medicine
Mahlet Tadesse, Georgetown University
Kai Wang, University of Pennsylvania School of Medicine
Mingyao Li, University of Pennsylvania School of Medicine

Copy number variations (CNVs) refer to gains and losses of genomic elements compared to a reference genome assembly. CNVs are common in humans and some CNVs are associated with human phenotypic variation and susceptibility to disease. Recent advances in genotyping technologies have made it possible to make high-resolution CNV calls using whole-genome SNP arrays. Various studies have demonstrated the heritability of CNVs, however, few have incorporated family structures in the analysis. Here we develop an integrated Bayesian approach that aims to incorporate family relationships when inferring CNVs in the context of parents-offspring trios. We assume that the copy number sequence along the chromosome follows a Markov model with transition probabilities dependent on genetic distances between adjacent SNPs. Specifically, we model parental CNVs through a standard hidden Markov model; given parental copy numbers at a SNP, the offspring's copy number is then modeled through Mendelian inheritance and the dependence on the copy number at the previous SNP is modeled through recombination fraction between the two SNPs. Due to cell-line artifacts or segmental mutation events during recombination, many CNVs might be non-Mendelian inherited. To accommodate such CNVs, we allow for de novo events in offspring's CNV calls and use another HMM to account for the dependence with neighboring SNPs in the same de novo CNV region. Our Bayesian approach yields posterior distributions of copy number configurations for the three family members, thus providing an uncertainty measure for the inferred CNV calls. By incorporating both family data and allowing for de novo events, our method provides flexibility for the analysis of a wide range of settings, and is expected to improve the accuracy of CNV calls as compared to methods that ignore family relationships. We evaluate the performance of the proposed method and illustrate its practical utility by applying it to the analyses of simulated datasets and the CEU trio data from HapMap.

email: chenz@mail.med.upenn.edu

ESTIMATING GENOME-WIDE COPY NUMBER USING ALLELE SPECIFIC MIXTURE MODELS

Shin Lin*, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine
Benilton Caravalho, Johns Hopkins Bloomberg School of Public Health
Wenyi Wang, Johns Hopkins Bloomberg School of Public Health
Aravinda Chakravarti, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine
Rafael Irizarry, Johns Hopkins Bloomberg School of Public Health

Copy number variants (CNV) are thought to be one of the major underlying causes of human phenotypic differences among normal and disease subjects. Instead of the expected two copies, a CNV at a particular section of the genome may represent zero copies (homozygous deletion), one copy (hemizygous deletions), or more than two copies (amplifications). More than a decade ago, comparative genomic hybridization (CGH) technology was developed to detect CNV in a high-throughput fashion. However, this technology only provides a 10 MB resolution which limits the ability to detect CNV spanning small regions. It is widely believed that CNV as small as one base can have significant downstream effects; thus, microarray manufacturers have developed technologies which provide much higher resolution. Unfortunately, strong probe effects and variation introduced by sample preparation procedures have made single-point (tens of bases) copy number estimates too imprecise to be useful. We propose a mixture model solution specifically designed for single-point estimation, which provides various advantages over the existing methodology. Software to implement this procedure will be available in the Bioconductor oligo package (http://www.bioconductor.org).

email: shin.lin@uphs.upenn.edu

CIRCULAR BINARY SEGMENTATION FOR THE ANALYSIS OF ARRAY CGH DATA

Venkatraman E. Seshan*, Columbia University
Adam Olshen, Memorial Sloan Kettering Cancer Center

DNA sequence copy number is the number of copies of DNA at a region of a genome. The development of malignant tumors and their progression often involve alterations in DNA copy number. We will present the motivation for the Circular Binary Segmentation algorithm we developed (Olshen et al Biostatistics, 2004) to segment the genome

into regions of equal copy number. We will also present refinements to the algorithm to handle the large arrays that are being used more commonly now (Venkatraman & Olshen Bioinformatics, 2007). We will present extensions to the problem such as clone specific copy number and the application to tumor data.

email: ves2111@columbia.edu

88. CHALLENGES IN LARGE, SIMPLE TRIALS

INTRODUCTION TO LARGE, SIMPLE TRIALS: WHAT ARE THEY AND WHEN SHOULD THEY BE DONE (OR NOT DONE)?

Julie E. Buring*, Harvard Medical School-Brigham and Women's Hospital

This presentation will review the first principles of large, simple trials (LSTs), in terms of what they are, the background of why they should be conducted, and when. Specific design issues will be discussed, such as the need for large sample sizes, use of broad entry criteria, use of streamlined protocols, and use of clinically important outcome measures. The limitations of large, simple trials will be addressed, as well as the impact of these trials on clinical practice. These principles will be illustrated by examples from the published literature.

email: jburing@rics.bwh.harvard.edu

PRACTICAL CHALLENGES TO KEEPING THE SIMPLE IN LARGE, SIMPLE STUDIES

James D. Neaton*, University of Minnesota

Many large trials require several hundred participating sites to enroll patients in a timely way. The inclusion of many investigators from heterogenous practice settings in research can enhance generalizability. However, the increasing amounts of defensve documentation required for trials and regulatory requirements are making such trials more difficult to do. Although the protection of the safety and confidentiality of study participants have to remain of utmost concern, for many questions, requiring very large sample sizes, with a small number of focused data collection requirements, imposing heavy administrative burdens on study staff and participating clinicians could have serious negative consequences for enrollment and retention rates. When the interest, as it is likely to be in the future, is in performing large, scientifically valid studies, within a reasonable time frame at a reasonable cost, the paradigm for study conduct needs careful rethinking.

email: jim@ccbr.umn.edu

ON SEQUENTIAL MONITORING AND INFERENCE OF SAFETY ENDPOINTS

Qing Liu*, Johnson & Johnson Pharmaceutical Research and Development

Consider a clinical trial with several interim analyses where the efficacy endpoints are evaluated via group sequential tests. Typically there are no specified guidelines for monitoring safety endpoints; rather, decisions to stop the trial or drop treatment groups for safety reasons are left to clinical judgments by a data monitoring committee. With recent scrutiny on drug safety, there is an increasing trend towards more rigorous safety testing and monitoring of on-going trials. We propose sequential safety monitoring guidelines by which a trial could stop early for safety concerns. We also propose formal sequential non-inferiority tests by which non-inferiority claims may be made upon final analysis. We develop new sequential p-values, point estimates, and asymmetrical confidence intervals for both efficacy and safety endpoints.

email: qliu2@prdus.jnj.com

89. MULTISTATE APPROACH TOWARDS MODELING COMPLEX BIOMEDICAL DATA

REGRESSION ANALYSIS FOR MULTISTATE MODELS BASED ON A PSEUDO-VALUE APPROACH, WITH APPLICATIONS TO BONE MARROW TRANSPLANTATION STUDIES

John P. Klein*, Medical College of Wisconsin

Typically, regression analysis for multistate models has been based on regression models for the transition intensities. These models lead to highly non linear and very complex models for the effects of covariates on state occupancy probabilities. We present a technique that models the state occupancy or transition probabilities in a multistate model directly. The method is based on the pseudo-values from a jackknife statistic constructed from non-parametric estimators for the probability in question. These pseudo-values are used as outcome variables in a generalized estimating equation to obtain estimates of model parameters. We examine this approach and its properties in detail for two special multistate model probabilities, the cumulative incidence function in competing risks and the current leukemia free survival used in bone marrow transplants. The latter is the probability a patient is alive and in either a first or second post transplant remission. The techniques are illustrated on a data set of CML patients given a marrow transplant. We also discuss extensions of the model that are of current research interest.

e-mail: klein@mcw.edu

NONPARAMETRIC ESTIMATION OF STATE WAITING TIME DISTRIBUTIONS IN A MARKOV MULTISTATE MODEL

Somnath Datta*, University of Louisville
Ling Lan, University of Louisville
Rajeshwari Sundaram, National Institute of Health

We describe novel nonparametric estimators of state waiting times distribution functions based on current status data in a Markov multistate model. These estimators are obtained from estimators of state entry and exit time distributions via a convolution formula. The state entry and exit time distributions are valid without the Markov assumption and are obtained using estimated transition counting processes and numbers "at risk" processes obtained using nonparametric regression and fractional weights. Finite sample behavior of our estimators is studied by simulation, in which we show that our estimators based on current status data compare well with those based on complete data. We also illustrate our method using a data set arising from a study on stages of sexual developments of children between the ages of age eight and eighteen.

e-mail: somnath.datta@louisville.edu

JOINT MODELING OF HIV MUTATIONS AND FAILURE TIME DATA

Chengcheng Hu*, Harvard School of Public Health

Antiretroviral therapy has been very successful in suppressing replication of HIV and slowing the progression to AIDS. Development of drug resistant strains of the virus is a major cause for treatment failures. In this talk we jointly model the progression of HIV resistance mutations and failure times. The time of mutation can not be observed directly, and hence is considered interval censored. The viral sequences are classified into a few genetic states, and the longitudinally measured genetic states are then modeled as a Markov process. The proposed method is applied to data from several clinical trials.

e-mail: cchu@hsph.harvard.edu

90. INFORMING HEALTH POLICY DECISIONS: BEST TEST STRATEGIES FOR COLORECTAL CANCER SCREENING

MICROSIMULATION MODELING OF COLORECTAL CANCER TO INFORM SCREENING GUIDELINES

Ann G. Zauber*, Memorial Sloan-Kettering Cancer Center
Iris Lansdorp-Vogelaar, Erasmus University-The Netherlands
Janneke Wilschut, Erasmus University-The Netherlands
Amy Kundsen, Massachusetts General Hospital
Marjolein Van Ballegooijen, Erasmus University-The Netherlands
Martin Brown, National Cancer Institute
Karen Kuntz, University of Minnesota

In 2002 the United States Preventive Task Force (USPSTF) recommended that all asymptomatic persons age 50 or older should have colorectal cancer screening. However, they also stated that there was insufficient evidence to date to recommend one screening strategy over another. In Nov 2007 the USPSTF will meet to evaluate whether there is sufficient evidence at this time to recommend one test strategy over another. For the first time the USPTF has asked for a decision analysis to complement its evidence based review. This session presents the methodology for microsimulation modeling of colorectal cancer to inform guidelines, the value of comparative modeling as developed by the NCI program Cancer Intervention and Surveillance Network (CISNET), the modeling of effectiveness across multiple modalities, and the implications of these analyses for health policy. The natural history of colorectal cancer is modeled according to the adenoma-carcinoma sequence with progression of an adenomatous polyp from small to larger size and then further progression to advanced stage of colorectal cancer and to death due to colorectal cancer. Screening can interrupt the adenoma-carcinoma sequence by detection and treatment of colorectal cancer at an early stage or, even better, by detection and removal of the precursor adenoma polyp prior to its progression to colorectal cancer.

e-mail: zaubera@mskcc.org

THE METHODOLOGY AND ADVANTAGES OF COMPARATIVE MODELING IN MICROSIMULATION ANALYSES

Iris Lansdorp-Vogelaar*, Erasmus University-The Netherlands
Ann G. Zauber, Memorial Sloan-Kettering Cancer Center
Janneke Wilschut, Erasmus University-The Netherlands
Amy Knudsen, Massachusetts General Hospital
Marjolein Van Ballegooijen, Erasmus University-The Netherlands
Karen M. Kuntz, University of Minnesota

Microsimulation modeling is built upon the best evidence of the adenoma carcinoma sequence and the effect of screening interventions. But most observations of the adenoma-carcinoma sequence are cross sectional rather than prospective. Different models make different assumptions concerning the processes which have are not observable. The assumptions for these processes are called “deep parameters”. In the interest of better understanding of the findings from different models, the CISNET modelers have developed exercises of base case analyses to provide diagnostic assessments of the model results and to understand where models may differ and due to what assumptions. In particular the MISCAN model assumes a shorter dwell time from adenoma to colorectal cancer than the SimCRC model. Comparative modeling provides for a more robust assessment of the consistencies of the model results and also suggests where current uncertainties suggest more caution in interpretation of the results. Graphical techniques have been developed to summarize model differences and results. The comparative modeling techniques as developed by CISNET will be presented in November 2007 to the United States Preventive Task Force to inform their deliberations concerning recommendations for different screening scenarios.

e-mail: zaubera@mskcc.org

COMPARATIVE EFFECTIVENESS OF COLORECTAL CANCER SCREENING STRATEGIES

Karen M. Kuntz*, University of Minnesota
Iris Lansdorp-Vogelaar, Erasmus University-The Netherlands
Janneke Wilschut, Erasmus University-The Netherlands
Amy B. Knudsen, Massachusetts General Hospital
Marjolein Van Ballegooijen, Erasmus University-The Netherlands
Ann G. Zauber, Memorial Sloan-Kettering Cancer Center

We used two independently-developed microsimulation models (MISCAN and SimCRC) to conduct a decision analysis of five colorectal cancer screening tests in current use (Hemoccult II, Hemoccult SENSA, fecal immunochemical test, flexible sigmoidoscopy with and without FOBT, and colonoscopy), by age to begin screening, age to end screening, and intervals for repeat screening. Our results were provided to the United States Preventive Services Task Force (USPSTF), alongside the standard evidence review, in their evaluation of colorectal cancer screening recommendations. We used the simulation models to incorporate the available evidence on the risks and benefits of each screening test and project outcomes for various screening strategies, such as life years gained relative to no screening and the expected number of colonoscopies required. The results and graphical techniques as presented to the USPSTF in November 2007 are presented in this session. (We are not at liberty to release these results prior to the March Biometrics meetings.)

e-mail: kmkuntz@umn.edu

POTENTIAL IMPACT OF MICROSIMULATION MODELING OF BEST TEST SCENARIOS FOR COLORECTAL CANCER SCREENING ON HEALTH POLICY RECOMMENDATIONS

Martin L. Brown*, National Cancer Institute

Guidelines and recommendations are generally based on evidence based review of the literature of retrospective studies, prospective studies, and randomized controlled trials. Integration of microsimulation modeling results to represent effectiveness of different ages to begin or end screening or interval to rescreen allows for more informed decisions in setting health policy.

e-mail: mb53o@nih.gov

91. QUANTITATIVE TRAIT LOCI MAPPING

ASYMPTOTIC TEST OF MIXTURE MODEL AND ITS APPLICATIONS TO QTL INTERVAL MAPPING

Dong-Yun Kim*, Virginia Tech University
Yuehua Cui, Michigan State University

The aim of this paper is to study the behavior of the profile log-likelihood ratio test statistic in a mixture model with unknown proportion. Using local asymptotic approach, we derive the limiting process of the test statistic for a class of one parameter exponential family. This result is applied to the problem of testing the presence of a Quantitative Trait Loci (QTL) in an interval. The method is illustrated using a real data set and via computer simulation. The test is applicable to data of moderate sample size, and the threshold value for the test can be directly obtained from a simple formula.

e-mail: dongyunkim@vt.edu

BAYESIAN SEMIPARAMETRIC MULTIPLE QUANTITATIVE TRAIT LOCI (QTL) MAPPING

Fei Zou*, University of North Carolina at Chapel Hill
Fuxia Cheng, Illinois State University
Haibo Zhou, University of North Carolina at Chapel Hill
Ina Hoeschele, Virginia Tech University
Hanwen Huang, University of North Carolina at Chapel Hill

In linkage and association mapping, it is often necessary to include covariates such as age or weight to increase power or avoid spurious false positive findings. However, if a covariate term in the model is specified incorrectly (e.g. as linear), then the inclusion of the covariate may adversely affect power and precision of the QTL identification. We propose to implement a semiparametric model for multiple QTL analysis which includes an unspecified function of any covariate found or suspected to have a more complex than linear but unknown relationship with the response variable. This analysis will be performed in a Bayesian inference framework using Markov chain Monte Carlo. The advantages of the proposed method will be demonstrated via extensive simulations.

e-mail: fzou@bios.unc.edu

GENETIC MAPPING OF QUANTITATIVE TRAIT LOCI IN AUTOTETRAPLOIDS

Jiahan Li*, University of Florida
Rongling Wu, University of Florida

Autotetraploids include many plant species, such as potatoes, which are of paramount importance to agricultural production and biological research. Quantitative trait locus (QTL) mapping in autotetraploids is challenged by their unique cytogenetic properties, such as preferential pairing and double reduction. In this talk, we present a statistical model for mapping autotetraploid QTLs by considering these cytogenetic properties. Our model is built in the mixture model-based Bayesian framework and implemented with the Markov chain Monte Carlo algorithm. The model allows simultaneous estimation of QTL positions, QTL effects, the chromosomal pairing factor and the degree of double reduction as well as the assessment of the estimation precision of these parameters. Computer simulation is used to examine the statistical properties of the model. Our model will provide a useful tool for QTL mapping in autotetraploids that undergo double reduction.

e-mail: jiahanli@ufl.edu

MODELING A NON-SEPARABLE COVARIANCE STRUCTURE IN HIGH-DIMENSIONAL FUNCTIONAL MAPPING OF QUANTITATIVE TRAITS

John Stephen Yap*, University of Florida
Rongling Wu, University of Florida

Wu et al. (2007) studied the genetic architecture of reaction norms to multiple environmental signals with the functional mapping model. They considered a deterministic genotype-specific mean model for photosynthetic rate as a function of temperature and irradiance incidence and assumed a separable covariance structure using an AR(1) model for each factor. Although computationally attractive, this assumption may not be reasonable in cases when the factors are dependent. By considering temperature and irradiance incidence as two different processes, we propose using non-separable covariance functions, as developed in the spatial-temporal literature, to model the structure of temperature-and irradiance incidence-dependent covariance. We will show how simple structures such as those proposed by Cressie and Huang (1999) can be useful in implementing functional mapping. We provide extensive simulations to validate our approach and analyze a real data to illustrate its usefulness.

e-mail: jyap@stat.ufl.edu

MAPPING QUANTITATIVE TRAIT NUCLEOTIDES ENCODING COMPLEX DISEASES IN A NATURAL POPULATION WITH FAMILY STRUCTURE

Qin Li*, University of Florida
Arthur Berg, University of Florida
Rongling Wu, University of Florida

Genetic mapping has been developed to a point at which specific DNA sequence variants that encode a phenotypic trait can be identified with multiple single nucleotide polymorphism (SNP) markers. We present a sampling strategy and statistical model for detecting nucleotide combinations responsible for a complex disease. Our sampling strategy considers the inherent nature of many diseases, such as cancer, including a set of families (each composed of both parents and theiroffspring) randomly sampled from a natural population. By integrating the natural association of different markers (measured by the linkage disequilibrium) and their co-segregation during meioses (measured by the recombination fraction) in a two-stage hierarchical mixture setting, our statistical model provides the estimation and test of the distribution frequencies and quantitative effects of different nucleotide combinations on a phenotypic trait. Computer simulation is used to demonstrate the statistical behavior of the new model and its utilization.

e-mail: qli@biostat.ufl.edu

BAYESIAN QTL MAPPING FOR MULTIPLE TRAITS

Samprit Banerjee *, University of Alabama at Birmingham
Nengjun Yi, University of Alabama at Birmingham

Understanding the genetic architecture of complex traits continues to pose a bewildering challenge to contemporary statistical geneticists. Typically in a QTL mapping experiment data is collected on several correlated traits. Suprisingly, there is a lack of a comprehensive genome wide mapping strategy for correlated traits in the literature. Here, we develop the bayesian multiple traits QTL mapping using two multivariate models. One which assumes the same genetic model for all traits, the traditional multivariate model, and the other known as the Seemingly Unrelated Regression (SUR) model allows different genetic models for different traits. We compare the performance of these two methods with that of a single variate trait-by-trait analysis. We develop a model selection approach to map multiple QTL across the entire genome for both the multivariate models. There are a couple of strategies one could adopt to perform the MCMC with the SUR model. There is very little understanding on the theoretical behavior of these two strategies. So, we conduct an extensive simulation study to assess the performance of the traditional multivariate model and the SUR model with two different sampling strategies to the conventional univariate model in an objective manner.

e-mail: samban@uab.edu

A STATISTICAL MODEL FOR CHARACTERIZING cis- AND TRANS-ACTING REGULATION BY eQTL

Yao Li*, University of Florida
Jianhan Li, University of Florida
Ronglin Wu, University of Florida

Two different types of eQTL regulation have been thought to contribute to genetic variation in gene expression. For the first type, cis-regulating, the DNA variants of a gene directly influence the transcript levels of that gene and, therefore, the underlying eQTL can be expected to map to the structural gene producing the transcript. For the second type, trans-regulating, the DNA variants of a gene exert an indirect effect on its expression, in which the underlying eQTL should co-localize to the structuring gene. In this study, we propose a general statistical model based on linkage analysis to discriminate between cis- and trans-regulating eQTL and estimating their additive, dominant and epistatic effects, for the purpose of the identification of positional candidates in QTL studies. This model provides a useful tool for studying the genomic architecture of complex traits.

e-mail: yaoli@ufl.edu

92. APPLIED DATA ANALYSIS

CHANGING APPROACHES OF PROSECUTORS TOWARDS JUVENILE REPEATED SEX-OFFENDERS: A BAYESIAN EVALUATION

Dipankar Bandyopadhyay*, Medical University of South Carolina
Debajyoti Sinha, Florida State University
Stuart Lipsitz, Brigham and Women's Hospital
Elizabeth Letourneau, Medical University of South Carolina

Existing well-maintained state-wide data bases on prosecutor's decisions about charging and prosecuting juvenile offenders are important, however, yet un-explored resources to assess how the judicial decision-making processes change their patterns in reaction to societal changes and legislation of new laws. To investigate whether there is any dramatic change in judicial behavior towards youths charged with sexual offenses, following the enactment of a new set of mandatory sentencing laws between 1992 and 1996, we analyze the data on the prosecutor's decision on moving forward with the initial charge of sexual offense for adolescent offenders in South Carolina. We use a novel model for multivariate binary data with a random effects logistic regression model (controlling for repeated sexual offense charges from each youth) and incorporating an unknown change-point year. Our main model and analysis allow the logit structure both conditionally and marginally (integrating out random effects). Using a Bayesian perspective, we make a posteriori conclusions about whether a change-point has occurred in between 1992 to 1996 (inclusive) and the magnitude of the effects of the change-point and other factors on judicial decisions.

e-mail: bandyopd@musc.edu

VARIANCE ESTIMATION IN REGRESSION MODELS

Eugenio Andraca-Carrera*, University of North Carolina at Chapel Hill
Bahjat F. Qaqish, University of North Carolina at Chapel Hill

We introduce a new variance estimator for regression models. The estimator is a member of a larger class that includes several known estimators. The emphasis is on small-sample performance of confidence intervals based on the variance estimator. The estimator is evaluated in simulation studies and applications to real data are given.

e-mail: eandraca@bios.unc.edu

LIMITING THE IMPACT OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION VIA WEIGHT FUNCTIONS

Tamekia L. Jones*, University of Alabama at Birmingham
David T. Redden, University of Alabama at Birmingham

When utilizing Ordinary Least Squares within linear regression, statisticians often encounter diagnostic issues such as outliers, leverage points, influential observations, and the masking and swamping effects. These diagnostic issues can affect estimation of regression parameters leading to incorrect inference and inaccurate predictions in linear regression. We utilize weight functions that allow unequal weighting of atypical observations, which are identified by the Robust Forward Detection method, within a robust regression framework. Simulations are conducted to evaluate and compare the performance of the Robust Forward Detection method using weight functions to that of Ordinary Least Squares. Because the weights are functions of random variables, bootstrapping procedures are conducted to make inference. A dataset well-known to the robust regression literature is presented as an application of the Robust Forward Detection method utilizing the proposed weighting functions. We illustrate that our proposed approach allows robust estimation of parameters via weight functions by downweighting atypical observations identified by the Robust Forward Detection method.

e-mail: tljf81@uab.edu

GENOTYPE ADJUSTED FAMILIAL CORRELATION ANALYSIS USING THREE GENERALIZED ESTIMATING EQUATIONS

Hye-Seung Lee*, University of South Florida
Myunghee Cho Paik, Columbia University
Joseph H. Lee, Columbia University

We analyze familial correlation of a memory score from Caribbean Hispanic families that have multiple family members affected with Alzheimer's disease, adjusting for having at least 1 APOE-ε4 allele, as well as other confounders. To enhance the efficiency of correlation model, this paper proposes an efficient three generalized estimating equation for correlation model, extending the regression approach by Yan and Fine (2004). The efficiency of correlation model is evaluated through the asymptotic relative efficiency computation and simulations.

e-mail: leeh@epi.usf.edu

DIRECTIONAL DEPENDENCE OF TRUNCATION INVARIANT FGM COPULA FUNCTIONS

Yoon-Sung Jung*, Kansas State University
Jong-Min Kim, University of Minnesota at Morris
Engin A. Sungur, University of Minnesota at Morris

Directional dependence situations can be encountered in many research areas such as finance, biostatistics and bioinformatics. The directional dependence using copula functions has been introduced by Sungur (2005). We propose a new copula family which incoporates the truncation invariance into the generalized Farlie-Gumbel-Morgenstern (FGM) distributions which are two variables under truncation applied to the third variable. The directional dependence of the new truncated invariant FGM copula model will be studied in this paper. We will show that there exists a directional dependence in our truncated invariant FGM copula family. Since the dependence structure using a copula does not need the normality and independence assumptions, the existence of the directional dependence in our truncated invariant FGM copulas will be helpful to investigate the dependence structure of various fields such as financial or survival data. An example of the truncation invariant dependence structures is illustrated by using foreign exchange currency data.

e-mail: ysjung72@gmail.com

PRACTICAL APPLICATION OF STATISTICS IN CLINICAL STUDIES

Jay Mandrekar*, Mayo Clinic College of Medicine

Research in academic medical centers offer opportunities to collaborate on clinical projects that involve novel applications of both common and uncommon statistical methods. In this talk illustrative examples from diverse clinical areas of medical education, imaging, clinical microbiology and physical medicine and rehabilitation will be presented. Specifically, statistical approaches including; 1) Graeco Latin Square design to compare the performance of four liquid crystal displays under typical medical center lighting conditions, 2) factor analysis to evaluate stability of factorial structure of clinical teaching assessments among medical specialties, 3) determination of the predictive utility of continuous versus categorical scaling of duration of post-traumatic amnesia, and 4) Deming's regression to compare performance of various assays in clinical microbiology studies, will be discussed. Findings from these studies directly impact patient care and/or offer cost savings with improved efficiency.

e-mail: mandrekar.jay@mayo.edu

EXPERIENCES IN A MASTERS-LEVEL COURSE IN BIOSTATISTICAL CONSULTING

Stephen W. Looney*, Medical College of Georgia
Jennifer L. Waller, Medical College of Georgia

In this presentation, we describe a core course in the Master of Science in Biostatistics program at the Medical College of Georgia entitled "Biostatistical Consulting in Research." The purpose of this course is to provide the students with instruction in the basic principles of effective statistical consulting and collaboration and to give them some 'hands-on' experience as statistical collaborators. At the beginning of the course, students are provided with instruction on how to make an effective oral presentation and preferred methods for presenting the results of statistical analyses in written and oral reports. Each student then participates as a full team member in at least one collaborative research project in the health sciences, under the supervision of a faculty member in the Department of Biostatistics. The student is required to submit a written report and make a 20-minute oral presentation summarizing the study findings at the end of the course. The emphasis throughout the course is on the effective communication of statistical results to non-statisticians. The "human side" of statistical consulting is also a primary focus of the course. In this presentation, we describe our experiences as co-directors of the course, and offer suggestions on how it might be improved.

e-mail: slooney@mcg.edu

93. LONGITUDINAL MODELS-DISCRETE

ANALYSIS OF THE SECOND LONGITUDINAL STUDY OF AGING

Hyokyoung Hong*, University of Illinois at Urbana-Champaign

We develop a predictive model for the functional status of the elderly people based on data from the Second Longitudinal Study of Aging (LSOAII). The functional status is an ordinal response variable. The ordered probit model has been moderately successful in analyzing ordinal response data, but it relies on the assumption of normal distributions for its latent variable. Our approach focuses on the prediction of conditional quantiles based on a more general transformation model. We show that our proposed model can achieve better accuracy in prediction without model misspecification bias. Cross-validation is used to assess the performance of competing models and methods. Monte Carlo simulations are also used to demonstrate the merits of our approach.

e-mail: hhong3@uiuc.edu

MARGINALIZED MODELS FOR LONGITUDINAL COUNT DATA

Keunbaik Lee*, Louisiana State University-Health Science Center
Michael Daniels, University of Florida
Yongsung Joo, University of Florida

Generalized linear models with serial dependence are often used for short longitudinal series. Heagerty (2002) and Lee and Daniels (2007) have proposed marginalized transition models for the analysis of longitudinal binary data and ordinal data, respectively. In this paper, we extend these work to accommodate longitudinal count data. We also propose a model to handle overdispersed data. Fisher-scoring algorithms are developed for estimation. Methods are illustrated with a real dataset and are compared with other statistical methods.

e-mail: klee4@lsuhsc.edu

LIKELIHOOD ANALYSIS OF JOINT MARGINAL AND CONDITIONAL MODELS FOR LONGITUDINAL CATEGORICAL DATA

Baojiang Chen*, University of Waterloo
Grace Y. Yi, University of Waterloo
Richard J. Cook, University of Waterloo

Analyses of longitudinal categorical data are typically based on semiparametric models in which covariate effects are expressed on marginal probabilities and estimation is carried out based on generalized estimating equations (GEE). Methods based on GEE are motivated in part by the lack of tractable models for clustered categorical data. However, such marginal methods do not yield fully efficient estimates, nor consistent estimates when data are missing at random since the model is not fully specified. We develop a Markov model for the analysis of longitudinal categorical data which facilitates modeling marginal and conditional structures separately. A likelihood formulation is employed for inference, so the resulting estimators enjoy the optimal properties such as efficiency and consistency, and remain consistent when data are missing at random. Simulation studies demonstrate that the proposed method performs well under a variety of situations. Application to data from a smoking prevention study illustrates the utility of the model and interpretation of covariate effects.

e-mail: b7chen@uwaterloo.ca

INTEGER-VALUED AUTOREGRESSIVE MODELS FOR ANALYSIS OF LONGITUDINAL COUNT DATA

Mohamed Alosh*, U.S. Food and Drug Administration

Longitudinal count data arise in many scientific disciplines including clinical trials. In this presentation we consider a class of integer-valued autoregression (INAR) models to capture the serial correlation in count data. In particular we introduce an extension of the Poisson INAR model to account for the over-dispersion phenomena usually observed in data of the active treatment arm in a clinical trial setting. An application of the proposed model to longitudinal data from a clinical trial will be presented.

e-mail: Mohamed.Alosh@fda.hhs.gov

A NON-HOMOGENEOUS MARKOV PROCESS MODEL FOR ALZHEIMER'S DISEASE PROGRESSION

Rebecca A. Hubbard*, University of Washington
Xiao-Hua Zhou, University of Washington

Identifying individuals at high risk of developing Alzheimer's disease (AD) is important for understanding the natural history of the disease and effectively targeting interventions. Subjects suffering from mild cognitive impairment (MCI) are one group with high probability of progression to AD. Estimating rate of progression from normal cognition to MCI and AD is challenging because cognitive status is ascertained only at irregular follow-up times giving rise to interval censored data. Moreover, progression rates are known to be non-constant with respect to age. Markov process models are useful for characterizing transition rates in multi-state disease processes with interval censoring. However, limited methods exist for temporally non-homogeneous multi-state processes. We propose a non-homogeneous Markov process model to characterize transitions between disease states defined by normal cognition, MCI, AD, and death. Risk factors for increased rates of transition are introduced via a regression model for elements of the baseline transition intensity matrix. We apply this model to a longitudinal study of subjects evaluated at the Alzheimer's Disease Centers.

e-mail: rhubb@u.washington.edu

SEMIPARAMETRIC MODELS FOR BIVARIATE PANEL COUNT DATA

Li-Yin Lee*, University of Wisconsin
KyungMann Kim, University of Wisconsin

Bivariate panel count data arise when two types of recurrent events are under investigation and each study subject is examined only at discrete time points. The data consist only of event counts that have occurred prior to observation times, while the exact event times remain unknown. While numerous methods of analyzing univariate panel count data have been proposed, methods for analysis of bivariate panel count data have not been investigated. We propose semiparametric methods for bivariate panel count data with covariates based on the isotonic regression estimators (Sun and Kalbfleisch, 1995; Wellner and Zhang, 2000). We assume that, given a vector of covariates and a frailty variable, the underlying bivariate counting process are two independent nonhomogeneous Poisson processes. The dependence of event processes is modeled by a frailty variable which takes into account both the subject heterogeneity and the correlation of event types. The methods provide both model-based inference of the covariates and the nonparametric inference of baseline mean functions for the bivariate counting process. The methods are illustrated via an analysis of data from a cancer chemoprevention trial on the effectiveness of difluoromethylornithine (DFMO) in preventing the non-melanoma skin cancer in the population with previous skin cancer.

e-mail: liyinlee@wisc.edu

TESTING THE EFFECT OF AN EXPOSURE ON LONGITUDINAL BINARY OUTCOMES WHEN EVENTS ARE RARE

Xiaonan Xue*, Albert Einstein College of Medicine
Mimi Y. Kim, Albert Einstein College of Medicine
Tao Wang, Albert Einstein College of Medicine
Howard D. Strickler, Albert Einstein College of Medicine

This work is motivated by a large prospective cohort study of HIV-seropositive women in which multiple Human Papillamavirus (HPV) types, squamous intraepithelial lesions (SIL) and high level squamous intraepithelial lesions (HSIL) were repeatedly assessed at 6-month intervals. The goal of the study is to examine the effects of human leukocyte antigen (HLA) polymorphisms on HPV infection, SIL and HSIL in both HIV + and HIV- women. However, the HPV outcomes such as HSIL are rare, and even more so in the HIV - group. Consequently, traditional methods such as generalized estimating equation (GEE) models on binary outcome fail to converge. Alternative methods that can be used in such situation are exact methods but they are often too conservative and involve intensive computations without readily accessible software. In this paper, we propose a Monte-Carlo approach to obtain an empirical p-value for testing the exposure and outcome association. The performance of this approach is evaluated via simulations studies and is compared with GEE methods.

e-mail: xxue@aecom.yu.edu

94. SURVIVAL ANALYSIS - COMPETING RISKS

REGRESSION ANALYSIS FOR BIVARIATE FAILURE TIME ASSOCIATIONS IN THE PRESENCE OF A COMPETING RISK

Jing Ning*, M.D. Anderson Cancer Center
Karen Bandeen-Roche, Johns Hopkins Bloomberg School of Public Health

This talk proposes regression modeling for the conditional cause-specific hazard ratio, a generalization of the conditional hazard ratio that accommodates competing risks data. We model the conditional cause-specific hazard ratio as a parametric regression function of time, event causes and other covariates. We develop a pseudo-likelihood estimation procedure for fitting and inference and compare this with a moment estimating equations approach. Asymptotic properties are developed, and small sample performance is evaluated through a simulation study. Data from the Cache County Study on dementia are used to illustrate the proposed methodology.

e-mail: jning@mdanderson.org

CLASSIFICATION TREES FOR SURVIVAL DATA WITH COMPETING RISKS

Fiona M. Callaghan*, University of Pittsburgh
Chung-Chou H. Chang, University of Pittsburgh

Classification trees are the most popular tool for categorizing individuals into groups and subgroups based on particular outcomes of interest. To date, trees have not been developed to deal with survival data involving competing risks. In this study, we propose two classification trees to analyze data with competing risks: a tree that maximizes between-node heterogeneity and a tree that maximizes within-node homogeneity. After we describe the methods used in growing and pruning the trees, we demonstrate and compare their performance with simulations in a variety of competing risk model configurations. We also illustrate their use by analyzing survival data concerning patients who had end-stage liver disease and were on the waiting list to receive a liver transplant.

e-mail: fmc2@pitt.edu

MULTI-CANCER EMERGENCE RATE ESTIMATION WITH SCHEDULED DIAGNOSES: A TRANSITION PROBABILITY APPROACH

Junfeng Liu*, University of Medicine and Dentistry of New Jersey and Cancer Institute of New Jersey
Weichung Joe Shih, University of Medicine and Dentistry of New Jersey and Cancer Institute of New Jersey

In order to empirically calculate the so-called ‘crude’ survival functions, multiple terminal events in competing risks model are usually assumed to be reported right at initial emergence. However, certain realistic data collection modes from competing risks model may involve ‘inactiveness’, where terminal states will only be reported at scheduled check points rather than emergence time. Data obtained herein may not be sufficient for straightforward routine survival analysis. However, we demonstrate that a one-way discrete multi-state Markov transition model could be applied to such seemingly ‘incomplete’ competing risks models in connection to latent emergence time estimation and ‘crude’ survival function calculation. The applied parameters are time-varying transition probabilities from the starting blank state to multiple terminal states at different check points, which interestingly lead to a novel expectation-maximization (EM) algorithm. The proposed algorithm is evaluated under different model configurations for practical purposes. Our results justify the feasibility of estimating survival functions from scheduled state report, thus real-time data monitoring may be avoided in large-scale medical research involving competing risks such as temporal multi-cancer emergence rate estimation at the population level.

e-mail: ljfbacc@yahoo.com

MULTIPLE IMPUTATION INSPIRED BY PSEUDO-VALUE APPROACH FOR CENSORED SURVIVAL DATA

Lyrica X. Liu*, University of Michigan
Susan Murray, University of Michigan
Alex Tsodikov, University of Michigan

One of the more prominent methods for handling censored data is through imputation of censored events consistent with the observed data. Existing methods either ignore patient characteristics when imputing a likely event time, or place quite restrictive modeling assumptions on the method for imputation. This research develops a method of retaining patient characteristics when making imputation choices, while substantially reducing the assumptions needed to impute. There is a theoretical link in this research to previous literature generating pseudo-values for analysis.

email: lyrica@umich.edu

TESTING AND ESTIMATION OF TIME-VARYING CAUSE-SPECIFIC HAZARD RATIOS WITH COVARIATE ADJUSTMENT

Yanqing Sun, University of North Carolina at Charlotte
Seunggeun Hyun*, University of South Carolina Upstate
Peter Gilbert, Fred Hutchinson Cancer Research Center and University of Washington

In the evaluation of efficacy of a vaccine to protect against disease caused by a genetically diverse infectious pathogen, it is often important to assess if and how vaccine protection depends on variations of the exposing pathogen. This problem can be viewed within the framework of a competing risks model where the endpoint event is pathogen-specific infection and the cause of failure is the strain type determined after the infection is diagnosed. The Cox model with time-dependent coefficients is used to relate the cause-specific outcomes to explanatory variables to allow for time-varying treatment effects. The strain specific vaccine efficacy can be defined in terms of one minus the cause-specific hazard ratios. We develop inferential methods for testing if the vaccine affords some protection against at least one pathogen strain, and for testing equal vaccine protection against the strains, adjusting for covariate effects. We also consider estimation of covariate-adjusted time varying strain-specific vaccine efficacy. The methods are applied to data from an oral cholera vaccine trial and the performances of the proposed tests are studied through simulations. These techniques apply more generally for testing and estimation of time-varying cause-specific hazard ratios.

e-mail: shyun@uscupstate.edu

BAYESIAN SEMI-PARAMETRIC REGRESSION UNDER COMPETING RISKS SETTING

Xiaolin Fan*, Medical College of Wisconsin
Purushottam W. Laud, Medical College of Wisconsin

Competing risks data have appeared frequently in medical studies. Regression on cause-specific hazards is a standard analysis in such cases to assess covariate effects. Recently, Fine and Gray(1999) developed a proportional hazards model directly on the sub-distribution (cumulative incidence function) of the cause of interest. From Bayesian nonparametric standpoint, the baseline(normalized) cumulative incidence function can be assigned a prior. We use a mixture of Polya trees process as a prior. Following the work of Hanson(2005) we implement, using the full likelihood, inference for the Fine and Gray model through Metropolis-Hastings chains. Other regression models on cumulative incidence functions can be analyzed similarly by adapting the general method through appropriate changes in the likelihood function. Such models include additive hazards, proportional odds and accelerated failure time. Model extension to hierarchical structure is also discussed, implemented and illustrated via an example. We also note that the methods used here easily adapt to models for cause-specific hazards.

e-mail: xfan@mcw.edu

NONPARAMETRIC ESTIMATION OF CAUSE-SPECIFIC CROSS HAZARD RATIO WITH BIVARIATE COMPETING RISKS DATA

Yu Cheng*, University of Pittsburgh
Jason P. Fine, University of Wisconsin-Madison

We propose an alternative representation of the cause-specific cross hazard ratio for bivariate competing risks data. The representation leads to a simple plug-in estimator, unlike an existing ad hoc procedure. The large sample properties of the resulting inferences are established. Simulations and a real data example demonstrate that the proposed methodology may substantially reduce the computational burden of the existing procedure, while maintaining similar efficiency properties.

e-mail: yucheng@pitt.edu

95. CLINICAL TRIALS - GENERAL

PREDICTING THE DURATION OF A SEQUENTIAL TRIAL FROM INTERIM BLINDED DATA

W. J. Hall*, University of Rochester

Consider a two-arm group-sequential clinical trial, with survival-analysis-based monitoring, and interim analyses scheduled at multiples of event counts (information). While the trial continues, those directing the trial may want an estimate, from blinded data, when the trial will end. If nonsequential, it is purely a question of using the current data, including recruitment information, losses to follow-up, and event occurrences to predict how long it will take to accumulate the needed total number of events. In a group-sequential trial, one can do the same for each analysis number (and event count) at which stopping could occur. A final step is to relate each analysis number to the range of estimated hazard ratios likely to obtain should stopping occur at that analysis. We thus develop a relationship between the hazard ratio, to be estimated when the trial ends, and trial duration. This system is easy to communicate to medical investigators, as duration is associated with a guess at what the estimated hazard ratio will be, a quantity they are more comfortable with than a true hazard ratio. We develop formulas to facilitate such predictions, and illustrate them with data from a recent clinical trial.

e-mail: hall@bst.rochester.edu

ON NON-INFERIORITY ASSESSMENT FROM BINARY DATA IN THE COMBINATION OF 2 BY 2 TABLES

Kallappa M. Koti*, U.S. Food and Drug Administration

We propose a stratified analysis for testing whether a new treatment is at least as effective as the standard treatment in comparative binomial trials. We focus on relative risk. We apply the restricted maximum likelihood methods in deriving the new test. We discuss the power and sample size determination in terms the ratio of averages of strata proportions. We illustrate the test procedure and explain its interpretation using a dataset with eight 2 by 2 tables.

e-mail: kallappa.koti@fda.hhs.gov

A SURVEY OF THE LIKELIHOOD APPROACH TO BIOEQUIVALENCE

Leena Choi*, Vanderbilt University School of Medicine
Brian Caffo, Johns Hopkins University Bloomberg School of Public Health
Charles Rohde, Johns Hopkins University Bloomberg School of Public Health

Bioequivalence trials are abbreviated clinical trials whereby a generic drug or new formulation is evaluated to determine if it is ‘equivalent’ to a corresponding previously approved brand-name drug or formulation. In this manuscript, we survey the process of testing bioequivalence and advocate the likelihood paradigm for representing the resulting data as evidence. We emphasize the unique conflicts between hypothesis testing and confidence intervals in this area - which we believe are indicative of the existence of the systemic defects in the frequentist approach - that the likelihood paradigm avoids. We suggest the direct use of profile likelihoods for evaluating bioequivalence. We discuss how the likelihood approach is useful to present the evidence for both average and population bioequivalence within a unified framework. This is the fundamental step that is missing in the current practice of bioequivalence trials. We also examine the main properties of profile likelihoods and estimated likelihoods under simulation. This simulation study shows that profile likelihoods are a reasonable alternative to the (unknown) true likelihood for a range of parameters commensurate with bioequivalence research.

e-mail: leena.choi@vanderbilt.edu

BAYESIAN SCREENING FOR PHARMACOGENETIC EFFECTS IN CLINICAL TRIALS

Mengye Guo*, University of Pennsylvania
Daniel F. Heitjan, University of Pennsylvania

Statistical methods for pharmacogenetic analysis aim to identify gene-by-treatment interactions by testing the significance of the interaction terms in a model predicting outcome from treatment and genotype. Standard frequentist tests of interaction are typically too sensitive, especially when the sample size is large. Moreover, they fail to make use of prior information, such as information on the likely roles of particular

genes in the phenotype in question. To address these concerns, we develop a Bayesian hypothesis testing method (Berger 1985). Our testing criterion, which is based on Bayes factors for comparing the null and alternative models (Kass and Raftery, 1995), is more conservative than an uncorrected frequentist test, but less conservative than a multiplicity-corrected version of the test that adjusts for the number of markers under study. We apply our method to a randomized trial of pharmacotherapy for smoking cessation, in which a number of SNPs were evaluated as potential pharmacogenetic markers.

e-mail: mengyego@mail.med.upenn.edu

STATISTICAL MODELING AND GRAPHICAL ANALYSIS OF SAFETY DATA IN CLINICAL TRIALS

Michael O'Connell*, Insightful
Dawn Woodard, Duke University

Rigorous assessment of drug safety, pre- and post- drug approval, is essential to protect and promote public health. However, clinical trial design is focused on the efficacy endpoints, and safety data analysis is typically an afterthought. For example, large amounts of safety data are collected in clinical trials, but basic information such as which types of patients have adverse events or elevated lab values, is not well captured or summarized in statistical analyses and clinical study reports. This talk will focus on statistical and graphical methods for analysis and reporting of safety data. Statistical methods considered include hierarchical Bayes models and and exploratory analyses such as trees and forests. Results from these analyses are presented as S-PLUS® graphical summaries that highlight key aspects of drug safety, such as risk difference and effect probabilities for adverse events. Such safety data analyses and reports have immense value for pharmaceutical companies, drug safety monitoring boards and regulatory agencies such as the FDA.

e-mail: moconnell@insightful.com

ASSESSMENT OF KEY FACTORS FOR OVERALL HEALTH RELATED QUALITY OF LIFE

Peng Zhao*, University of Southern California

Health-related quality of life (HRQOL) is emerging as a conceptualization that is measurable and could be used as a quality indicator. The purpose of this analysis was to examine empirically the association of the four main dimensions (socio-demographic characteristics, symptom amplification, biological factors and psychosocial factors) with HRQOL, particularly, through the mediation of general health perceptions. Based on Wilson and Cleary's model, a series of revised simplified conceptual models were proposed. Through testing these hypotheses, it was found that the best fitted model by conducting structural equation modeling analysis, correlations between each pair of the four first level factors and their direct and indirect effect on the general health perceptions and overall quality of life were explored. The analysis results showed that model 3 overall was a compelling explanation of overall quality of life, and the evidence favors the mediation of general health perceptions. It confirmed that overall the data were consistent with the revised Wilson and Cleary conceptual model, as results accounted for 84 percent of the variance in overall quality of life.

e-mail: pengzhao08@hotmail.com

BENEFIT IN PROGRESSION-FREE SURVIVAL (PFS) WITH A GENASENSE-DACARBAZINE (GENASENSE-DTIC) REGIMEN IN ADVANCED MELANOMA: A CASE STUDY ON ASSESSMENT SYMMETRY

Richard Kay*, R.K. Statistics-United Kingdom
Erard Gilles, Genta Incorporated
Jane Wu, Genta Incorporated
Alexander M.M. Eggermont, Erasmus University Medical Center-The Netherlands

In a randomized study (N=771) of patients with advanced melanoma, DTIC was administered alone or following a 5-day Genasense infusion. Patients were assessed on the same schedule by design. Intent-to-treat analysis of PFS showed significant benefit with Genasense+DTIC (HR=0.73, p=0.0003). Unlike survival, unequivocal determination of the timing of progression is difficult. When these PFS data were reviewed by the Food and Drug Administration (FDA), the agency performed a series of simulations to evaluate PFS assessment bias. The FDA model was based on the unrealistic assumption that all assessments in both groups occurred on the same day. This presentation will compare the assumptions and results of the FDA model with those of a model based on the realistic assumption of a normal distribution of assessments with a standard deviation of 10 days. Simulation can be useful in assessing the potential for bias. However, simulation studies must be based on realistic assumptions or the results can be very misleading. Sensitivity analyses and consistency of efficacy results are meaningful measures of the robustness of PFS findings. In the randomized study of Genasense+ DTIC, the PFS benefit was due to a treatment effect of Genasense when added to DTIC and was not the result of assessment bias.

e-mail: wu@genta.com

96. METHODS IN SPATIAL MODELING

HIERARCHICAL MODELS FOR LARGE SPATIALLY-REFERENCED BINARY DATA WITH AN APPLICATION TO SURVIVAL OF TROPICAL TREE SEEDLINGS

Sang Mee Lee*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Liza Comita, University of Minnesota

With accessibility to geocoded locations where scientific data are collected through Geographical Information Systems (GIS), investigators are increasingly turning to hierarchical spatial process models for carrying out statistical inference. However, estimating hierarchical spatial models involves expensive matrix decompositions whose computational complexity increases dramatically with the number of spatial locations rendering them infeasible for large spatial data sets. In this talk we consider hierarchical spatial logistic models for estimating the probability of survival for seedlings from a large trial as a function ofseveral spatially referenced covariates. We consider two approaches: the first models the spatial association of observations from each cell as a function of its neighbors using a conditional autoregressive (CAR) model; the second approach introduces a dimension-reducing geostatistical process that directly models spatial correlations. These methods help produce probability maps over the spatial domain that can assist in identifying zones of high and low survival rates.

e-mail: leex2919@umn.edu

RAMPS: AN R PACKAGE FOR UNIFIED GEOSTATISTICAL MODELING OF COMPLEX SPATIOTEMPORAL DATA

Brian J. Smith*, University of Iowa
Jun Yan, University of Connecticut
Mary K. Cowles, University of Iowa

We introduce and demonstrate the R package 'ramps' which implements reparameterized and marginalized posterior sampling (RAMPS) for complex Bayesian geostatistical models. Our package allows joint modeling of areal and point-source data arising from the same underlying spatial process. The reparametrization of variance parameters facilitates a slice sampling algorithm based on simplexes, which can be useful in general when multiple variances are present. Prediction at arbitrary points can be made, which is critical in applications where maps are needed. The implementation takes advantage of sparse matrix operations in the 'Matrix' package and can provide substantial savings in computing time for large datasets. A user-friendly interface, similar to the linear mixed effect model package, enables users to analyze datasets and plot results with little programming effort. Support is provided for numerous spatial and spatiotemporal correlation structures, user-defined correlation structures, and non-spatial random effects. The package features are illustrated throughout the presentation via an analysis of residential radon measurements collected in Iowa.

e-mail: brian-j-smith@uiowa.edu

SMOOTHED ANOVA WITH SPATIAL EFFECTS AS A COMPETITOR TO MCAR IN MULTIVARIATE SPATIAL SMOOTHING

Yufen Zhang*, University of Minnesota
James S. Hodges, University of Minnesota
Sudipto Banerjee, University of Minnesota

Smoothed ANOVA (SANOVA; Hodges et al 2007 Technometrics) is a way to smooth ANOVA that shrinks interactions by embedding SANOVA in a hierarchical model to do shrinkage. Instead of simply shrinking effects without any structure, SANOVA can use spatial structure to smooth effects. In this talk, we will extend SANOVA to model cases in which one factor is a spatial lattice, which is smoothed using a conditional autoregressive model (CAR), and a second factor is, for example, type of cancer. As such, SANOVA may be a competitor to the multivariate CAR (MCAR) model. Simulations are done to compare SANOVA under different design matrix settings versus MCAR under different prior settings, with advantages of each approach discussed. A cancer-surveillance dataset, describing incidence of 3 cancers in Minnesota's 87 counties, is analyzed using different methods and their results are discussed.

e-mail: yufenz@biostat.umn.edu

MINING EDGE EECTS IN AREALLY REFERENCED SPATIAL DATA: A BAYESIAN MODEL CHOICE APPROACH

Pei Li*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Alexander M. McBean, University of Minnesota

Statistical models for areal (or region-specific) data are primarily used for smoothing maps revealing spatial trends. Subsequent interest often resides not in the statistically estimated maps themselves, but on the formal identification of 'edges' or 'boundaries' on the map. Here boundaries or edges refer not to the geographical, political or administrative boundaries, but rather to 'difference boundaries' representing significant differences between adjacent regions. One approach is to develop hierarchical models that attempt to assign stochastic distribtions to the adjacency matrix. Another approach is to treat this as a model comparison problem where the models correspond to different underlying edge configurations across which we wish to smooth (or not). We incorporate these edge configurations in spatial autoregression models (SAR and CAR) and demonstrate how the Bayesian Information Criteria (BIC) can be used to detect difference boundaries in the map. We illustrate and compare these strategies with public health data extracted the SEER-Medicare merged database concerning pneumonia hospitalizations in elderly patients.

e-mail: lixx0525@umn.edu

NONSTATIONARY SPATIAL GAUSSIAN MARKOV RANDOM FIELD PRIORS

Yu Yue*, University of Missouri-Columbia
Paul L. Speckman, University of Missouri-Columbia

Thin-plate splines have been widely used as spatial smoothers. In this paper, we present a Bayesian version of a discretized adaptive thin-plate spline suitable for modeling nonstationary spatial data. The main idea is to extend two dimensional intrinsic Gaussian Markov random fields (IGMRFs) by using a spatially adaptive variance component and taking a further IGMRF prior for this variance function. Fully Bayesian inference can be carried out through efficient Markov chain Monte Carlo simulation. Performance is demonstrated with a simulated data set and by an application to a rainfall data set.

e-mail: yytc9@mizzou.edu

LOGISTIC JOINPOINT REGRESSION MODELS IN COHORT STUDIES

Ryan Gill*, University of Louisville
Grzegorz Rempala, University of Louisville

In studying trend data such as cancer mortality and incidence data we are frequently concerned with detecting a change in recent trend. In this talk, we consider the extension of the simple Gaussian joinpoint regression model to logistic regression and possibly non-homogenous dispersion parameters. We derive the maximum likelihood estimates of the joinpoint as well as the regression parameters. Since the location of the joinpoints (change points) in the model is unknown, the method employs an iterative conditional maximization algorithm in seeking the solutions of the likelihood equations. In order to test the validity of the final model as well as to assess the significance of the final set of detected joinpoints, we sequentially apply the parametric bootstrap method using a backward and forward algorithm. Additionally, we compare the performance of the joinpoint logistic regression model versus that of penalized splines (P-splines). Finally, we also apply the developed model to the longitudinal dataset on cancer mortality among the members of the Louisville VC cohort of now retired chemical workers up until 1996 using an R package that implements the methods discussed in this talk.

e-mail: rsgill01@louisville.edu

CONSISTENT NONPARAMETRIC INTENSITY ESTIMATION FOR INHOMOGENEOUS SPATIAL POINT PROCESSES

Yongtao Guan*, Yale University

Nonparametric intensity estimators based on local smoothing are generally inconsistent for the true intensity function for inhomogeneous spatial point processes. In this paper, we introduce a new estimator based on covariate smoothing. When the intensity function of the spatial point process is a continuous function of some observed covariates, we prove that the proposed estimator is consistent for the true intensity under some mild conditions. To handle the situation with multiple covariates, we also extend sliced inverse regression, a popular dimension reduction technique, to the spatial point process setting.

e-mail: yongtao.guan@yale.edu

97. COLLABORATION IN HIV RESEARCH: TWO CASE STUDIES

USING MATHEMATICAL-STATISTICAL MODELING TO INFORM THE DESIGN OF HIV TREATMENT STRATEGIES AND CLINICAL TRIALS

Eric S. Rosenberg*, Clinician/Scientist, Massachusetts General Hospital and Harvard Medical School
Marie Davidian*, Statistician, North Carolina State University

Much progress has been made in management of HIV infection using highly active antiretroviral (ARV) therapies, but continuous treatment involves significant cost, burden, toxicities, drug resistance, and adherence problems. This has led to interest in 'structured treatment interruption' (STI) strategies involving cycles of ARV withdrawal and initiation; however, studies of STI have been inconclusive, in part because the strategies evaluated may not have been the most advantageous to study. The speakers are part of a multidisciplinary team comprising a statistician and an immunologist/infectious disease clinician (the speakers) as well as mathematicians and control theorists who are investigating the use of mathematical dynamical systems models of the within-subject interplay between HIV and the immune system, embedded in a statistical framework that facilitates simulation of the disease progression in 'virtual subjects', to design HIV treatment strategies and clinical trials to study them. In this presentation, the first speaker will review the relevant background regarding HIV infection. The second speaker will describe the mathematical-statistical model and simulation framework. Both speakers will emphasize how statistical, mathematical, biological, and clinical science are integrated to support this unique collaboration.

e-mail: davidian@stat.ncsu.edu

MODELING AND ESTIMATION OF KINETIC PARAMETERS AND REPLICATIVE FITNESS OF HIV-1 FROM FLOW-CYTOMETRY-BASED GROWTH COMPETITION EXPERIMENTS

Carrie Dykes*, Clinician/Scientist, University of Rochester School of Medicine and Dentistry
Hulin Wu*, Statistician, University of Rochester School of Medicine and Dentistry

Growth competition assays have been developed to quantify the relative fitness of HIV-1 mutants. In this talk we illustrate how to develop mathematical models, statistical methods and computational software tool to estimate viral fitness parameters from in vitro experiments by close collaboration between experimentalists and statisticians/modelers. The nonlinear differential equation (NDE) models are developed to describe viral/cellular dynamic interactions in the assay system from which the competitive fitness indices or parameters are defined.
The model identifiability will be discussed and parameter estimation methods for NDE models are developed. A web-based computational tool is developed for HIV virologists to use for viral fitness index calculations. In this talk we (Dr. Dykes, HIV Virologist and Dr. Wu, Statistician/Modeler) will share our experience for a multidisciplinary team (virologists/experimentalists, statisticians, modelers and software developers) to collaborate and communicate to each other to solve an important biological problem in biomedical research.

email: hwu@bst.rochester.edu

98. JOINT MODELING APPROACHES FOR LONGITUDINAL DATA UNDER COMPLEX STUDY DESIGNS

CASE-CONTROL STUDIES WITH LONGITUDINAL COVARIATES

Honghong Zhou*, Schering-Plough Corporation
Bin Nan, University of Michigan
Xihong Lin, Harvard School of Public Health

The case-control design is commonly used for studying rare diseases. In a case-control study, subjects are recruited according to their case or control status, and thus the sampling is outcome-based, retrospective and biased. An emerging problem in case-control studies is the presence of a longitudinal covariate, which is collected longitudinally but retrospectively. We consider 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 study several estimation procedures, such as the Two-stage method, the Best Linear Unbiased Predictor (BLUP) method, the Maximum Likelihood Estimation (MLE) based on the true retrospective likelihood, and the Sufficiency Score method. The asymptotic properties of these methods are also discussed. The proposed methods are illustrated through simulation studies and by application to a case-control study of breast-cancer in postmenopausal women with weight as the longitudinal covariate.

e-mail: honghongz@gmail.com

MODELING MULTIVARIATE LATENT TRAJECTORIES AS PREDICTORS OF A UNIVARIATE OUTCOME

Sujata Patil*, Memorial Sloan Kettering Cancer Center
Trivellore E. Raghunathan, University of Michigan
Jean T. Shope, University of Michigan

Changes in the two serum markers, alpha-fetoprotein (AFP) and human chorionic gonandotrophin (HCG), while on therapy is one criterion that has been used to determine whether germ cell cancer relapse is likely to have occurred. The interrelationships between the two longitudinal markers and how they are associated with relapse are of clinical interest. A two-stage approach is one way to investigate these types of research questions and can be implemented easily. First each longitudinal marker is summarized by a few latent trajectory variables (such as subject-specific intercepts and slopes). In stage two of the model, these summary latent trajectory variables are used as predictors of the endpoint. It is important to note that the latent trajectory variables are estimated and have measurement error. Since ignoring measurement error can affect statistical inference, we develop a fully Bayesian joint approach to obtain estimates for model parameters. In this talk, we will describe a joint estimation approach for a model where two or more longitudinal markers are predictors of a subsequent univariate outcome. The proposed approach is applied to an example dataset of germ-cell cancer patients. Results from a simulation study and methodological challenges in fitting these models will be discussed.

e-mail: patils@mskcc.org

STATISTICAL METHODS FOR ADJUSTING SELECTION BIAS IN LONGITUDINAL SURVEY STUDIES IN AGING RESEARCH

Wen Ye*, University of Michigan

The Health and Retirement Study (HRS), aiming at the American population over age 50, tracks a cohort of individuals from working ages into retirement, collecting economic, demographic and biomedical information every two years. These longitudinal survey data sets allow researchers to describe and make inference about the dynamics of health changes in the aging process that can not be drawn from cross-sectional data. However, several complicated sampling issues are involved in utilizing these longitudinal data, e.g. informative dropout and cohort heterogeneity. Another less addressed problem is that, comparing to a subject from a more recent cohort, a subject from an earlier cohort must survive longer to be included in the study. This cause selection bias and leads to biased estimate of the population longitudinal trajectory. It also often leads to a paradox that the older cohorts seem healthier than the younger cohorts. To adjust the bias we use a maximum likelihood joint modeling approach. In this model, the selection process and informative dropout is modeled by a cox proportional hazards model with left truncation. The change of health indices over age, e.g. functional status (activities of daily living (ADL) and instrumental activities of daily living (IADL) difficulties), is modeled by a simple linear mixed model. EM algorithm is used to estimate model parameters.

e-mail: wye@umich.edu

99. MULTI-TASK LEARNING FOR BORROWING INFORMATION FROM DISPARATE DATA SOURCES

SEMIPARAMETRIC BAYES BORROWING OF STRENGTH IN RELATED ANALYSES

David B. Dunson*, National Institute of Environmental Health Sciences

With improvements in biotechnology, it has become routine to collect a wide variety of measurements on each subject in a study, with the number of measurements often large relative to the number of subjects. In such settings, it is important to consider statistical methods that allow borrowing of strength across multiple, related analyses. In the machine learning literature, such problems are referred to as multi-task learning. Although multi-task learning is related to meta analysis, traditional meta analysis techniques are often insufficiently flexible, since the different tasks often have fundamentally different scales and are not exchangeable. To address this problem, this talk considers methods based on semiparametric Bayes methods, relying on local clustering to borrow strength. In particular, we propose a matrix stick-breaking process (MSBP), which is more flexible than the Dirichlet process, and apply the MSBP to several applications. One application is to toxicology studies assessing chemical effects on tumor development in multiple organ sites.

e-mail: dunson@stat.duke.edu

WHAT NEXT IN GENE SET ANALYSIS?

Giovanni Parmigiani*, Johns Hopkins University
Luigi Marchionni, Johns Hopkins University
Sierra Li, Johns Hopkins University
Dongmei Liu, London School of Hygiene and Tropical Medicine
Leslie Cope, Johns Hopkins University

Gene set analysis considers whether genes that form a set from a specific biological standpoint, also behave similarly in a high throughput genomic experiment. This simple cross referencing is very powerful and creative definition of sets has allowed combined analysis and interpretation of very disparate sources of knowledge. In this presentation, I will provide a brief review of concepts, our of ongoing research on models for gene set analysis, and of remaining challenges. I will also present in some more detail an R tool called funcBox that facilitates “comparative gene set analysis” by allowing users to simultaneously analyze and visualize multiple gene set analyses performed on related comparisons or experiments.

e-mail: gp@jhu.edu

ESTIMATING VARIABLE STRUCTURE AND DEPENDENCE IN MULTI-TASK LEARNING VIA GRADIENTS

Sayan Mukherjee*, Duke University

We consider the problem of learning gradients in the supervised setting where there are multiple, related tasks. Gradients provide a natural interpretation to the geometric structure of data, and can assist in problems requiring variable selection and dimension reduction. By extending this idea to the multi-task learning (MTL) environment, we present methods for simultaneously
learning structure within each task, and the shared structure across all tasks. Our methods are placed within the framework of Tikhonov regularization, providing (a) robustness to high-dimensional data, and (b) a mechanism for incorporating a priori knowledge of task (dis)similarity. We provide an implementation for multi-task gradient learning for classification and regression, and demonstrate the utility of our algorithms on simulated and real data. In particular modeling of tumor progression.

email: sayan@stat.duke.edu

100. DYNAMIC TREATMENT REGIMES: PRACTICE AND THEORY

A TWO-STAGE SELECTION TRIAL WITH INTRAPATIENT SEQUENTIAL RANDOMIZATION IN PROSTATE CANCER

Randall E. Millikan*, M. D. Anderson Cancer Center
Peter F. Thall, M. D. Anderson Cancer Center
Sijin Wen, M. D. Anderson Cancer Center

Physicians typically switch therapies unless clinically relevant thresholds of response are observed, and treatments that produce high quality responses and that are active in the salvage setting are generally felt to be promising. With the goal of efficiently selecting promising regimens for more advanced trials, we applied a novel trial design that takes explicit account of first-line and salvage activity. Patients were initially randomized equally among 4 regimens. At each 8-week evaluation, patients could continue with the same treatment or be randomized to an alternative (from the remaining 3), in keeping with the attainment of clinically relevant thresholds of response at 8 weeks and at 16 weeks. Overall success (i.e. meeting the 16 week response threshold) by means of initial treatment assignment was observed in 35 patients, with an addtional 9 by means of salvage therapy. We found evidence that some patients respond to particular therapies, and advanced some hypotheses for definitive comparative trials. The dataset is now mature, and predictions based on response criteria were reinforced by survival data. A statistical model formalizing intuitive clinical concepts fit the data well and allowed additional hypothesis generation from the trial data.

e-mail: rmillika@mdanderson.org

TWO-STAGE TREATMENT STRATEGIES BASED ON SEQUENTIAL FAILURE TIMES

Peter F. Thall*, M.D. Anderson Cancer Center
Leiko H. Wooten, M.D. Anderson Cancer Center
Nizar Tannir, M.D. Anderson Cancer Center
Randall E. Millikan, M.D. Anderson Cancer Center
Christopher Logothetis, M.D. Anderson Cancer Center

For many diseases, therapy involves multiple stages, with treatment in each stage chosen adaptively based on the patient's current disease status and history of previous treatments and outcomes. Physicians routinely use such multi-stage treatment strategies, also called dynamic treatment regimes or treatment policies. In this talk, I will present a Bayesian framework for a clinical trial comparing several two-stage strategies based on the time to overall failure, defined as either second disease worsening or discontinuation of therapy. The design was motivated by a trial of two-stage strategies for treating advanced kidney cancer. Each patient is randomized among a set of treatments at enrollment, and if disease worsening occurs the patient is then re-randomized among a set of treatments excluding the treatment received initially. The goal is to select the two-stage strategy giving the largest mean overall failure time. A parametric model is formulated to account for non-constant failure time hazards, regression of the second failure time on the patient's first worsening time, and the complications that the failure time in either stage may be interval censored and there may be a delay between first and second stage of therapy. A simulation study in the context of the kidney cancer trial is presented.

e-mail: rex@mdanderson.org

SCREENING EXPERIMENTS FOR DYNAMIC TREATMENT REGIMES

Susan Murphy*, University of Michigan
Derek Bingham, Simon Fraser University

Dynamic treatment regimes individualize the treatment level and type via decision rules that specify when and how the intensity or type of treatment should change; the decision rules input outcomes collected during treatment and output recommended treatment alterations. These regimes are of great interest to clinical scientists as they operationalize the adaptive decision making inherent in clinical practice. However in this setting dynamic treatment regimes tend to be multi-component treatments and include the usual components (medications, counseling, adjunctive therapies, mode of delivery) in addition to components involving the timing of when to switch or augment treatment and which treatment to attempt first. Most of the treatment components will have already been evaluated for efficacy and safety. Here question centers on how best to combine the treatment components. Screening experiments represent a natural experimental approach that can be used to reduce the number of components, pinpoint which components may interact and which components and their interactions require further investigation. We discuss issues in defining effects and in conducting an analysis that not only permits determination of the aliasing but also produces appropriate causal inference.

e-mail: samurphy@umich.edu

OPTIMAL TREATMENT AND TESTING STRATEGIES WITH POSSIBLY NON-IGNORABLE OBSERVATION PROCESSES

Andrea Rotnitzky*, Di Tella University-Argentina and Harvard School of Public Health
James Robins, Harvard School of Public Health
Liliana Orellana, Facultad de Ciencias Exactas, Universidad de Buenos Aires
Miguel Hernan, Harvard School of Public Health

We review recent developments in the estimation of an optimal treatment strategy or regime from longitudinal data collected in an observational study. We propose novel methods for using the data obtained from an observational database in one health care system to determine the optimal treatment regime for biologically similar subjects in a second health care system when, for cultural, logistical, and financial reasons, the two health care systems differ (and will continue to differ) in the frequency of, and reasons for, both laboratory tests and physician visits. Finally, we propose a novel method for estimating the optimal timing of expensive and/or painful diagnostic or prognostic tests. Diagnostic or prognostic tests are only useful in so far as they help a physician to determine the optimal dosing strategy, by providing information on both the current health state and the prognosis of a patient because, in contrast to drug therapies, these tests have no direct causal effect on disease progression. Our new method explicitly incorporates this no direct effect restriction.

e-mail: arotnitzky@utdt.edu

101. TEXT DATA MINING

GO-DRIVEN LITERATURE-BASED DISCOVERY USING SEMANTIC ANALYSIS

Anthony E. Zukas*, Science Applications International Corporation
Jeffrey L. Solka, George Mason University
Jennifer W. Weller, University of North Carolina-Charlotte

This talk will discuss a new methodology to identify gene ontology (GO)-gene relationships. The methodology combines the previous work of Khatri et al., 2005 on the application of latent semantic indexing (LSI) to the gene-GO identification problem with the literature-based discovery work of Stegmann and Grohmann, 2003. The talk will illustrate how one can use our newly developed refinement on these approaches to help identify new gene-GO relationships.

e-mail: jeffrey.solka@navy.mil

TEXT ANALYSIS WITH ITERATIVE DENOISING

Kendall Giles*, Virginia Commonwealth University
David Marchette, Naval Surface Warfare Center
Carey Priebe, Johns Hopkins University

The performance of a classifier, given the real-world constraints of finite computational resources and limited training observations, is especially of concern in the case of unsupervised statistical learning. Iterative Denoising is a tree-based unsupervised statistical learning framework, and in this paper we investigate the empirical performance of one non-parametric implementation of this methodology. In order to do so, we propose an unsupervised metric for use as one stopping criteria for tree growth, and characterize the resulting Iterative Denoising trees from simulated and empirical text data using two indicators of a good model, accuracy and simplicity. The purpose is to gain a better understanding of how the bias-variance tradeoff can be realized in statistical learning using the Iterative Denoising
framework, especially when analyzing text.

e-mail: kendallgiles@gmail.com

TRACKING TRENDS IN HEALTH ARTICLES

Elizabeth L. Hohman*, Naval Surface Warfare Center

We use a corpus of health news articles collected over two years and attempt to detect changes and trends in the news stories. In order to treat the articles in time, we develop a generalization to the commonly used vector space model for text representation in the environment of a changing corpus of documents. This is achieved by using an exponential window for computing the document frequency of words and by managing a changing lexicon. A graph model is developed in order to provide a reduced representation of the documents. The vertices in the graph represent document topics and evolve in time. Methods are presented for visualizing the data and the document topics.

e-mail: elizabeth.hohman@navy.mil

102. NONPARAMETRIC BAYES: THE PRACTICAL USE FOR GENOMIC DATA

USING GENE ANNOTATION AS PRIOR INFORMATION IN BAYESIAN NONPARAMETRIC MODELS

David B. Dahl*, Texas A&M University

Integration of data from several sources and technologies is a burgeoning field in bioinformatics. Annotation information, such as Gene Ontology, has been used to validate clustering results of microarray gene expression data. We argue that annotation information should be an integral part of the estimation procedure. Indeed, Pan (2006) proposed a finite mixture model for microarray data that using gene function as prior information. This talk shows how the usual Dirichlet process prior can be extended to permit unequal prior probabilities that two genes are clustered based on annotation information. We show how the resulting Bayesian nonparametric model incorporates both gene expression data and annotation information to simultaneously infer clustering and detect of differential expression.

e-mail: dahl@stat.tamu.edu

NONPARAMETRIC GRAPHICAL MODELS WITH APPLICATIONS TO MICROARRAY EXPERIMENTS

Abel Rodriguez*, University of California, Santa Cruz

Nonparametric and clustering procedures have been extensively used for the analysis of microarray experiments. Depending of the context, their goal is to partition genes and/or experimental conditions in groups with similar expression profiles. Graphical models have also been repeatedly used in this context, but the goal in this case is typically to understand the dependence between genes in order to identify pathways. This work aims at combining both ideas by developing infinite mixtures of graphical models in the context of Dirichlet process models. Although the idea is relatively straightforward, computational issues arise in large dimensional problems like microarray analysis. We discuss these and show illustrations in simulated and real data sets.

e-mail: abel@ams.ucsc.edu

A GAUSSIAN PROCESS MODELING APPROACH TO SURVIVAL DATA WITH BAYESIAN VARIABLE SELECTION

Naijun Sha*, The University of Texas at El Paso
Marina Vannucci, Rice University

In this paper we propose a Gaussian process modeling approach to survival data that facilitates variable selection. Our specific interest is in survival analysis with high-dimensional data, such as those encountered in high-throughput genomic studies. We employ Bayesian variable selection methods to select predictors among the prohibitively vast number of variables. Inference is done via MCMC techniques. Our method simultaneously leads to the estimate of the survival function as well as the identification of the factors that affect the survival outcome. We describe strategies for posterior inference and explore their performance on simulated and real data. We also discuss possible extensions to additive-multiplicative survival models.

e-mail: nsha@utep.edu

INTEGRATIVE DIRICHLET PROCESS MIXTURES FOR CONSTRUCTING TRANSCRIPTIONAL MODULES

Mario Medvedovic*, University of Cincinnati
Xiandgong Liu, University of Cincinnati
Siva Sivaganesan, University of Cincinnati

Transcriptional modules (TM) consist of groups of co-regulated genes and transcription factors (TF) regulating their expression. Two high-throughput (HT) experimental technologies, gene expression microarrays and Chromatin Immuno-Precipitation on Chip (ChIP-chip), produce data informative about expression regulatory mechanism on a genome scale. The optimal method for identifying and characterizing TMs based on such data is an important open problem in computational biomedicine. We developed and validated a novel Dirichlet process-based statistical model identifying TMs by jointly analyzing gene expression and ChIP-chip binding data. The model uses nested Dirichlet processes to cluster genes based on the expression and binding data and to assign regulators to groups of co-regulated genes. Parameters of the model are estimated using the Gibbs sampler and inference is based on marginal posterior pairwise probabilities of co-expression and posterior binding probabilities. We demonstrate an improved functional coherence of the TMs produced by the new method when compared to either analyzing expression or ChIP-chip data separately or to alternative approaches for joint analysis. The software implementing the computational algorithm is incorporated is freely available at http://eh3.uc.edu/gimm.

e-mail: medvedm@email.uc.edu

103. METHODS IN CAUSAL INFERENCE

EVALUATING LONGITUDINAL TREATMENTS USING REGRESSION MODELS ON PROPENSITY SCORES

Achy-Brou CE Aristide*, Johns Hopkins University
Frangakis Constantine, Johns Hopkins University
Griswold Michael, Johns Hopkins University

We characterize admissible estimators for comparing longitudinal treatments using regression on longitudinal propensity scores. Assuming knowledge of the variables used in the treatment assignment, such estimators are important for reducing the large dimension of covariates. We show that if the regression models on the longitudinal propensity scores are correct, then our estimators are superior to estimators based on weights. If our models are incorrect, the misspecification can be limited through easier model checking. Thus, our estimators can also be superior when used in place of the regression on the full covariates. We use our methods to compare longitudinal treatments for type 2 diabetes mellitus.

e-mail: aachybro@jhsph.edu

STRUCTURAL PRINCIPAL EFFECTS MODELS FOR RANDOMIZED TRIALS WITH CONTINUOUS MEASURES OF COMPLIANCE

Yan Ma*, University of Rochester
Jason Roy, Geisinger Center for Health Research

In many clinical trials, compliance with assigned treatment could be measured on a continuous scale (e.g., the proportion of assigned treatment actually taken). In general, inference about principal causal effects (Frangakis and Rubin, 2002) can be challenging, particularly when there are two active treatments; the problem is exacerbated for continuous measures of compliance. We address this issue by first proposing a structural model for the principal effects. We then specify compliance models conditional on covariates within each arm of the study. These marginal models are identifiable. The joint distribution of the observed and counterfactual compliance variables is assumed to follow a Gaussian copula model, which links the two marginal models and includes a dependence parameter that cannot be identified. This dependence parameter can be varied as part of a sensitivity analysis. We illustrate the methodology with an analysis of data from a smoking cessation trial. As part of the analysis, we estimate causal effects at particular levels of the compliance variables and within subpopulations that have similar compliance behavior.

e-mail: yan_ma@urmc.rochester.edu

CAUSAL INFERENCE IN MEDIATOR ANALYSIS: OLD PROBLEMS AND NEW SOLUTIONS

Andreas G. Klein*, University of Western Ontario, Canada

In this paper, we discuss the difficulties related to the causal interpretation of mediator models. The problem of causal inference arises from the fact that the mediating variable plays the double role of an independent and a dependent variable and as such cannot be randomized. We adopt Holland's (1988) critique on the causal interpretation of conventional mediator models. As the main result, a new probability weighting technique is proposed that provides a novel solution to the problem and includes an estimation formula for the direct and indirect causal effects. The approach is illustrated by an example using depression data.

e-mail: aklein25@uwo.ca

COMPARE MEDIAN TREATMENT DIFFERENCE IN CASUAL INFERENCE

Jing Qin*, National Institute of Allergy and Infectious Diseases

Comparing mean treatment effect is one of the most popular methods in casual inference. In this paper we compare median treatment difference in observational studies when the observed data are skewed or there are some outliers. Empirical likelihood method is used to combine baseline covariate information effectively. Economic and medical data are used for illustration. Applications in missing data problem and survey sampling will be discussed briefly.

e-mail: jingqin@niaid.nih.gov

CONFOUNDING IN OBSERVATIONAL STUDIES COMPARING PROPENSITY SCORE AND TRADITIONAL REGRESSION ANALYSES

Megan E. Price*, Emory University
Vicki Hertzberg, Emory University
Michael Frankel, Emory University

In non-randomized studies, treatment groups may not be comparable due to potential confounders. Traditional regression analysis adjusts for this by including confounders in the final model. Propensity score analyses adjust for this by estimating the probability of treatment given a set of covariates, then adjusting for this probability in the final model. The effect of race on length of hospital stay after stroke was estimated using traditional and propensity score analyses. Traditional linear regression analyses included all potential confounders identified by propensity score analyses (hypertension, hyperlipidemia, coronary artery disease, and afibrillation). Simulations were conducted to systematically alter the balance of these covariates between black and white patients, then average length of stay was estimated for whites vs. blacks using both methods. Including propensity score as covariate resulted in statistically significant differences between methods in 12%

of simulations. When regression models were stratified by propensity score, the weighted average effect across propensity scores differed significantly from traditionally estimated effect size in approximately 30% of simulations. Overall, propensity score analyses may differ significantly from traditional linear regression controlling for potential confounders.

e-mail: meprice@sph.emory.edu

OPTIMAL PROPENSITY SCORE STRATIFICATION

Jessica A. Myers*, Johns Hopkins Bloomberg School of Public Health
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health

Stratifying on propensity score in observational studies of treatment is a common technique used to control for bias in treatment assignment; however, there have been few studies of the relative efficiency of the various ways of forming those strata. The standard method is to use the quintiles of propensity score to create subclasses, but this choice is not based on any measure of performance either observed or theoretical. In this paper, we investigate the optimal subclassification of propensity scores for estimating treatment effect with respect to mean squared error (MSE) of the estimate. We consider the optimal formation of subclasses within formation schemes that require either equal frequency of observations within each subclass or equal variance of the effect estimate within each subclass. Under these restrictions, choosing the partition is reduced to choosing the number of subclasses. We also consider an overalll optimal partition that produces an effect estimate with minimum MSE among all partitions considered. To create this stratification, the investigator must choose both the number of subclasses and their placement. Finally, we present a stratified propensity score analysis of data concerning insurance plan choice and its relation to satisfaction with asthma care.

e-mail: jamyers@jhsph.edu

A MIXTURE MODEL APPROACH TO ASSESSING THE PERFORMANCE OF INSTRUMENT VARIABLE ANALYSIS IN MEASURING THE EFFECT OF HORMONE THERAPY ON PROSTATE CANCER SURVIVAL

Dirk F. Moore*, University of Medicine and Dentistry, New Jersey-School of Public Health
Grace Lu-Yao, Cancer Institute of New Jersey

Instrument variable analysis is a technique for analyzing observational data that allows one to assess the impact of an intervention on an outcome while controlling for unmeasured covariates. The technique works if the population under study may be subdivided into clusters that vary in the proportions of subjects given the intervention, and if the subject's outcome depends only on the intervention, and not on proportion of individuals in that subject's cluster that receives the intervention. We use a mixture model to study the impact of cluster size on estimation bias in the presence of an unmeasured covariate. We apply the method to a SEER/Medicare population study, where we use IVA to estimate the survival of prostate cancer patients who receive or don't receive hormone therapy.

e-mail: mooredf@umdnj.edu

104. BIOASSAY AND CANCER APPLICATIONS

STOCHASTIC AND STATE SPACE MODELS OF HUMAN EYE CANCER

W. Y. Tan, University of Memphis
H. Zhou*, Arkansas State University

In 1975, Knudson had proposed a two-stage model for human eye cancer. The specific assumption is that when both copies of the RB gene have been mutated or deleted, then the stem cell grows instantaneously into a cancer tumor. While the discovery by Knudson has been confirmed by molecular biology and genetics (Cavenee et al. 1985; Tan 1991, Chapter 3, Weinberg, 2007), there are some basic questions that remain to be answered: (1) Is the two-stage model really the true model for human eye cancer? Recent studies by molecular biologists have indicated that the human eye cancer are best described by a three-stage model with the last stage involving apoptosis (DiCiommo et al. 2000). (2) If the RB gene can be inherited from parents to children, then what is the frequency of the mutated gene in the population? To answer these questions, in this paper we will proceed to develop a more realistic model for human eye cancer incorporating the recent biological findings by DiCiommo et al. (2000). We will introduce the state space model and develop a generalized Bayesian approach to estimate the unknown genetic parameters and to predict the state variables.

e-mail: zhou.hong@csm.astate.edu

IMPROVING BIOLOGICAL PLAUSIBILITY OF CLASSIFICATION TREES FOR CANCER RESEARCH

Douglas Landsittel*, Duquesne University
Megan McLaughlin, Independent Consultant
Anna Lokshin, University of Pittsburgh School of Medicine and University of Pittsburgh Cancer Institute

Classification trees have proven to be a useful discrimination technique with advantages such as automated variable selection and an easily interpretable decision rule. Recent advances in computational methods, such as boosting and bagging, have further improved classification accuracy and model stability. However, these methods result in a far more complex sequence of trees which may not retain any intuitive interpretation. Further, underlying biological knowledge is not utilized in fitting the tree(s); therefore, we may end up with a statistical rule without any biological plausibility. For instance, CA-125 is an antigen currently believed to be the best predictor of ovarian cancer, but, in our previous work, we have found scenarios where 1) the selected trees do not include CA-125 and 2) when CA-125 is included, the splits are at baseline levels not likely to be associated with disease. For this study, we aim to improve the biological plausibility of classification trees via modifying the associated algorithm. Specifically, we propose to force in variables of known clinical significance and impose limitations on ranges of data for the variable splits. Simulation studies will be conducted to assess model stability and generalization error of these methods.

e-mail: landsitteld@duq.edu

ON THE ROBUSTNESS OF THE Ploy-k TEST FOR LIFE TIME ANIMAL STUDIES

Mulugeta G. Gebregziabher*, Medical University of South Carolina
David Hoel, Medical University of South Carolina

The statistical analysis of cancer bioassay data has traditionally depended on the pathological determination of the experimental animal's cause of death. Recently the poly-k statistical test has provided a method of statistical analysis of animal bioassay data without the need for cause of death information. The test has been shown to have good statistical properties in the typical two-year cancer bioassay. However, for chronic life-time animal studies it has not been evaluated but is being applied in several studies. We observed in one recent life-time study of the gasoline additive MTBE that the poly-k test is not statistically robust. In this work, we assess the robustness of the poly-k method on simulated life-time bioassay data.

e-mail: gebregz@musc.edu

INTERVAL APPROACH TO ASSESSING ANTITUMOR ACTIVITY FOR TUMOR XENOGRAFT STUDIES

Jianrong Wu*, St Jude Children's Research Hospital

In tumor xenograft studies, the tumor growth inhibition ratio (T/C) is a commonly used measurement to assessing antitumor activity. Unfortunately, this measurement can discard useful data and result in a high false-negative rate. Furthermore, the degree of antitumor activity based on T/C ratio is assessed on the basis of an arbitrary cutoff point. To overcome these drawbacks, we propose an adjusted area-under-the-curve (AUC) ratio. Three nonparametric intervals of the adjusted AUC ratio are also proposed to assessing the significance of the antitumor activity of the treatment. The proposed method is then applied to a real tumor xenograft study.

e-mail: jianrong.wu@stjude.org

COMPARISON OF PROPERTIES OF TESTS FOR ASSESSING TUMOR CLONALITY

Irina Ostrovnaya* Memorial Sloan-Kettering Cancer Center
Venkatraman Seshan, Columbia University
Colin Begg, Memorial Sloan-Kettering Cancer Center

Cancer patients often develop second malignancies that may be either metastatic spread of a previous cancer or a new primary cancer. This distinction is of considerable clinical significance and may lead to radically different treatments. Sometimes the diagnosis cannot be made easily on the basis of comparison of the gross pathologic specimens. Recent advances in molecular profiling offer the possibility of comparison of the patterns of somatic mutations in the tumors at candidate genetic loci to see if the patterns are sufficiently similar to indicate a clonal (metastatic) origin. We have developed a likelihood ratio test that can be used to compare the mutational profiles. This test requires pre-specification of the marginal mutation probabilities at each locus, parameters for which some information will typically be available in the literature. This test was developed to circumvent possible validity problems in the Concordant Mutations Test, published in a recent article by Begg et al. (2007), due to simplifying assumptions about the nature of and expected frequencies of the somatic mutations. In simulations this test is shown to be valid, but to be considerably less efficient than the Concordant Mutations Test for sample sizes (numbers of informative loci) typical of this problem. It is shown that the efficiency can be improved by restricting the parameter space to a pre-specified range.

e-mail: ostrovnaya@mskcc.org

DETECTING GROUP DIFFERENCES WITH RIGHT-CENSORED COUNTS FROM SERIAL DILUTION ASSAYS

Tim Bancroft*, Iowa State University
Dan Nettleton, Iowa State University

This talk proposes a method to estimate the amount of bacteria present in a sample based on complete or partially censored observations on a series of diluted subsamples. These serial dilution assays are widely employed for quantifying bacteria concentration whenever concentrations in the original undiluted samples may be too high for accurate measurement. Our method, which is an extension of Fisher's Most Probable Number model (1922), requires the counts of bacterial colonies and/or a lower bound on the number of colonies which form on growth medium at each dilution level rather than just an observation on the presence/absence of the bacteria. We focus on the problem of testing for group differences when serial dilution assays provide the data for each experimental unit. Although our method requires a more rigorous data acquisition, it is shown through simulation to have higher power than Fisher's method and is also superior to another popular, more ad hoc procedure that uses only a subset of the serial dilution data.

e-mail: timbancroft2000@yahoo.com

AN OPTIMAL DILUTION EXPERIMENT DESIGN FOR SAMPLE DNA CONCENTRATION ESTIMATION

Ming Li*, Vanderbilt University Medical Center
Robbert Slebos, Vanderbilt University Medical Center
Yu Shyr, Vanderbilt University Medical Center

Mutations in microsatellite markers provide direct assessment of susceptibility to DNA damage. Because they reflect individual exposure and differences in carcinogen metabolism, studies of microsatellite mutations may provide more accurate prediction of cancer risk. When comparing different treatment groups, since the samples may have different DNA concentrations, it is appropriate to compare the mutation rates but not the counts of mutations. For each sample, we need to know the number of mutant alleles as well as the total allele count. The former can be recorded accurately through PCR amplification followed by dilution. However, to obtain the total allele count, or equivalently, the allele density, the current estimation scheme of using a series of dilution experiments is neither efficient nor accurate. To improve the estimation accuracy and efficiency, we propose a new dilution and estimation strategy, in which we determine the level of dilution so that the corresponding estimator of allele density has the minimum variance. The dilution strategy is modeled by a binomial experiment with the success probability derived from a Poisson distribution. Both mathematical derivations and simulations support that our proposed dilution strategy ensures an optimal estimator on the total number of alleles with minimal variance.

e-mail: ming.li@vanderbilt.edu

105. MISSING DATA AND IMPUTATION

ESTIMATING SPATIAL INTENSITY FROM LOCATIONS COARSENED BY INCOMPLETE GEOCODING

Dale L. Zimmerman*, University of Iowa

The estimation of spatial intensity is an important problem in spatial epidemiologic studies. A standard data assimilation component of these studies is the assignment of a geocode, i.e. point-level spatial coordinates, to the address of each subject in the study population. Unfortunately, when geocoding is performed by the standard automated batch method of street-segment matching to a georeferenced road file and subsequent interpolation, it is rarely completely successful. Typically, 10-30% of the addresses in the study population, and even higher percentages in particular subgroups, fail to geocode, potentially leading to a selection bias, called geographic bias, and an inefficient analysis. However, there is almost always some geographic information coarser than a point (e.g. a zip code) observed for the addresses that fail to geocode. In this presentation, I develop methodology for kernel-based spatial intensity estimation from coarsened geocoded data. Substantial improvements in intensity estimation quality relative to analyses using only the observations that geocode are demonstrated via simulation and an example from a rural health study in Iowa.

e-mail: dale-zimmerman@uiowa.edu

IMPUTATION MODEL ASSESSMENT USING POSTERIOR PREDICTIVE CHECKING

Yulei He*, Harvard Medical School
Alan M. Zaslavsky, Harvard Medical School

Multiple imputation is often used to facilitate analyses in studies with missing data. Following the strategy proposed by Gelman et al. (2005), we use posterior predictive checking for assessing the adequacy of imputation models. The focus of this method is to apply the analyses of interest to both the completed data with imputations and a simulated copy of the completed data under the imputation model and to compare the results across simulations. Posterior predictive P-values can be calculated for estimates under the targeted analyses, testing the fit of the imputation model. In addition, point estimates of the differences and the associated credible intervals show the effect of the lack of fit on the analysis results. We study the patterns of the assessment performances under different scenarios of uncongeniality between the true model, imputation model, and analysis model using a simulation study. An application to a cancer care dataset is presented.

e-mail: he@hcp.med.harvard.edu

SEQUENTIAL SEMI AND NONPARAMETRIC REGRESSION MULTIPLE IMPUTATIONS

Irina Bondarenko*, University of Michigan
Trivellore Raghunathan, University of Michigan

Multiple imputation is a general purpose method for analyzing data with missing values. Under this approach the missing set of values is replaced by several plausible sets of missing values to yield completed data sets. It is fairly well established that the imputations should be draws from a predictive distribution of the missing values and should condition on as many covariates as possible. A sequential regression imputation method uses a Gibbs sampling style iterative process of drawing values from a predictive distribution corresponding to a sequence of conditional regression models to impute the missing values in any given variable with all other variables as predictors. The conditional regression models are usually parametric. In practice, however, many variables have distribution that very difficult to classify or transform to satisfy standard parametric distribution assumptions. We develop and evaluate a modification of this method. We construct propensity score for missing the given variable and the predicted value of that variable. We stratify the sample based on these two scores and then within each stratum, we use approximate Bayesian Bootstrap or Tukey's GH distribution to impute the missing values conditional on the observed values. We illustrate proposed method using actual and simulated data sets.

e-mail: ibond@umich.edu

IMPUTATION ADJUSTED FOR COVARIATE FOR NONRESPONDENTS WITH APPLICATIONS

Juan Li* Amgen Inc.
Shein-Chung Chow, Duke University
Amy Feng, Amgen Inc.
Jr-Rung Lin, Duke University
Eric Chi, Amgen Inc.

In clinical research, instruments consisting a number of items are often used to assess treatment effects, e.g., quality of life assessment. In many situations, instead of individual component, it is of interest to provide an assessment of the treatment effect in some overall measures, e.g., cumulative scores. In practice, these types of data often suffer from incompleteness. A common method is to simply ignore all the item non-respondents from the analysis. Although this method is statistically valid under the assumption of MCAR, it suffers from decreasing power / efficiency and from excluding too many evaluable patients from the analysis. In our study, we propose a regression imputation approach adjusted for covariates with item non-respondents in the instrument. The proposed method provides consistent estimators, which are asymptotically normal. A bootstrap procedure is also proposed to estimate the asymptotic variance of the derived estimators. A simulation study was conducted to study the finite sample performance of the derived estimators. It is also shown that the estimators based on the imputed data set are more efficient than the estimators based on the completers only. The proposed methodology was illustrated through a real example concerning the STAMINA-HFP Registry.

e-mail: juli@amgen.com

FRACTIONAL IMPUTATION FOR CATEGORICAL VARIABLES WITH MISSING VALUES

Michael D. Larsen*, Iowa State University

Imputation is used to fill in missing values so that analyses based on complete data methods can be completed. Random imputation methods can add imputation variance to the results. Not accounting for the fact that some records are imputed can lead to understatement of uncertainty in conclusions. Fractional imputation (FI) imputes multiple values for each missing value and assigns fractional weights to the replicates. Under some forms of nonresponse, FI should reduce imputation variance while allowing accurate estimation of uncertainty. FI methods for categorical variables with missing values aredescribed. The method is applied to data from a longitudinal study of families in Iowa.

e-mail: larsen@iastate.edu

MULTIPLE IMPUTATION METHODS FOR TREATMENT NONCOMPLIANCE AND MISSING DATA IN CLUSTERED ENCOURAGEMENT DESIGN STUDIES

Leslie L. Taylor*, University of Washington and Axio Research
Xiao-Hua Zhou, University of Washington

Well-designed randomized clinical trials are a powerful tool for investigating causal treatment effects, but in human trials there are oftentimes problems of noncompliance. This is particularly a problem in encouragement design studies, where encouragement to take the treatment, rather than the treatment itself, is randomized. We consider a 'clustered encouragement design', meaning that the randomization is at the level of the clusters (e.g. physicians), but the compliance with assignment is at the level of the units (e.g. patients) within clusters (Frangakis et al. 2002). Furthermore, there is a problem of outcome nonresponse, as is typical in most clinical trials. Frangakis et al. (2002) proposed a Bayesian methodology for causal inference in a clustered encouragement design setting. We extend their setting to one in which there is outcome nonresponse, and we propose an alternative approach to causal inference using multiple imputation methods. We illustrate our methods using data from a study exploring the role of computer-based care suggestions in managing patients with chronic heart failure.

e-mail: taylorl@u.washington.edu

MULTIPLE IMPUTATION OF ORDINAL AND COUNT OUTCOMES IN A MULTIPLE SCLEROSIS CLINICAL TRIAL USING DATA AT DROPOUT

Peter B. Imrey*, Cleveland Clinic-Lerner College of Medicine at Case Western Reserve University
John Barnard, Cleveland Clinic-Lerner College of Medicine at Case Western Reserve University
Matthew Karafa, Cleveland Clinic Foundation

A survey of clinical trial reports in premier medical journals in 2001found 'almost no use of modern missing data methods' (Wood, White, and Thompson, 2004). Further development of accessible software has since eased application of analyses using multiple imputation (MI), pattern mixture models, and selection models when data are missing at random (MAR), as well as facilitated some approaches to inference in the presence of nonrandom dropout (MNAR). However, the situation often remains complex when variables are neither normal nor dichotomous. Moreover, the clinical trial literature remains dominated by traditional but inherently biased analyses using only complete cases, or mean or last-observation-carried-forward single imputation. We illustrate MI by predictive mean matching in a four-arm trial of therapies for relapsing-remitting multiple sclerosis. Multiple imputations were generated for proportional odds analysis of the ordinal primary outcome, (categorized) new and/or enlarging T2-hyperintense MRI lesions, and for negative binomial regression analysis of relapse counts. Imputation models merged available data at dropout with imputed data between dropout and target follow-up time, obtained by matching using complete case models with predictors from baseline data and interim data preceding the imputed patient's dropout time.

e-mail: imreyp@ccf.org

106. LONGITUDINAL MODELS-CONTINUOUS

THE MIXTURE OF NONLINEAR MODELS FOR GASTRIC EMPTYING STUDIES

Inyoung Kim*, Virginia Tech University
Noah D. Cohen, College of Veterinary Medicine, Texas A&M University
Allen Roussell, College of Veterinary Medicine, Texas A&M University
Naisyin Wang, Texas A&M University

Studying the impact of medications, procedures, and the composition or size of diets on gastric emptying is of great interested in equine medicine. The standard method for assessing gastric emptying is by using scintigraphy and summarizing the non-linear emptying of the radioisotope. Currently, the most popular model for fitting gastric emptying data is based on a power exponential model. This model has the limitation of poorly describing localized intragastric events that can occur during emptying. Hence, we develop a new model for gastric emptying studies to improve population and individual inference using the mixture of a nonlinear mixed-effect model and a nonlinear model: one model captures the globally decreasing pattern with random coefficients and the other captures local 'bumping' (i.e., apparent filling rather than emptying) or rapid decay (emptying), without random coefficients. Two methods are developed to fit the mixture model: 1) the mixture of an Expectation Maximization algorithm and a global two-stage method; and, 2) the mixture of an Expectation Maximization algorithm and the Monte Carlo Expectation Maximization algorithm. We compare our methods using simulation, showing that the two methods appear approximately equally efficient. An example of gastric emptying data from equine medicine that motivated the work is used to demonstrate our model and approaches.

e-mail: inyoungk@vt.edu

COMPARISONS OF METHODS FOR HANDLING MISSING CONTINUOUS LONGITUDINAL OUTCOME WITH MIXED EFFECTS MODELS

Tulay Koru-Sengul*, McMaster University-Canada

Researchers are frequently faced with the problem of analyzing data with missing values. Missing values are practically unavoidable in long term longitudinal studies especially in medicine and incomplete data sets make the statistical analyses very difficult. Mixed effects models are widely used for modeling longitudinal continuous outcomes with missing values. In my talk, I will discuss the missing-data problem, implications of missing values for data analysis and interpretation specifically for mixed effects models. Implications of different methods of handling missing longitudinal continuous outcome on the parameter estimates will be studied for different missing data mechanisms. Complete case analysis, LOCF method, mixed effects model, and multiple imputation will be studied with simulation studies. A longitudinal randomized clinical trial on cancer will be used for illustration purposes.

e-mail: korut@mcmaster.ca

MODELING COVARIANCE STRUCTURE IN UNBALANCED LONGITUDINAL DATA

Min Chen*, Texas A&M University

Modeling covariance structure is important in efficient estimation for longitudinal data models. Pourahmadi (1999,2000) promoted to use modified Cholesky decomposition as an unconstrained reparameterization of the covariance matrix. The new parameters have transparent statistical interpretations and are easily modeled using covariates. However, this approach is not directly applicable when the longitudinal data are unbalanced, because a Cholesky factorization for observed data coherent across all subjects usually does not exist. We overcome the difficulty by treating the problem as a missing data problem and employing a generalized EM algorithm to compute the ML estimators. We illustrate our method by reanalyzing Kenward's (1987) cattle data and conducting a simulation study.

e-mail: mchen@stat.tamu.edu

THE GLM WITH A GENERALIZED AR(1) COVARIANCE STRUCTURE

Sean L. Simpson*, University of North Carolina at Chapel Hill
Lloyd J. Edwards, University of North Carolina at Chapel Hill
Keith E. Muller, University of Florida
Pranab K. Sen, University of North Carolina at Chapel Hill

Repeated measures designs are very common in biostatistical applications. The general linear model for repeated measures data is often employed when standard multivariate techniques do not apply. In many repeated measures experiments, the within-subject correlation tends to decrease exponentially in time or space. This correlation pattern is generally modeled with the AR(1) covariance structure. However, there are many situations in both longitudinal and imaging studies in which the within-subject correlations decay at a rate faster or slower than that imposed by the AR(1) structure. We propose a three-parameter covariance model for the general linear model with structured covariance that is a generalization of the continuous-time AR(1) structure. Through simulations we show that this new covariance model, termed the 'generalized autoregressive (GAR) covariance structure', is more appropriate for many types of data than the AR(1) model and other comparable models. We employ the GAR covariance model in the analysis of a real imaging data example.

e-mail: ssimpson@bios.unc.edu

NONPARAMETRIC ESTIMATION FOR CONDITIONAL DISTRIBUTION FUNCTIONS AND TIME-VARYING TRANSFORMATION MODELS WITH LONGITUDINAL DATA

Colin O. Wu, National Heart, Lung and Blood Institute
Xin Tian*, National Heart, Lung and Blood Institute
Jarvis Yu, National Heart, Lung and Blood Institute

Regression methods for longitudinal analysis have mostly focused on conditional mean based models. In many situations, the relevant scientific questions could be better studied by modeling the conditional distributions of the outcome variables as a function of time and other covariates. In this paper, we propose a time-varying nonparametric approach for modeling the conditional cumulative distribution functions (CDF) and the time-varying covariate effects, and develop a two-step method for estimating the conditional CDF's and the time-varying parameters. Applications and finite sample properties of our modeling and estimation procedures are demonstrated through a prospective cohort study of obesity and cardiovascular risk factors and a simulation study.

e-mail: tianx@nhlbi.nih.gov

INFERENCE FOR CENSORED QUANTILE REGRESSION MODELS IN LONGITUDINAL STUDIES

Huixia Judy Wang*, North Carolina State University
Mendel Fygenson, University of Southern California

We develop inference procedures for longitudinal data where some of the measurements are censored by fixed constants. We consider a semi-parametric quantile regression model that makes no distributional assumptions. Our research is motivated by the lack of proper inference procedures for data from biomedical studies where measurements are censored due to a fixed quantification limit. In such studies the focus is often on testing hypotheses about treatments equality. To this end, we propose a rank score test for large sample inference on a subset of the covariates. We demonstrate the importance of accounting for both censoring and intra-subject dependency and evaluate the performance of our proposed methodology in a simulation study. We then apply the proposed inference procedures to data from an AIDS related clinical trial. We conclude that our framework and proposed methodology is very valuable for differentiating the influences of predictors at different locations in the conditional distribution of a response variable.

e-mail: wang@stat.ncsu.edu

CHARACTERIZATION OF VARIABILITY IN LONGITUDINAL DATA USING THE SPECTRUM

Wei Yang*, University of Pennsylvania School of Medicine
Marshall Joffe, University of Pennsylvania School of Medicine
Steven Brunelli, University of Pennsylvania School of Medicine

In longitudinal data analysis, there has been an increasing interest in characterizing the variability of physiologic parameters, and to explore the association between variability and outcome (Joffe 2007; Elliott 2007). One challenge in defining the variability is to distinguish it from the long-term trend (e.g., gradual change) and the short-term oscillations due to measurement error (e.g. random noise). Spectrum analysis can be useful in this regard, and functions by decomposing variance into a series of frequencies: low-frequency components represent the gradual excursions; high-frequency components represent the short-term activities (including random noise). Therefore, by removing both the low- and high-frequency spectral components, we can characterize the variability independent of the long-term temporal trends and short-term variation caused by random noise. The specific frequency band in which the variability will be defined depends on the research of interest or the intrinsic frequency of the measurement. We use this approach to characterize blood pressure variability in hemodialysis patients and to explore its association with mortality.

e-mail: weiyang@mail.med.upenn.edu

107. BAYESIAN AND MULTI-LEVEL SURVIVAL ANALYSIS

HIERARCHICAL DYNAMIC TIME-TO-EVENT MODELS FOR POST-TREATMENT PREVENTIVE CARE DATA ON BREAST CANCER SURVIVORS

Freda W. Cooner*, U.S. Food and Drug Administration
Xinhua Yu, University of Minnesota
Sudipto Banerjee, University of Minnesota
Patricia L. Grambsch, University of Minnesota
A. Marshall McBean, University of Minnesota

This article considers modeling data arising in post-treatment preventive care settings, where cancer patients who have undergone disease-directed treatment discontinue seeking preventive care services. Clinicians and public health researchers are interested in explaining such behavioral patterns by modeling the time to receiving care while accounting for several patient and treatment attributes. A key feature of such data is that many patients would never return for screening,

a concept subtly different from censoring, where an individual does not return for screening in the given time frame of the study. Models distinguishing between these two concepts are known as cure models and are often preferred for data where a significant part of the population never experienced the endpoint. Building upon recent work on hierarchical cure models we propose modeling a sequence of latent events with a piecewise exponential distribution that remedies over-smoothing encountered in existing models. We investigate simultaneous regression on the cure fraction and the latent event distribution and derive a flexible class of semi-parametric cure models.

e-mail: freda.cooner@fda.hhs.gov

BAYESIAN SEMIPARAMETRIC MODELING USING MIXTURES FOR STRATIFIED SURVIVAL DATA

Bo Cai*, University of South Carolina
Renate Meyer, University of Auckland

Survival analysis based on mixture models have certain advantages over classical parametric approaches because mixture models provide a convenient and flexible semiparametric framework to model unknown distributions shapes. We describe a generalized mixture model based on B-splines for modeling monotone function such as the integrated baseline hazard function and covariate link in a proportional hazard model, which includes beta mixtures by Diaconis and Ylvisaker (1985, Bayesian Statistics, 2, 133-156) and triangular mixtures by Perron and Mengersen (2001, Biometrics, 57, 518-528). A Bayesian hierarchical semiparametric proportional hazard model is developed by using such mixtures for fitting stratified survival data. Data from a multicenter AIDS clinical trial are used for illustration and comparison of hierarchical proportional hazards regression models based on different mixtures.

e-mail: bocai@gwm.sc.edu

A TRANSFORMATION APPROACH FOR THE ANALYSIS OF INTERVAL-CENSORED FAILURE TIME DATA

Liang Zhu*, University of Missouri
Xingwei Tong, Beijing Normal University
Jianguo Sun, Universityof Missouri

This paper discusses the analysis of interval-censored failure time data, which has recently attracted a great amount of attention (Li and Pu, 2003; Sun, 2006; Tian and Cai, 2006; Zhang et al., 2005). Interval-censored data mean that the survival time of interest is observed only to belong to an interval and they occur in many fields including clinical trials, demographical studies, medical follow-up studies, public health studies and tumorgenicity experiments. A major difficulty with the analysis of interval-censored data is that one has to deal with a censoring mechanism that involves two related variables. For the inference, we present a transformation approach that transforms general interval-censored data into current status data, for which one only needs to deal with one censoring variable and the inference is thus much easier. We apply this general idea to regression analysis of interval-censored data using the additive hazards model and numerical studies indicate that the method performs well for practical situations. An illustrative example is provided.

e-mail: lzkn7@mizzou.edu

BAYESIAN THRESHOLD REGRESSION WITH RANDOM EFFECTS

Michael L. Pennell*, The Ohio State University
Mei-Ling Ting Lee, The Ohio State University

Threshold regression models are a conceptually appealing approach to relating covariates to time to a medical endpoint, such as death. A common approach is to model the health status of a subject by a Wiener diffusion process whose parameters are related to a set of covariates through generalized link functions. The medical endpoint is reached when the process reaches a threshold for the first time. In this presentation, we propose a Bayesian methodology which extends previous fixed effect models to include random effects that account for between subject heterogeneity in both the initial state and drift of the Wiener process. Posterior inference proceeds via an MCMC methodology which includes data augmentation steps to sample the final health status of censored observations. We will exemplify our method using mortality data of patients following operation for malignant melanoma.

e-mail: mpennell@cph.osu.edu

THE APPLICATION OF THE BAYESIAN DYNAMIC SURVIVAL MODEL

Jianghua He*, University of Kansas Medical Center
Daniel L. McGee, Florida State University
Xufeng Niu, Florida State University

Different estimation approaches has been proposed for the Bayesian Dynamic Survival Model, a time varying coefficient model from the Bayesian prospective. Unlike the other time-varying coefficients survival modeling approches such as the additive survival model or the time-varying coefficient Cox model, the Bayesian Dynamic Survival Model is not widely discussed or used. In this paper, we assess and compare the performances of these methods with a designed simulation study and propose a method of implementing of the Bayesian Dynamic Survival Model for medical research. The proposed approach is applied to the some examples from clinical and epidemiological research to investigate the consequences of ignoring time-varying effects.

e-mail: hejiang75@yahoo.com

LINEAR REGRESSION ANALYSIS OF SURVIVAL DATA FROM STRATIFIED CASE-COHORT STUDIES

Lan Kong*, University of Pittsburgh
Jianwen Cai, University of North Carolina at Chapel Hill

In a case-cohort design, the expensive covariates are assembled only for a randomly selected subcohort of the entire cohort, and any additional cases outside the subcohort. We developed a rank-based estimating equation approach for stratified case-cohort studies under a semiparametric linear regression model. We showed that the proposed estimators were consistent and asymptotically normally distributed. We studied the finite sample properties of case-cohort estimator and its relative efficiency to full cohort estimator via simulation studies. A real example from a study of cardiovascular disease was provided for illustration.

e-mail: lkong@pitt.edu

ON RISK REVERSAL DUE TO HETEROGENEITY

David Oakes*, University of Rochester

Unbiased random error usually has the effect of attenuating the magnitude and or strength of associations. However there are circumstances where unbiased random error can result in an (apparent) reversal of the direction of an effect. We construct some specific examples to illustrate this phenomenon.

e-mail: oakes@bst.rochester.edu

108. MULTIPLE TESTING IN GENOMICS

A SEQUENTIAL TESTING PROCEDURE FOR DETECTION OF ASSOCIATION BETWEEN DISEASE AND HAPLOTYPE BLOCKS: A SIMULATION STUDY

Andres Azuero*, University of Alabama at Birmingham
David T. Redden, University of Alabama at Birmingham

With the advent of the international HapMap project, there has been increased development of statistical methods for haplotype phasing using genotype data of multiple linked SNPs. Regardless of the haplotype inference algorithms, the issue of multiple testing of large numbers of haplotype blocks in the same group of subjects remains a major obstacle in Genetic Association Studies and Genome-Wide Association Scans. A Likelihood Ratio Sequential Test (LRST) provides an efficient way to sequentially test a simple null hypothesis versus a composite alternative hypothesis as sample size accumulates. We attempt to develop an LRST-based group-sequential procedure for testing association between disease and haplotype blocks, that combines the efficiency of the LRST with a flexible correction for multiplicity, in the context of Genetic Association Studies and Genome-Wide Association Scans. We use simulated datasets to study the behavior of the procedure in regards to false positive rate, power, and cost-efficiency.

e-mail: andreo@uab.edu

A PARAMETRIC PERMUTATION TEST FOR REGRESSION COEFFICIENTS IN LASSO REGULARIZED REGRESSION FOR HIGH DIMENSIONAL DATA

Michael C. Wu*, Harvard School of Public Health
Tianxi Cai, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

Variable selection in high dimensional data arises frequently in modern biomedical research. Regularized regression using the LASSO procedure (Tibshirani, 1996) is a popular method that has been found to be effective in selecting models with good predictive ability even when the dimensionality is high. However, in many biomedical studies, testing for associations between an outcome and predictors is the primary goal of interest. Standard methods for evaluating the significance of model parameters fail for the LASSO. We propose a LASSO based parametric permutation test procedure for testing the significance of regression coefficients in high dimensional data. Asymptotic theory is presented, and finite sample simulations show that our test has correct size and good power. We apply the proposed method to a study relating HIV drug resistance to viral codon mutations. Theoretical results are extended to generalized linear models.

e-mail: mwu@hsph.harvard.edu

EMPIRICAL BAYES MODELING WITH OPTIMAL DISCOVERY PROCEDURE FOR IMPROVED SIGNIFICANCE TESTING OF MICROARRAY DATA

Xiting Cao*, University of Minnesota
Baolin Wu, University of Minnesota

Empirical Bayes method has proven to be a very useful approach for simultaneous significance testing of large-scale microarray data. The intuitive idea is to borrow information across genes to improve the individual gene inference. The recently proposed optimal discovery procedure for multiple significance testing is another way to improve the overall testing power. In this paper, we propose to develop an improved optimal discovery procedure based on Empirical Bayes modeling. Through simulation studies and applications to public microarray data, we will illustrate the competitive performance of the proposed methods.

e-mail: baolin@umn.edu

A FLEXIBLE BAYESIAN FRAMEWORK FOR LARGE SCALE SIMULTANEOUS TESTING

Seungbong Han*, University of Wisconsin - Madison
Adin Cristian Andrei, University of Wisconsin - Madison
Kam Wah Tsui, University of Wisconsin - Madison

When drawing large scale simultaneous inference, such as in genomics and imaging problems, multiplicity adjustments should be made, since, otherwise, one would be faced with an inflated type I error. However, when interested in a particular hypothesis, it might be more advantageous to carry out inference based on test-specific evidence. We develop a joint probability model across the set of p-values, while incorporating a latent binary random variable indicating whether or not the respective null hypothesis is true. Under the null hypothesis, we assume that each p-value follows a uniform distribution, while under the alternative it has a general beta distribution, whose parameters are not subject to restrictions. Our methodology makes use of the raw p-values, without transforming them to p-values based on other test statistics, such as the z-statistic. Such transformations are sometimes based on assumptions that may be unreasonable. This is a very flexible approach, in that it does not impose restrictions on the distribution of the p-values under the alternative hypothesis. An MCMC method is proposed to carry out the posterior inference to detect the true alternative hypotheses. Method performance is assessed in simulated settings. Its practical usefulness is illustrated in two applications to a B-cell lymphoma study and a breast cancer example, where p-values based on Cox regression models are used.

e-mail: hanseung@stat.wisc.edu

TWO-STAGE GROUP SEQUENTIAL ROBUST TESTS IN FAMILY-BASED ASSOCIATION STUDIES: CONTROLLING TYPE I ERROR

Lihan K. Yan*, George Washington University; The EMMES Corporation
Gang Zheng, National Heart, Lung and Blood Institute
Zhaohai Li, George Washington University, National Institute of Child Health and Human Development

In family-based association studies, an optimal test statistic with asymptotic normal distribution is available when the underlying genetic model is known (e.g., recessive, additive/multiplicative or dominant). In practice, however, genetic models for many complex diseases are often unknown. Using a single test statistic optimal for one genetic model may

lose substantial power when the model is mis-specified. When a family of genetic models is scientifically plausible, the maximum of several tests, each optimal for a specific genetic model, is robust against the model mis-specification. This robust test is preferred over a single optimal test. Recently, cost-effective group sequential approaches have been introduced to genetic studies. The group sequential approach allows interim analyses and has been applied to many test statistics, but not to the maximum statistic. When the group sequential method is applied, Type I error should be controlled. We propose and compare several approaches of controlling Type I error rates when group sequential analysis is conducted with the maximum test for family-based association studies. For a two-stage group sequential robust procedure with a single interim analysis, two critical values for the maximum tests are provided based on a given alpha spending function to control the desired overall Type I error.

e-mail: lihanyan@yahoo.com

IMPROVED ESTIMATION OF FALSE DISCOVERY RATES BASED ON SUBSAMPLING WITH APPLICATIONS TO MICROARRAY DATA

Long Qu*, Iowa State University
Dan Nettleton, Iowa State University
Jack CM Dekkers, Iowa State University

False discovery rate control procedures used in microarray data analysis are often conservative primarily due to the bias in estimating the proportion of true null hypotheses (r). We develop an improved procedure for estimating r through novel use of data subsampling, with analogy to but extending the jackknife. Our procedure repeatedly deletes a random set of biological samples to produce many subsamples of various sample sizes. For each subsample, the same set of hypotheses is tested and the p-value density at 1 is estimated. We develop a flexible regression function to regress the p-value density at 1 over the sample size and then extrapolate it to infinity to get the estimate of r. This corresponds to the p-value density at 1 with an infinite sample size, which is exactly r in theory. The new estimator achieves smaller mean squared error (MSE) than the currently most widely used q-value smoother method by greatly reducing bias but mildly increasing variance in most simulation settings. In conclusion, our new procedure leads to improved statistical power and has smaller MSE compared to the q-value smoother
method. (Supported by USDA-NRI-2005-3560415618)

e-mail: longor@iastate.edu

DATA COMPLEXITY DEPENDENCE OF FALSE DISCOVERY RATE ESTIMATES IN DIFFERENTIAL GENE SELECTION PROCEDURES

Michael G. Schimek*, Medical University of Grazaustria
Tomas Pavlik, Masaryk University-Czech Republic

The false discovery rate FDR is most popular for global type I error control in studies of differential gene expression. Gene variance heterogeneity is usually tackled by ‘correcting’ constants in the estimation of the pooled variance. But there is also the influence of correlated measurements due to co-expression. Correlation effects can only be studied in simulation experiments. Here we demonstrate for the most popular SAM approach (Chu et al. (2005) Tech. Rep., Stanford Univ.) that the FDR estimates and thus the sets of selected genes much depend on the underlying data model. We specifically focus on a model due to Qiu, Xiao, Gordon and Yakovlev (2006, BMC Bioinformatics, 7:50) that stimulated a controversial discussion. Moreover we are studying alternative SAM procedures (Schwender, Krause and Ickstadt (2003) Tech. Rep. SFB 475, Univ. of Dortmund; Grant, Liu and Stoeckert (2005) Bioinformatics 21, 2684-90). In our simulation study a FDR-free empirical Bayes thresholding EBT procedure (Johnstone and Silverman (2004) Annal. Statist., 32, 1594-1649) was included as well. Our conclusion is that SAM procedures can cope better with correlation compared to EBT but fail for data of high structural complexity.

e-mail: michael.schimek@meduni-graz.at

109. ADVANCES IN THE DESIGN AND ANALYSIS OF RANDOMIZED CLINICAL TRIALS

RE-FORMULATING NON-INFERIORITY TRIALS AS SUPERIORITY TRIALS: THE CASE OF BINARY OUTCOMES

Valerie Durkalski*, Medical University of South Carolina
Vance Berger, National Cancer Institute

A continuously debated design feature of non-inferiority trials is the choice of the margin of non-inferiority. The non-inferiority setting presents the challenge of deciding on a balance between a suitable reduction in efficacy in return for a gain in other important treatment characteristics. The reduction in efficacy is pre-specified as the clinically relevant un-important difference between two treatments, also referred by the lower case delta symbol, d. The choice of delta has received great attention in the literature and continues to be a critical issue in the design of non-inferiority trials. It would be ideal to alleviate the dilemma by reverting to a traditional superiority trial design where a single p-value for both the most important endpoint (efficacy) and the most important finding (superiority). In some cases, this can be done even when superiority is not expected for efficacy. We discuss how this can be done using the information-preserving composite endpoint approach and consider binary outcome cases in which the combination of efficacy and treatment characteristics, but not efficacy itself, paints a clear picture that the novel treatment is superior to the active control.

e-mail: durkalsv@musc.edu

PLACEBO EFFECTS OR MEDICATION COMPLIANCE?

William Grant*, James Madison University

In randomized experiments, placebo effects occur when subjects’ expectations are responsible for part of the treatment effect. I examine one factor which may confound the measurement of placebo effects: medication compliance. This paper (1) provides a model of compliance effort as a choice variable for maximizing net utility and (2) conducts a test for compliance effects and placebo effects. For a sample of statin drug trials, regression analysis indicates that medication compliance effects are statistically significant but placebo effects are insignificant.

e-mail: grantwc@jmu.edu

ENROLLMENT PROJECTIONS: ESTIMATING WITH CONFIDENCE

Venita DePuy*, INC Research, Inc.

The enrollment of subjects is major component of the critical path for most clinical trials. Current practice to estimate enrollment timelines is typically no more than estimating a given number of people per month enrolled per site; for example, clinicians approximate that 5 patients per month per site are recruited in similar trials, so it is calculated that it will take 10 months for 4 sites to recruit 200 patients. If this timeline is not met, it will affect not only schedules, budgets, and submission deadlines for the current trial, but can also delay an entire program of work, from affecting the start-up of other studies to delay in time until market. One key drawback to this method of enrollment projection is that it does not account for variation in the timeline estimate. We explore enrollment predictions based on exponential simulation models, resulting in an estimated timeline based on the number of sites and different confidence intervals for that estimate. By using the upper confidence limit of that estimate, it is possible to improve trial planning by better estimating both the timeline for enrollment and the number of sites needed to meet enrollment projections within a given timeline.

e-mail: vdepuy@incresearch.com

SELECTION BIAS AND COVARIATE IMBALANCE IN RANDOMIZED CLINICAL TRIALS

Vance Berger*, National Cancer Institute

The quality of randomized trials is often evaluated for the purpose of synthesizing the results of several studies. It is common to use a check list for this purpose, with some check lists being more complete than others. Most ask about randomization itself, and some go further and ask how a trial was randomized, so as to verify that it actually was randomized, and not alternated. While these check lists do discriminate between truly randomized trials and pseudo-randomized trials, they do not discriminate further among the various ways to randomize. Yet it is a fact that some randomization methods are better than others. We propose a novel quantification of the quality of randomization based on a notion of degrees of freedom to measure the extent of randomization. The quality of randomization is then the adjusted degrees of freedom (i.e., the degrees of freedom divided by the sample size). This measure applies even to deterministic designs, which fit on the scale, but with very low scores (possibly zero). The adjusted degrees of freedom would be maximized by unrestricted randomization, for which a full 100% credit is assigned. Permuted blocks of size two involve true randomization of only half the patients (the first one in each block), so the score would be 50%. With larger block sizes and more complicated schemes one must distinguish between the design and the result. For example, with blocks of size four, the third allocation in a block may or may not be truly random depending on the first two allocations in that block, so we need two different measures. These will be discussed in the context of scoring a trial for quality.

e-mail: bergerv@mail.nih.gov

110. RECENT ADVANCES IN GRAPHS/GRAPHICAL MODELS FOR GENETIC NETWORK ANALYSIS

NETWORK STRUCTURE INFERENCE IN BIOLOGY

Wing H. Wong*, Stanford University

The reconstruction of signaling and regulatory networks from experimental data is a major challenge in current biology. In this talk we discuss the statistical issues in this problem. We believe that successful network structure inference will require an approach that integrates recent advances in Monte Carlo Bayesian computation, statistical learning, as well as novel hardware architecture.

e-mail: whwong@stanford.edu

THE MODAL ORIENTED STOCHASTIC SEARCH FOR GRAPHICAL MODELS

Adrian Dobra*, University of Washington

We present a novel stochastic search method for exploring regions of high posterior probability for Gaussian graphical models and hierarchical log-linear models. We illustrate the application of our method for inferring gene expression networks and SNP-SNP association networks.

e-mail: adobra@u.washington.edu

HIGH DIMENSIONAL GRAPHS INFERENCE BY SPARSE REGRESSION

Jie Peng*, University of California-Davis
Pei Wang, Fred Hutchinson Cancer Research Center,
Ji Zhu, University of Michigan

In this talk, we propose a joint sparse regression approach for the selection of non-zero entries in the inverse covariance matrix of a multivariate normal distribution under the setting of $p>n$. This method depends on the assumption of the overall sparsity of the concentration matrix. We study the properties of this approach by extensive simulation studies, as well as asymptotic analysis. We also apply this method on high dimensional array data for genetic network inference. We demonstrate that, compared to some existing methods, our method is particularly favorable when hubs (genes with many connections) exist.

e-mail: jie@wald.ucdavis.edu

111. NEW STATISTICAL METHODS FOR BIOMEDICAL IMAGING DATA

STATISTICAL ANALYSIS OF NEUROIMAGING DATA USING ADJUSTED EXPONETIALLY TILTED LIKELIHOOD

Hongtu Zhu*, University of North Carolina at Chapel Hill
Haibo Zhou, University of North Carolina at Chapel Hill
Jiahua Chen, University of British Columbia- Canada
Yimei Li, University of North Carolina at Chapel Hill
Martin Styner, University of North Carolina at Chapel Hill

We develop an adjusted exponentially tilted likelihood along with a nonparametric likelihood ratio statistic to test linear hypotheses of unknown parameters, such as the associations of brain measures (e.g., cortical and subcortical surfaces) with their potential genetic determinants. The exponentially tilted likelihood is a nonparametric method, and thus it avoids parametric assumptions in standard linear regression including that imaging data follow a Gaussian distribution. We also propose an novel adjustment to exponential tilting likelihood. This adjustment not only dramatically improves the finite performance of the exponentially tilted likelihood, but also it ensures all theoretical properties of the estimate based on the exponentially tilted likelihood.

Simulation studies show that our adjusted exponentially tilted likelihood ratio statistic performs as good as the t test when imaging data are symmetric distributed and it is better than the t test when imaging data are skewed distributed. We apply our methods to the detection of statistical significance of the difference in the m-rep model of hippocampus between schizophrenia patients and healthy subjects in a schizophreniza study.

e-mail: hzhu@bios.unc.edu

ANALYZING HIGH TEMPORAL RESOLUTION fMRI DATA

Martin A. Lindquist*, Columbia University

Understanding the neural basis of human brain function requires a detailed knowledge of both the spatial and temporal aspects of information processing. Functional magnetic resonance imaging (fMRI) provides the capability of visualizing changes in neuronal activity with high spatial resolution, but is lacking in temporal resolution. Most statistical techniques for analyzing fMRI data are based on studying oxygenation patterns that are far removed from the underlying event we wish to base our conclusions on (i.e. the neural activity). Therefore, one faces the statistically intractable task of sorting out possibly unknown confounding factors influencing the timing of these oxygenation patterns across different regions of the brain, in comparison to the actual ordering of their neuronal activity. Recently, new techniques have been introduced that can dramatically increase the temporal resolution of fMRI studies, and provide temporal activation profiles with significantly more detail than has previously been available in fMRI studies. Proper analysis of this high temporal resolution data necessitates the development of statistical methods for scoring observed activation responses that can be used for determining the absolute order of activation between different brain regions. We will discuss such techniques and compare them with standard metrics for determining temporal ordering in fMRI.

e-mail: martin@stat.columbia.edu

PHARMACOLOGIC IMAGING USING PRINCIPAL CURVES IN SINGLE PHOTON EMISSION COMPUTED TOMOGRAPHY

Brian S. Caffo*, Johns Hopkins University
Lijuan Deng, Boston Scientific Company
Ciprian Crainiceanu, Johns Hopkins University
Craig Hendrix, Johns Hopkins School of Medicine

In this talk we are concerned with functional imaging of the colon to assess the kinetics of a microbicide lubricant. The overarching goal is to understand the penetration of the lubricant after anal coitus. Such information is crucial for understanding the potential impact of the microbicide on viral transmission. The experiment was conducted by simulating coitus in a subject after injecting a radiolabeled lubricant. After coital simulation, the subject was imaged via Single photon emission computed tomography (SPECT), a non-invasive, in-vivo functional imaging technique. We use a highly modified version of the principal curve algorithm to construct a three dimensional curve through the colon images. The algorithm is developed on several difficult two dimensional images of familiar curves. The final curve fit the colon data is compared to experimental sigmoidoscope collection.

e-mail: bcaffo@jhsph.edu

A UNIFIED FRAMEWORK FOR GROUP INDEPENDENT COMPONENT ANALYSIS FOR MULTI-SUBJECT fMRI DATA

Ying Guo*, Rollins School of Public Health, Emory-University

Independent component analysis (ICA) is a popular tool for analyzing functional magnetic resonance imaging (fMRI) data. The extension of ICA for group inferences in fMRI is a challenging topic because the spatio-temporal structure underlying measured brain activity is not pre-specified in ICA and may vary across subjects. Current group ICA approaches assume different structures of the group spatio-temporal processes, and their estimation and inference procedures are tailored specifically for a particular model assumption. I propose a unified framework for fitting group ICA models with different underlying group structures. The proposed method could incorporate covariate information in the ICA decomposition of multi-subject data and hence allow for group comparison of brain activity between subject subpopulations. A global as well as a local statistical test are proposed to select an appropriate group structure for the whole brain and for a local functional network, respectively. Simulation studies and application to an fMRI data example would be discussed.

e-mail: yguo2@sph.emory.edu

112. STATISTICAL CHALLENGES IN GENOMEWIDE ASSOCIATION STUDIES

SEARCHING FOR DISEASE SUSCEPTIBILITY VARIANTS IN STRUCTURED POPULATIONS

Kathryn Roeder*, Carnegie Mellon University

Data for genome-wide association studies are being collected for a myriad of phenotypes. Many of these studies do not include control samples selected to reflect ancestry similar to the case samples. At the same time “control databases” are becoming available to be utilized as a common resource. These data are often being genotyped using a large-scale SNP array. Human populations exhibit complex structure due to heterogeneous ancestry which can lead to spurious associations if not properly handled. How to couple case and control databases effectively is a pressing question. We develop a method to match, by genetic ancestry, controls to cases. The method relies on principal component analysis and Laplacian eigenmaps for dimension reduction. We illustrate the method by matching Americans with Type 1 diabetes to controls from Germany. Despite the complex study design, these analyses identify numerous loci known to confer risk for diabetes.

e-mail: roeder@stat.cmu.edu

ARE WE READY TO LOOK AT COPY NUMBER VARIATION IN WHOLE-GENOME ASSOCIATION STUDIES?

Eleanor Feingold*, University of Pittsburgh

Cancer geneticists have looked at copy number variation in tumor cell lines for a number of years now, and the next generation of genome-wide SNP chips will give us the hypothetical capability of including copy number variants in genetic association studies. But what do we mean by copy number variation in the context of a genetic association study, how do we actually assay it, and how do we analyze the data? This talk will give both statistical and genetic perspectives on the challenges of including copy number variation in whole-genome association studies.

e-mail: feingold@pitt.edu

USING GENOTYPE IMPUTATION TO COMBINE DATA ACROSS STUDIES

Goncalo Abecasis*, University of Michigan
Yun Li, University of Michigan
Paul Scheet, University of Michigan

With millions of single nucleotide polymorphisms (SNPs) identified and characterized, genome-wide association studies have begun to identify susceptibility genes for complex traits and diseases. These studies involve the characterization and analysis of very high-resolution SNP genotype data for hundreds or thousands of individuals. Nevertheless, and despite continuing improvements in SNP genotyping technologies, most genome-wide association studies are only powered to detect a subset of the alleles associated with any one complex trait only and only directly genotype a subset of all existing SNPs. Combining multiple studies is an attractive way to gain power, but requires comparison across studies that may have used substantially different marker sets. I will review computationally efficient approaches for estimating unmeasured

genotypes, evaluating the association between these unmeasured genotypes and relevant traits, and combining results across studies. These approaches all rely on the intuition that even apparently unrelated individuals will share stretches of chromosome that include many SNPs. Once one of these stretches has been characterized in detail in a few individuals, the alleles it contains can be imputed in other carriers, with different degrees of accuracy. I illustrate the performance of the method and its potential utility using data from ongoing genome-wide association scans.

email: goncalo@umich.edu

113. RECENT ADVANCES IN ANALYZING BIOMARKER DATA WITH LIMITS OF DETECTION

RELATING A REPRODUCTIVE HEALTH OUTCOME TO SUBJECT-SPECIFIC FEATURES BASED ON LEFT- OR INTERVAL-CENSORED LONGITUDINAL EXPOSURE DATA

Kathleen A. Wannemuehler*, Centers for Disease Control and Prevention
Robert H. Lyles, Emory University
Amita K. Manatunga, Emory University
Renee H. Moore, Emory University
Metricia L. Terrell, Emory University
Michele Marcus, Emory University

Data consisting of longitudinal exposure measurements lends itself to the use of random effects models to obtain subject-specific measure of exposure that are potentially associated with health-related outcomes. Outcomes and exposure information can be tied together through either a two-stage or a joint modeling approach. In the two-stage approach, interest lies in the properties of predictors of random effects and their relative performances as covariates at the second stage. A joint or unified modeling approach arguably provides an inherent and efficient adjustment for covariate measurement error. The complexity of both approaches increases when the exposure data are subject to detection limits, are coarse due to rounding, or are otherwise interval-censored. Our research is motivated by a need to associate unobservable subject-specific exposures at a given point in time, ti*, with either a dichotomous or a continuous reproductive health outcome. We compare the use of empirical Bayes predictions in the two-stage approach with results from a joint modeling approach, with and without an adjustment for left-and interval-censored data. We further investigate the two approaches via a simulation study.

e-mail: kpw9@cdc.gov

A BAYESIAN APPROACH ESTIMATING TREATMENT EFFECTS ON BIOMARKERS CONTAINING ZEROS WITH DETECTION LIMITS

Haitao Chu*, University of North Carolina at Chapel Hill
Lei Nie, Georgetown University
Thomas W. Kensler, Johns Hopkins Bloomberg School of Public Health

Often in randomized clinical trials and observational studies in occupational and environmental health, a non-negative continuously distributed response variable denoting some metabolites of environmental toxicants is measured in treatment and control groups. When observations occur in both unexposed and exposed subjects, the biomarker measurement can be bimodally distributed with an extra spike at zero reflecting those unexposed. In the presence of left censoring due to values falling below biomarker assay detection limits, those unexposed with true zeros are indistinguishable from those exposed with left censored values. Since interventions usually do not enhance or eliminate exposure, they do not have any impact on those unexposed. Thus, only the subset of individuals who are exposed should be used to make comparisons to estimate the effect of interventions. In this article, we present Bayesian approaches using nonstandard mixture distributions to account for true zeros. The performance of the proposed Bayesian methods is compared to the maximum likelihood methods presented in Chu, Kensler and Muñoz (Statistics in Medicine, 2005: 2053-67) through simulation studies and a randomized chemoprevention trial conducted in Qidong, People's Republic of China.

e-mail: hchu@bios.unc.edu

A COMBINED EFFICIENT DESIGN BASED ON DATA SUBJECT TO DETECTION LIMIT

Enrique F. Schisterman*, National Institute of Child Health and Development-National Institutes of Health
Albert Vexler, The State University of New York at Buffalo

Pooling biospecimens, a well accepted sampling strategy in biomedical research, can be applied to reduce the cost of studying biomarkers. In some cases, even if the cost of measuring a single assay is not a major restriction in the evaluation of biomarkers, pooling is shown to be a powerful design that increases efficiency of estimation and/or testing based on data with incomplete observations due to the instruments' limit of detection. Pooled assays commonly have less probability of being below a limit of detection (LOD) than measurements of unpooled assays. However, there are situations when the pooling design strongly aggravates the detection limit problem. Moreover, there is not a straightforward solution to reconstructing statistical characteristics of individual assays based on pooled data. We propose and examine a cost-efficient hybrid design that involves taking a sample of both pooled and unpooled data in an optimal proportion in order to efficiently estimate the unknown parameters of the biomarker distribution based on data subject to LOD. We demonstrate that this design can be utilized to estimate and account for measurement error, without the need to collect validation data or repeated measurements.

e-mail: schistee@mail.nih.gov

114. BIOLOGICAL APPLICATIONS OF MACHINE LEARNING

MODELING AND DETECTION OF SPATIAL PATTERNS OF BRAIN ACTIVATION FROM fMRI DATA

Polina Goland*, Massachusetts Institute of Technology

In this talk, I will present a novel approach to computational modeling of spatial activation patterns observed through fMRI. Functional connectivity analysis is widely used in fMRI studies for detection and analysis of large networks that co-activate with a user-selected `seed' region of interest. In contrast, our method is based on clustering; it simultaneously identifies interesting seed time courses and associates voxels with the respective networks. This generalization eliminates the sensitivity to the threshold used to classify voxels as members of a network and enables discovery of co-activated networks without user selection of seed regions.

Based on the empirical observation that the detected patterns of co-activation are inherently hierarchical, we propose a new representation for spatial patterns of functional organization. Just like the anatomical hierarchies represent the structure of the brain as a tree of increasingly simple systems, we believe that the functional description of the brain should also be of a hierarchical nature. We introduce Functional Hierarchy, a top-down representation that encapsulates the notion that functionally defined regions should be viewed at different resolutions, as dictated by the observed activation pattern. We construct the functional hierarchy through an iterative decomposition that utilizes clustering for network subdivision at each step.

The experimental results demonstrate that the functional region hierarchy provides a robust and anatomically meaningful model for spatial patterns of co-activation in fMRI. The hierarchical representation leads to insights into the structure of the functional networks that are not immediately apparent from flat representations that segment the brain into a large number of small regions. In addition, subject-specific region hierarchies tend to share common tree structure, further confirming the validity of this representation for modeling group-wise patterns of co-activation.

email: polina@csail.mit.edu

REGULARIZATION APPROACH TO SCREENING BIOMARKERS

Yoonkyung Lee*, The Ohio State University

In medical studies involving such genomic data as gene expression levels and SNPs, a large number of potential biomarkers are available for prognosis or diagnosis of a disease. Finding a subset of relevant biomarkers is important for understanding of the disease and construction of effective prognostic or diagnostic rules. The method of regularization in the form of convex optimization is considered as a computationally viable alternative to the combinatorial search of the relevant biomarkers. In this talk, we present a fast and efficient linear programming algorithm for tracking the path of varying subsets of biomarkers in the regularization framework. It allows a complete exploration of the vast space of composite biomarkers and facilitates screening and selection. The algorithm will be illustrated with applications.

e-mail: yklee@stat.osu.edu

LEARNING PREDICTIVE MODELS OF GENE REGULATION

Christina Leslie*, Memorial Sloan-Kettering Cancer Center

Studying the behavior of gene regulatory networks by learning from high-throughput genomic data has become one of the central problems in computational systems biology. Most work in this area focuses on learning structure from data -- e.g. finding clusters or modules of potentially co-regulated genes, or building a graph of putative regulatory "edges" between genes -- and generating qualitative hypotheses about regulatory networks.

Instead of adopting the structure learning viewpoint, our focus is to build predictive models of gene regulation that allow us both to make accurate quantitative predictions on new or held-out experiments (test data) and to capture mechanistic information about transcriptional regulation. Our algorithm, called MEDUSA, integrates promoter sequence, mRNA expression, and transcription factor occupancy data to learn gene regulatory programs that predict the differential expression of target genes. MEDUSA does not rely on clustering or correlation of expression profiles to infer regulatory relationships. Instead, the algorithm learns to predict up/down expression of target genes by identifying condition-specific regulators and discovering regulatory motifs that may mediate their regulation of targets. We use boosting, a technique from machine learning, to help avoid overfitting as the algorithm searches through the high dimensional space of potential regulators and sequence motifs. We will describe results of a recent gene expression study of hypoxia in yeast, in collaboration with the lab of Li Zhang. We used MEDUSA to propose the first global model of the oxygen and heme regulatory network, including new putative context-specific regulators. We then performed biochemical experiments to confirm that regulators identified by MEDUSA indeed play a causal role in oxygen regulation.

email: cleslie@cbio.mskcc.org

115. HEALTH SERVICES AND POLICY RESEARCH

INFORMATIVE SCREENING

Joshua M. Tebbs*, University of South Carolina
Christopher R. Bilder, University of Nebraska

When estimating the prevalence of a rare trait, the use of pooled (group) testing can confer substantial benefits when compared to individual testing. In addition to screening experiments for infectious diseases, pooled testing has also been used in other biomedical applications such as drug discovery, epidemiology, and genetics. Until recently, nearly all biomedical studies using pooled testing have treated individuals as homogenous. However, in most real problems, researchers have access to covariate information on each individual subject, and a major thrust of the analysis is to understand how this information can be exploited. In this talk, we will introduce a new approach called "informative group testing," whereby observed covariates are used to determine how pooled responses can be efficiently converted into individual responses. We will illustrate our approach using chlymadia screening data collected by the Nebraska State of Public Health Laboratory.

e-mail: tebbs@stat.sc.edu

SMALL-AREA ESTIMATION OF MENTAL ILLNESS PREVALENCE FOR SCHOOLS

Fan Li*, Harvard Medical School
Alan M. Zaslavsky, Harvard Medical School
Ronald Kessler, Harvard Medical School

We use data collected in the National Comorbidity Survey - Adolescent (NCS-A) to develop a methodology to estimate the small-area prevalences of serious emotional disturbance (SED) in schools in the United States, exploiting the clustering of the main NCS-A sample by school. The NCS-A instrument includes both a short (6-item) screening scale, the K6, and extensive diagnostic assessments of the individual disorders and associated mental distress that determine the diagnosis of SED. We fitted a Bayesian bivariate multilevel regression model with correlated effects for the probability of SED and the K6 score at the individual, school, and geographical levels. Under this model we obtain a prediction equation for the rate of SED based on the mean K6 score and covariates. Our results provide evidence for the existence of variation in the prevalence of SED across schools and geographical regions. Furthermore, although the concordance between the K6 scale and SED is only modest for individuals, the school-level random effects for the two measures are strongly correlated. This finding supports the feasibility of using short screening scales like the K6 as alternative to more comprehensive lay assessments in estimating school-level rates of SED among adolescents. These methods may be applicable to other studies aiming at small-area estimation for geographical units.

e-mail: li@hcp.med.haravard.edu

BAYESIAN ORDERING OF A TRINOMIAL SAMPLE SPACE FOR CLASSICAL HYPOTHESIS TESTING

A. John Bailer, Miami University
Robert B. Noble, Miami University
Douglas A. Noe*, Miami University

In the classical hypothesis testing framework, a p-value is defined as the probability that a random sample from the distribution assumed under the null hypothesis will be 'as extreme as or more extreme than' the data actually observed. In many applications, 'extremity' is a natural concept; however, in other cases it must be carefully defined. In our application, we conduct inference on the ratio of two probabilities defining a trinomial distribution. Given a random sample of a particular size, we need to order the multinomial vectors in the sample space in terms of extremity relative to the null hypothesis ratio of probabilities. Though several seemingly natural ad-hoc measures exhibited difficulties, our Bayesian approach produced the most qualitatively acceptable ordering. This application is motivated and then illustrated with data from a study of the agreement of raters of nursing home residents.

e-mail: noeda@muohio.edu

BACKWARD ESTIMATION OF MEDICAL COST IN THE PRESENCE OF A FAILURE EVENT

Kwun Chuen Gary Chan*, Bloomberg School of Public Health-Johns Hopkins University
Mei-Cheng Wang, Bloomberg School of Public Health-Johns Hopkins University

The backward medical cost process is introduced for modeling end-of-life-cost. Estimator for mean backward cost process is proposed for right censored and left truncated right censored data. The proposed estimator converges weakly to a Gaussian process and we propose a method for constructing simultaneous confidence band by simulations. We illustrate the proposed methodologies by analyzing the SEER-Medicare linked dataset. The results shows that among ovarian cancer patients, half of the final year of life cost is spent in the final three months of life.

e-mail: kcchan@jhsph.edu

SOME ISSUES IN SUICIDE ATTEMPT PREDICTION

Steven P. Ellis*, Columbia University

Suicide is a major cause of mortality worldwide. Therefore, constructing a practical method for predicting suicides, or even suicide attempts, is an important project. A prediction method (call it a 'rule') takes patient attributes as input and produces as output some statement about future suicidal behavior of the patient. We want to develop such a rule using data. A binary prediction rule is one whose output is either 'yes, the patient will make an attempt' (in the time period of interest) or 'no, the patient will not make an attempt'. It is widely accepted that binary prediction of suicide attempts cannot be done well. More promising are 'fine grained' prediction rules, two examples of which are (1) rules that output probabilities, or (2) rules that output actual suicide prevention treatment recommendations (a 'dynamic treatment regime', Murphy et al JASA '01). Construction of either kind of rule would be aided by, even a simple, analysis of the cost of suicide attempt and costs and benefits (in terms of reduced probability of attempt) of preventive treatment. In this talk, I will examine these issues, but will not discuss the technical problems of actually computing rules from data.

e-mail: spe4@columbia.edu

GAMMA SHAPE MIXTURES FOR HEAVY-TAILED DISTRIBUTIONS

Sergio Venturini*, Bocconi School of Management-Italy
Francesca Dominici, Johns Hopkins Bloomberg School of Public Health, Baltimore
Giovanni Parmigiani, The Sidney Kimmel Comprehensive Cancer Center-Johns Hopkins University

An important question in health services research is the estimation of the proportion of medical expenditures that exceed a given threshold. Typically, medical expenditures present highly skewed, heavy tailed distributions, for which a) simple variable transformations are insufficient to achieve a tractable low-dimensional parametric form and b) nonparametric methods are not efficient in estimating exceedance probabilities for large thresholds. In this paper we propose a general Bayesian approach for the estimation of tail probabilities of heavy-tailed distributions, based on a mixture of gamma distributions in which the mixing occurs over the shape parameter. This family provides a flexible and novel approach for modeling heavy-tailed distributions, it is computationally efficient, and it only requires to specify a prior distribution for a single parameter. By carrying out simulation studies, we compare our approach with commonly used parametric and non-parametric alternatives. We found that the mixture-gamma model significantly improves predictive performance in estimating tail probabilities, compared to these alternatives. We also applied our method to the Medical Current Beneficiary Survey (MCBS), for which we estimate the probability of exceeding a given hospitalization cost for smoking attributable diseases.

e-mail: sergio.venturini@sdabocconi.it

A BAYESIAN LOCATION SHIFTED MIXTURE MODEL FOR HMO PHARMACOVIGILANCE RESEARCH DATABASE

Fang Zhang*, Harvard University
Martin Kulldorff, Harvard University

Early detection of potential adverse effects of medications is crucial and existing methods to detect adverse drug reactions may be limited. With over 11 million enrollees, the HMO Research Network CERT provides a well-defined population with relatively complete drug dispensing and outcome data. Such databases provide an important complement to the existing spontaneous reporting systems because they allow calculation of denominators of drug users and better estimated ‘expected counts’ of adverse events among unexposed health plan members. To fully utilize such large claim-based databases, we propose to use a new mixed structure method based on the two distinguished components where one component is from a distribution with mean 1 and the other is always larger than 1, with certain threshold mean based on one’s experiences or beliefs. Both empirical and fully Bayesian approach of such frame work are investigated. We compare them with the existing empirical Bayesian tool for AERS using simulation. We employ a large pool of prior selections for the fully Bayesian approach to investigate the sensitivity of the method. Several statistical criteria would be utilized to compare the performances. Lastly, we analyze available HMO datasets using the newly proposed method and existing data mining methods.

e-mail: fang_zhang@hphc.org

116. MEASUREMENT ERROR

MISCLASSIFICATION ADJUSTMENT IN THRESHOLD MODELS FOR THE EFFECTS OF SUBJECT-SPECIFIC EXPOSURE MEANS AND VARIANCES

Chengxing Lu*, Emory University
Robert H. Lyles, Emory University

In environmental epidemiology, researchers sometimes assume the existence of exposure thresholds above which the risk of adverse effects begins to increment. In this work, we assume exposures are measured repeatedly over time, and the research question of interest is to identify the relationship between a health-related outcome and whether or not the subject-specific mean and/or variability of the exposure exceed known thresholds. As a subject’s true exposure mean and variability cannot be observed directly, misclassification typically arises in the categorization of whether these quantities are above their respective thresholds. Building off of random effects models for repeated exposure measurements and assuming balanced data, methods based on the well-known regression calibration and matrix methods are demonstrated for the case of the mean exposure only. For unbalanced data, and to incorporate categorizations based on both the exposure mean and variance, a maximum likelihood approach is introduced. Simulation results and a real study example from the Mount Sinai Study of Women Office Workers (MSSWOW) are presented to demonstrate the performance of the methods.

e-mail: clu@emory.edu

AVERAGE CAUSAL EFFECT ESTIMATION ALLOWING COVARIATE MEASUREMENT ERROR: A FINITE MIXTURE MODELING FRAMEWORK

Yi Huang*, University of Maryland,
Karen Bandeen-Roche, Johns Hopkins University
Constantine Frangakis, Johns Hopkins University

In many applications of the propensity score subclassification approach, the underlying true covariates, which might confound the average causal effect (ACE) assessment of a binary treatment on health outcomes, are not directly observable, but rather are measured with error. A naive approach is to use the observed covariates to estimate propensity scores and then use these to subclassify and estimate ACEs. The work we report introduces appropriate causal assumptions for applications with errors-in-covariates and demonstrates that the naive approach for ACE estimation typically produces biased results. We also propose a flexible finite mixture framework for ACE estimation reflecting a covariate-balancing criterion in a joint likelihood, which unifies subgroup membership assessment and subgroup-specific treatment effect evaluation allowing for measurement error in the covariates. In all, this talk aims to improve causal inferences in situations involving unobserved measurement error in covariates, in analyses based on propensity score subclassification.

e-mail: yihuang@umbc.edu

MEASUREMENT ERROR MODEL WITH UNKNOWN ERROR

Peter Hall, University of Neuchatel,
Yanyuan Ma*, Australian National University

We consider functional measurement-error models where the measurement-error distribution is estimated nonparametrically. We derive a locally efficient semiparametric estimator, but propose not to implement it due to its numerical complexity. Instead, a plug-in estimator is proposed, where the measurement-error distribution is estimated through nonparametric kernel methods based on multiple measurements. The root-n consistency and asymptotic normality of the plug-in estimator is derived. Despite the theoretical inefficiency of the plug-in estimator, simulations demonstrate its near-optimal performance. Computational advantages relative to the theoretically efficient estimator make the plug-in estimator practically appealing. Application of the estimator is illustrated using the Framingham data example.

e-mail: yanyuanma@yahoo.com

MODELING HEAPING IN SELF-REPORTED CIGARETTE COUNTS

Hao Wang*, University of Pennsylvania
Daniel F. Heitjan, University of Pennsylvania

In studies of smoking behavior, some subjects report exact cigarette counts, whereas others report rounded-off counts, particularly multiples of 20, 10 or 5. This form of data reporting error, known as heaping, can bias the estimation of parameters of interest such as mean cigarette consumption. We present a model to describe heaped count data from a randomized trial of bupropion treatment for smoking cessation. The model posits that the reported cigarette count is a deterministic function of an underlying precise cigarette count variable and a heaping behavior variable, both of which are at best partially observed. To account for an excess of zeros, as would likely occur in a smoking cessation study where some subjects successfully quit, we model the underlying count variable with zero-inflated distributions. We study the sensitivity of the inference on smoking cessation by fitting various models that either do or do not account for heaping and zero inflation, comparing the models by means of Bayes factors. Our results suggest that sufficiently rich models for both the underlying distribution and the heaping behavior are indispensable to obtaining a good fit with heaped smoking data. The analyses moreover reveal that bupropion has a significant effect on the fraction abstinent, but not on mean cigarette consumption among the non-abstinent.

e-mail: haow@mail.med.upenn.edu

LINEAR MODEL COVARIATE MEASUREMENT ERROR CORRECTION VIA MULTIPLE IMPUTATION

Miguel A. Padilla*, University of Alabama at Birmingham
Jasmin Divers, Wake Forest University
Hemant K. Tiwari, University of Alabama at Birmingham

Statistical linear models are widely used to analyze of variety of data from the behavioral to the biological sciences. The typical assumptions of linear models are that the errors are normally and independently distributed with common variance. Even so, when the sample size is sufficiently large linear models tend to be robust to the normally and common variance assumption. Nevertheless, linear models have one additional assumption that is sometime forgotten: all variables in the model are measured without error. However, here we discuss measurement error in the covariate. Measurement error in the covariate tends to attenuate its associated parameter estimate and confound the residual variance. Multiple imputation is presented as a viable tool for correcting measurement error problems in linear models with emphasis on correcting a measurement error contaminated covariate. The method is assessed in terms of Type I error and power. Factors investigated are degree of measurement error, sample size, number of imputations, and degrees of freedom. Results indicate that multiple imputation can be used to correct covariate measurement error in linear models. However, the covariate should be reasonably measured because the method uses the existing data to borrow more information in which to make the measurement error corrections.

e-mail: mpadilla@uab.edu

RECURSIVE ESTIMATION METHOD FOR PREDICTING RESIDUAL BLADDER URINE VOLUMES TO IMPROVE ACCURACY OF TIMED URINE COLLECTIONS

David Afshartous*, University of Miami, Miller School of Medicine
Richard A. Preston, University of Miami, Miller School of Medicine

Clinical research studies often collect data via repeated measurements of collected urine. Unfortunately, the accuracy of timed urine collections is limited by the presence of a residual volume of urine remaining in the bladder following each timed void due to incomplete emptying of the bladder. This residual urine volume adds significant imprecision to the urine collection method, rendering an important and fundamental clinical research tool inaccurate. We present an unbiased method to estimate the residual bladder volumes via a mathematical model of the bladder process. Regardless of the substance of primary interest, the model leverages conservation of mass and conservation of concentration principles towards a substance of secondary interest in order to solve a system of recursive equations, resulting in our Recursive Residual Estimation to predict the residual volumes at each time point. We verify the model on simulated patients and also investigate the sensitivity of the model to initial value specification.

e-mail: dafshartous@med.miami.edu

USING ORTHOGONAL DECOMPOSITION TO ADJUST REGRESSION CALIBRATION IN PHYSICAL ACTIVITY STUDIES

Sanguo Zhang*, Vanderbilt University School of Medicine

Epidemiologic studies have demonstrated that physical activity(PA) is inversely associated with numerous chronic diseases. However, measurement of PA using questionnaires contains a significant amount of measurement errors(ME) which may lead to biased estimates of the relative risk. In a validation study imbedded in a large cohort study, Shanghai Men and Women's Health Study, we attempt to correct ME in questionnaires using an alloyed golden standard, 7-day PA logs. Due to correlation between the errors from the two measurements, the traditional regression calibration may also lead to biased estimates. We propose adjusting regression calibration methods by orthogonal decomposition of the errors. Simulation studies demonstrate that our methods can effectively correct for bias induced by the correlated errors. Our methods can be generally applied to studies in which both the error-prone measurement and the alloyed gold standard measurement are obtained more than once, preferably simultaneously, in the validation study. An advantage of our methods is that we do not need a third measurement as discussed extensively in the literature for such settings.

e-mail: sanguo.zhang@vanderbilt.edu

117. MICROARRAY DATA ANALYSIS

A GEOMETRIC APPROACH FOR DETECTING CELL CYCLE GENE EXPRESSION

Omar De la Cruz*, University of Chicago
Dan L. Nicolae, University of Chicago

We propose a method for detecting cell-cycle-regulated genes by studying the geometric structure of gene expression data. Since this method does not require a synchronization step (as has been used to study the cell cycle in yeast), it can in principle be applied to any growing cell population, e.g.: embryonic, stem, epithelial, apical, or tumor cells, from any species. Starting from a data set containing the expression level for m genes in n individual cells randomly sampled from a growing population, we consider it as a set of n points in m-dimensional Euclidean space. Under reasonable assumptions, these points cluster around a closed curve that represents the ideal evolution of expression levels during the cycle. The core of our method is finding

the (parameterized) curve that best fits the points. We will present some theoretical results and examples using simulated as well as existing time-course and single-cell data, as arguments for the usefulness of the method.

email: odlc@uchicago.edu

TWO-CRITERION: A NEW BAYESIAN DIFFERENTIALLY-EXPRESSED GENE SELECTION ALGORITHM

Fang Yu*, University of Nebraska Medical Center
Ming-Hui Chen, University of Connecticut
Lynn Kuo, University of Connecticut

Recently, the Bayesian method becomes more popular for analyzing high dimensional gene expression data as it allows us to borrow information across different genes and provides more powerful estimators for evaluating gene expression levels. It is crucial to develop a simple but efficient gene selection algorithm for detecting differentially expressed (DE) genes based on the Bayesian estimators. In this paper, we extend the idea of Chen, Ibrahim, and Chi (2008) to propose a new general gene selection algorithm, namely, the two-criterion, for any Bayesian model. The proposed two-criterion method first evaluates the posterior probability of a gene having different gene expressions between two competitive samples. If the obtained posterior probability is large enough, it declares the gene to be DE. The theoretical connection between the proposed two-criterion and the Bayes factor is established under the normal-normal-model with equal variances between two samples. The performance of the proposed method is examined and compared to those of several existing methods.

e-mail: fangyu@unmc.edu

A NON-PARAMETRIC META-ANALYSIS APPROACH FOR COMBINING INDEPENDENT MICROARRAY DATASETS: APPLICATION USING TWO MICROARRAY DATASETS PERTAINING TO CHRONIC ALLOGRAFT NEPHROPATHY

Xiangrong Kong*, Virginia Commonwealth University
Valeria Mas, Virginia Commonwealth University
Kellie J. Archer, Virginia Commonwealth University

With the popularity of DNA microarray technology, multiple groups of researchers have studied the gene expression of similar biological conditions. Different methods have been developed to integrate the results from various microarray studies, though most of them rely on distributional assumptions, such as the t-statistic based, mixed-effects model, or Bayesian model methods. However, often the sample size for each individual microarray experiment is small. Therefore, in this paper we present a non-parametric meta-analysis approach for combining data from independent microarray studies, which is based on an innovative application of a non-parametric pattern recognition method (K-Nearest Neighbor). The rationale behind the approach is logically intuitive and can be easily understood by clinical researchers not having advanced training in statistics. The simulation study comparing the non-parametric meta-analysis approach to a commonly used t-statistic based approach shows that the non-parametric approach has better sensitivity and specificity. We illustrate the application of our meta-analytic approach on two independent Affymetrix GeneChip studies that compared the gene expression of biopsies from kidney transplant recipients with chronic allograft nephropathy (CAN) to those with normal functioning allograft. Further study on the identified genes may lead to better understanding of CAN at the molecular level.

e-mail: kongx@vcu.edu

RXA: ANALYSIS OF GENE EXPRESSION MICROARRAYS BASED ON EXPRESSION ORDERINGS

Xue Lin*, Johns Hopkins University
Daniel Naiman, Johns Hopkins University
Leslie Cope, Johns Hopkins University
Giovanni Parmigiani, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University
Donald Geman, Johns Hopkins University

RXA(Relative Expression Analysis) is a family of rank-based methods for the statistical analysis of gene expression microarray data. The first and simplest form of RXA is the TSP(Top Scoring Pair) classifier for distinguishing between two classes (e.g., phenotypes). If A and B are the expression values of two genes, the TSP classifier associates the order A>B with one class and B>A with the other. Optimal pairs are selected by maximizing the score function f(A,B)=|Pr(A>B|class 1)-Pr(A>B|class 2)|. TSP is intuitive and has produced impressive results in various classification problems. We extend TSP to TST(Top Scoring Triplet) by bringing in one more gene. The TST method considers the six possible orderings among three genes; the decision rule is (again) a simple likelihood ratio test. Although the extension involves only one more gene, it raises interesting computational and modeling problems. We apply TST to two breast cancer data sets and it shows high accuracy in predicting BRCA1 mutations. Permutation tests and cross validation are used to validate the results.

e-mail: xlin@ams.jhu.edu

PROFILING TIME COURSE EXPRESSION OF VIRUS GENES

I-Shou Chang*, National Health Research Institutes-Taiwan
Li-Chu Chien, National Health Research Institutes-Taiwan
Chao Hsiung, National Health Research Institutes-Taiwan

We illustrate a hierarchical Bayesian shape restricted regression method in making inference on the time course expression of virus genes, using data from microarray experiments. The prior is introduced through Bernstein polynomials so as to take into consideration the geometry of the regression functions, which are assumed to be zero initially, increasing afte a while and finally decreasing. A MCMC algorithm is used to sample the posterior distribution. One advantage of this method is that it offers an assessment of the strength of the evidence provided by the data in favor of hypothesis on the shape of the regression; for example, the hypothesis that it is unimodal. Another advantage of this approach is that estimates of many salient features of the profile like onset time, inflection point, maximum value, time to maximum value, etc. can be obtained immediately.

e-mail: ischang@nhri.org.tw

IDENTIFY EQUIVALENTLY EXPRESSED GENES FROM MICROARRAY DATA USING A MIX-SCALED EQUIVALENCE CRITERION

Jing Qiu, University of Missouri-Columbia
Xiangqin Cui*, University of Alabama at Birmingham

Besides the common use in identifying differentially expressed genes, microarrays can be used to identify genes that are equivalently expressed across conditions. This application is important in identifying genes expressed at constant levels and evaluating the degree of equivalence in overall gene expression across conditions. There has been little consideration on formal statistical inferential methods for identifying equivalently expressed genes in the microarray data analysis literature. We applied the concept of equivalence testing required by the Food and Drug Administration (FDA) for establishing therapeutic equivalence and developed a mix-scaled criterion for establishing expression equivalence in microarray studies. An average equivalence criterion is used for naturally stable genes and a scaled average equivalence criterion is used for naturally variable genes. Corresponding to the mix-scaled criterion, our test procedure combines the two-one-sided-t (TOST) test and a uniformly-most-powerful-invariant (UMPI) test. We evaluated the false discovery rate and the power of our test procedure using two spike-in microarray data sets with technical replicates, one rat data set with biological replicates, and simulations based on these data sets. Our test procedure showed good power even for small sample sizes and the false discovery rate was under control.

e-mail: xcui@uab.edu

A STATISTICAL FRAMEWORK FOR INTEGRATING DIFFERENT MICROARRAY DATA SETS IN DIFFERENTIAL EXPRESSION ANALYSIS

Yinglei Lai*, George Washington University

Micorarrays have been widely used to detect differential expression in biological and medical studies. Different microarray data sets can be collected for the same study with different microarray platforms and/or from different laboratories. We expect to achieve more efficient detection of differential expression if an efficient statistical method can be developed to integrate these different microarray data sets. In this study, we propose a statistical framework for this purpose. Based on the expression measurements for the common genes in different data sets, we first evaluate the genome-wide p-values for each individual data set and then transform the p-values into their corresponding z-scores to achieve efficient normal mixture modeling. To avoid spurious results, it is necessary to perform the concordance/discordance tests before the data integration. Based on the test results, different subsequent actions are suggested. The data integration can be considered if we can reject the hypothesis of complete discordance; otherwise, the data integration is discouraged. We develop a bivariate normal mixture model based method to integrate different lists of genome-wide z-scores. The mixture model can be reduced if we cannot reject the hypothesis of complete concordance. Genes can be prioritized by their probabilities of being concordantly differentially expressed.

e-mail: ylai@gwu.edu

118. SURVIVAL ANALYSIS METHODS AND APPLICATIONS

ACCELERATED QUANTILE RESIDUAL LIFE MODEL

Hanna Bandos*, University of Pittsburgh
Jong-Hyeon Jeong, University of Pittsburgh

Non-uniqueness of lifetime distribution that corresponds to a specific percentile residual life function limits the usefulness of the function in practice, especially in the regression setting. In this work we establish a connection between the accelerated failure time (AFT) model and a new accelerated quantile residual life (AQRL) model. We demonstrate that the AFT model is a necessary condition for the AQRL model and identify the set of points for which both models are equivalent. We show that by restricting to a specific family of parametric distributions a priori, the non-uniqueness problem can be overcome. For some families this condition allows for a life distribution to be uniquely determined by the percentile residual life function under the AQRL model. We also show that under the same condition any location-scale family guarantees a one-to-one correspondence between the AFT and the AQRL model.

e-mail: bandos@nsabp.pitt.edu

BAYESIAN CASE INFLUENCE DIAGNOSTICS FOR SURVIVAL MODELS

Hyunsoon Cho*, University of North Carolina at Chapel Hill
Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Debajyoti Sinha, Medical University of South Carolina
Hongtu Zhu, University of North Carolina at Chapel Hill

We propose Bayesian case influence diagnostics for complex survival models. We develop case deletion influence diagnostics on both the joint and marginal posterior distributions based on the Kullback-Leibler divergence (K-L divergence). We present a simplified expression for computing the K-L divergence between the posterior with full data and the posterior based on single case deletion, as well as investigate its relationships to the Conditional Predictive Ordinate (CPO). All the computations for the proposed diagnostic measures can be easily done using Markov chain Monte Carlo (MCMC) samples from the full data posterior distribution. We consider the Cox model and a frailty model with a gamma process prior on the cumulative baseline hazard. We present a theoretical relationship between our case-deletion diagnostics and diagnostics based on Cox's partial likelihood. A simulated data example and two real data examples are given to demonstrate the methodology.

e-mail: hscho@bios.unc.edu

A ROBUST WEIGHTED KAPLAN-MEIER APPROACH USING LINEAR COMBINATIONS OF PROGNOSTIC COVARIATES

Chiu-Hsieh Hsu*, University of Arizona
Jeremy M.G. Taylor, University of Michigan

We consider combining multiple prognostic covarites into two risk scores. One is derived from a working model for the event times. The other is derived from a working model for the censoring times. These two risk scores are then categorized to define the risk groups to conduct the weighted Kaplan-Meier estimator. We show the weighted Kaplan-Meier estimator is robust to mis-specification of either one of the two working models. In a simulation study, we show the robust weighted Kaplan-Meier approach can reduce bias due to dependent censoring and improve efficiency.

e-mail: phsu@azcc.arizona.edu

MODELING DISCRETE SURVIVAL DATA WITH APPLICATION TO REPRODUCTIVE STUDY

Huichao Chen*, Harvard University
Amita K. Manatunga, Emory University
Limin Peng, Emory University
Michele Marcus, Emory University

In reproductive studies, time-to-pregnancy (TTP) is measured as the number of menstrual cycles it takes a woman to conceive a pregnancy. We are interested in investigating the influence of potential covariates such as age and body mass index etc. on time-to-pregnancy. We propose an adaption of proportional odds model for modeling discrete TTP subject to right censoring and left truncation. Furthermore, we consider an alternative approach which models TTP in terms of a conditional probability (hazard) model. Using data from the Mount Sinai Study of Women Office Workers (MSSWOW), we compare and contrast our results from different models.

e-mail: hchen@sdac.harvard.edu

GOODNESS-OF-FIT TESTING FOR THE COX PROPORTIONAL HAZARDS MODEL WITH TIME-DEPENDENT COVARIATES

Susanne May*, University of California-San Diego
Jennifer Emond, University of California-San Diego
Steve Edland, University of California-San Diego
Loki Natarajan, University of California-San Diego

Many graphical and formal goodness-of-fit tests for the Cox proportional hazards (PH) model have been developed. Only a few are easy to compute or implemented in statistical software packages and are typically aimed at models with covariates that are constant over time. Categorical or continuous covariates that vary over time play an important role for many research questions that use the Cox PH model, but some of the available goodness-of-fit tests can not be used in this setting. We review the goodness-of-fit tests that are applicable in the presence of time-dependent covariates and focus on those that are either easy to compute or are available in statistical software packages. We illustrate the use of the tests and challenges regarding their interpretation using simulations and data from the Alzheimer's Disease Research Center.

e-mail: smay@ucsd.edu

CENSORED MEDIAN REGRESSION AND PROFILE EMPIRICAL LIKELIHOOD

Sundar Subramanian*, New Jersey Institute of Technology

Profile empirical likelihood-based inference for censored median regression models is investigated. Inference for any specified subvector is carried out by profiling out the nuisance parameters from a "plugin" empirical likelihood ratio function. The limiting distribution is a sum of weighted chi square distributions having intractable covariance structure. To obtain the critical value, the bootstrap is employed and the resulting confidence intervals are compared with the ones obtained through a minimum dispersion statistic. An illustration using a lung cancer data set is provided.

e-mail: sundars@njit.edu

IT'S ALL ABOUT THE PROPORTIONAL HAZARDS ASSUMPTION

Kyung Y. Lee*, U.S. Food and Drug Administration
Yuan-Li Shen, U.S. Food and Drug Administration
Kallappa M. Koti, U.S. Food and Drug Administration

Cox's proportional hazards model is widely used for the analysis of covariate effects on a hazard ratio using censored survival data, but the proportional hazards (PH) assumption that is made is often inappropriate. The proportional hazards assumption must be assessed and dealt with if it is not true. We examine existing procedures for checking appropriateness of Cox's proportional hazards model using several real survival datasets. We argue that the graphical method of diagnosis such as SAS LS/LLS plot is simpler and more authentic in most cases for assessing proportional hazards assumption compared to the one based on the time-dependent covariate.

e-mail: yuan.li.shen@hhs.fda.gov

119. VARIABLE SELECTION AND MODEL BUILDING - APPLICATIONS

RECURSIVE PARTITIONING ANALYSIS AND PROPORTIONAL HAZARD MODELS IN PROGNOSTIC FACTORS ANALYSIS: A JOINT COOPERATIVE GROUPS STUDY FOR HIGH GRADE RECURRENT GLIOMA

Wenting Wu*, Mayo Clinic
Kathleen Lamborn, University of California-San Francisco
Paul Novotny, Mayo Clinic
Terry Therneau, Mayo Clinic

A major issue in the analysis of disease is the identification and assessment of prognostic factors relevant to the development of the illness. Statistical analyses within the Proportional hazards (PH) analysis framework suffer from a lack of flexibility due to stringent model assumptions such as additivity and linearity. Recursive Partitioning Analysis (RPA) provides additional flexibility and more readily highlights interactions among variables. A retrospective study including 27 clinical trials was conducted to evaluate how the different RPA algorithms compared to each other and to PH and also to explore methods for adjustment for nuisance variables. Time-to-event (OS, PFS) and binary (PFS6) endpoints were considered. RPA with cross validation, and PH and Logistic regression with validation using bootstrap samples were used to identify prognostic variables and to look for evidence of interactions. Comparisons were made among RPA splitting methods using different methods for adjusting for nuisance variables and between these RPA methods and conclusions based on PH/logistic regression models. RPA and PH models provided complementary information. RPA using different splitting methods could generate different models. Adjusting for nuisance parameters is very important in obtaining meaningful prognostic models.

email: wu.wenting@mayo.edu

NETWORK-CONSTRAINED REGULARIZATION AND VARIABLE SELECTION FOR ANALYSIS OF GENOMIC DATA

Caiyan Li*, University of Pennsylvania School of Medicine
Hongzhe Li, University of Pennsylvania School of Medicine

Graphs or networks are common ways of depicting information. In biology in particular, many different biological processes are represented by graphs, such as regulatory networks or metabolic pathways. This kind of a priori information gathered over many years of biomedical research is a useful supplement to the standard numerical genomic data such as microarray gene expression data. How to incorporate information encoded by the known biological networks or graphs into analysis of numerical data raises interesting statistical challenges. In this paper, we introduce a network-constrained regularization procedure for linear regression analysis in order to incorporate the information from these graphs into an analysis of the numerical data, where the network is represented as a graph and its corresponding Laplacian matrix. We define a network-constrained penalty function that penalizes the L_1-norm of the coefficients but encourages smoothness of the coefficients on the network. An efficient algorithm is also proposed for computing the network-constrained regularization paths, much like the Lars algorithm does for the lasso. We illustrate the methods using simulated data and analysis of a microarray gene expression data set of glioblastoma.

e-mail: licaiyan@mail.med.upenn.edu

CONFOUNDER-SELECTION USING THE CHANGE IN ESTIMATE APPROACH; IS 10% ALL WE NEED TO KNOW?

Gheorghe Doros*, Boston University
Robert Lew, V.A. Boston
Alexander Ozonoff, Boston University

A number of approaches for selecting a minimal set of variables to control for based on significance testing have been proposed in the literature. These approaches include the conventional model selection approaches based on significance testing. One common alternative that avoids using significance testing is the ‘change in estimate’ (CIE) approach. Some simulation studies, which compare a number of confounder-selection approaches, found the CIE approach with a 10% cutpoint to be best from among a number of confounder selection criteria that included CIE with various cutpoints of 5%, 10%, 15%, 20% and 25%. These results have been echoed in the applied epidemiology as ‘CIE with a 10% cutpoint is good practice’. In this paper we study the performance of the CIE approach with a fixed cutpoint and propose an alternative approach.

e-mail: doros@bu.edu

INCORPORATING PRIOR KNOWLEDGE OF GENE FUNCTIONAL GROUPS INTO REGULARIZED DISCRIMINANT ANALYSIS OF MICROARRAY DATA

Feng Tai*, University of Minnesota
Wei Pan, University of Minnesota

Discriminant analysis for high-dimensional and low-sample-sized data has become a hot research topic in bioinformatice, mainly motivated by its importance and challenge in applications to tumor classifications for high-dimensional microarray data. Two of the popular methods are the nearest shrunken centroids, also called predictive analysis of microarray (PAM), and shrunken centroids regularized discriminant analysis (SCRDA). Both methods are modifications to the classic linear discriminant analysis (LDA) in two aspects tailored to high-dimensional and low-sample-sized data: one is the regularization of the covariance matrix, and the other is variable selection through shrinkage. In spite of their usefulness, both PAM and SCRDA are possibly too extreme: the covariance matrix in the former is restricted to be diagonal while in the latter there is barely any restriction. We propose modified LDA methods to integrate biological knowledge of gene functions (or variable groups) into classification of microarray data. We group the genes according to their biological functions extracted from existing biological knowledge or data, and propose regularized covariance estimators that encourages between-group gene independence and within-group gene correlations while maintaining the flexibility of any general covariance structure. Furthermore, we propose a shrinkage scheme on groups of genes that tends to retain or remove a whole group of the genes altogether, in contrast to the standard shrinkage on individual genes.

e-mail: fengtai@biostat.umn.edu

HIERARCHICALLY PENALIZED COX MODEL FOR SURVIVAL DATA WITH GROUPED VARIABLES AND ITS ORACLE PROPERTY

Sijian Wang*, University of Michigan
Nengfeng Zhou, University of Michigan
Bin Nan, University of Michigan
Ji Zhu, University of Michigan

In many biological and other scientific applications, predictor variables are naturally grouped. For example, in biological applications, assayed genes or proteins are grouped by biological roles or biological pathways. When studying the dependence of survival outcome on these grouped predictor variables, it is desired to select variables at both the group level and variable-specific level. In this paper, we develop a new method to handle the group variable selection in Cox's proportional hazards model. Our method not only effectively removes unimportant groups, but also maintains the flexibility of selecting variables within the identified groups. We also show that the new method offers the potential for achieving the asymptotic oracle property as in Fan & Li (2002).

e-mail: sijwang@umich.edu

CLASSIFICATION OF FAMILIES OF LOCALLY STATIONARY TIME SERIES

Robert T. Krafty*, University of Pittsburgh
Wensheng, Guo, University of Pennsylvania

In real applications, such as the analysis of EEG, time series data are rarely stationary. We introduce a class of families of locally stationary time series that accounts for within-population spectral variability. This class is used to create a procedure for discriminating between different populations of time series that takes advantage of within-population variability. A hierarchical structure is placed on the log-spectra, smoothing spline ANOVA estimation procedures are developed for the estimation of the mean and variance structures of the log-spectra in each group, and asymptotic distributions are computed and used to create a quadratic classification rule. A Kullback-Leibler criterion is developed for the data driven selection of smoothing parameters. The procedure is applied to an EEG data set to predict the onset of epileptic seizures and illustrate the benefits in accounting for within-population spectral variability in the discrimination of non-stationary time series.

e-mail: krafty@pitt.edu

120. NONPARAMETRIC METHODS

NONPARAMETRIC ADDITIVE REGRESSION FOR NONPARAMETRIC ADDITIVE REGRESSION FOR REPEATEDLY MEASURED DATA

Raymond Carroll, Texas A&M University
Arnab Maity*, Texas A&M University
Enno Mammen, University of Mannheim-Germany
Kyusang Yu, University of Mannheim-Germany

There has been considerable recent interest and development in the fitting of nonparametric regression for repeated measures data. The literature has focused almost exclusively on the case that the function to be fitted nonparametrically has a scalar argument. In this paper, we develop an easily computed smooth backfitting algorithm for additive model fitting in such repeated measures problems. Our methodology easily copes with various settings, such as when some covariates are the same over repeated response measurements. We allow for a working covariance matrix for the regression errors, showing that our method is most efficient when the correct covariance matrix is used. The component functions achieve the known asymptotic variance lower bound for the scalar argument case. We also discuss a number of important extensions. We first describe the behavior of our estimators when the underlying model is not additive. We apply them to the partially linear additive repeated-measures model, deriving an explicit consistent estimator of the parametric component; if the errors are in addition Gaussian, the estimator is semiparametric efficient. Finally, we apply our basic methods to a unique testing problem that arises in genetic epidemiology; in combination with a projection argument we develop an efficient and easily computed testing scheme.

e-mail: amaity@stat.tamu.edu

LOCAL POST-STRATIFICATION AND DIAGNOSTICS IN DUAL SYSTEM ACCURACY AND COVERAGE EVALUATION FOR US CENSUS

Chengyong Tang*, Iowa State University
Song X. Chen, Iowa State University

This paper proposes a local post-stratification approach to the dual system estimation for evaluating coverage in a Census. The local stratification is achieved by nonparametric estimators for the Census enumeration and correct enumeration probabilities, which may be employed as an alternative to the existing post-stratification in the Census dual system estimation. We propose an imputation based estimator for general nonparametric regression with missing values, which can handle both continuous and categorical covariates and is used for model diagnostic and checking. Our theoretical analysis and simulation studies indicate attractive properties of the imputation based estimator. The proposed approach is applied to analyze a 20% research data file from the 2000 Census dual system surveys.

e-mail: yongtang@iastate.edu

ANALYZING FORCED UNFOLDING OF PROTEIN TANDEMS VIA ORDER STATISTICS

Efstathia Bura*, George Washington University
Dmitri Klimov, George Mason University
Valeri Barsegov, University of Massachusetts-Lowell

Mechanically active proteins are typically organized as homogeneous or heterogeneous tandems of protein domains. A large number of proteins perform important biological functions in their unfolded state. In current force-clamp atomic force microscopy (AFM), mechanical unfolding of protein tandems is studied by using constant stretching force and recording the unfolding transitions of individual domains. The main goals of these experiments are (a) to obtain the distributions of unfolding times for individual domains and (b) to probe interdomain interactions. Existing statistical methodology offers limited information gain as it ignores that the observable quantities are the ordered forced unfolding times. Order statistics based methodology will be presented for analyzing the unfolding times of protein tandems and to infer the parent unfolding time distributions of individual domains from ordered unfolding times. Statistical tests for independence of the unfolding times and equality of their (parent) distributions, applied to ordered data, will be presented. The proposed tests will enable experimentalists to detect presence of interdomain interaction.

e-mail: ebura@gwu.edu

ON DISTRIBUTION-FREE RUNS TEST FOR SYMMETRY USING MEDIAN RANKED SET SAMPLING

Hani M. Samawi*, Georgia Southern University-Jiann-Ping Hsu College of Public Health

Optimal ranked set sample design for distribution-free runs test of symmetry is introduced. Our investigation reviled that the optimal ranked set sample designs for runs test of symmetry are those with quantifying order statistics with labels when the set size r is odd and when the set size r is even. The exact null distribution is provided. Numerical analysis of the power of the proposed optimal designs is included. Illustration using real data is used.

e-mail: hsamawi@georgiasouthern.edu

BETA REGRESSION MODEL FOR THE AREA UNDER THE ROC CURVE

Lin Zhang*, Baylor University
Jack Tubbs, Baylor University

In this paper, we propose to use beta regression to model the nonparametric Area under the ROC Curve. The response variable in the model is the placement value which is the survivor function of the non-disease group for the disease group sample point. The model uses the relationship between the ROC and placement value. It allows for the use of readily available software for beta regression models and the inclusion of covariates. Because the response variables are not independent, the bootstrap method is used to make inference about the parameter of the model. The approach is demonstrated by the simulation studies and real data.

e-mail: linzhg78@yahoo.com

DENSITY ESTIMATION IN PHASE II STUDIES IN LARGE SCALE COMMUNITY-BASED RESEARCH

Wan Tang*, University of Rochester
Hua He, University of Rochester
Xin Tu, University of Rochester

It is of great interest to study the distribution of some characteristics in a population of interest in large scale community-based research. If a random sample is obtained from some subpopulation and if there is no missing data, kernel smoothing methods may be used to estimate the desired density. However, it is often the case that the membership of subject is unknown in advance, making it impossible to sample from the subpopulation directly. Instead, we may sample the whole population and use some gold standard test to obtain the membership of the sampled subject. The problem created is that the gold standard may be too costly or invasive to administer. However, it is often possible to choose a subset of the sampled subjects based on observed other characteristics to ascertain their membership. In such phase II studies, the membership is missing for the subject without the gold standard test. In this talk, we discuss a new kernel smoothing approach to estimate the density in such situations.

e-mail: wan_tang@urmc.rochester.edu

Index

Reference is to Session Number

Index

Index

Index

Notes

Books and Journals From

Wiley-Blackwell

Biometrics
JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY
VOLUME 63 • NUMBER 4
DECEMBER 2007
Impact Factor 1.489*
An International Society Devoted to the Development and Application of Statistical and Mathematical Theory and Methods in the Biosciences

*2006 Thomson ISI Ranking

Stop by the booth for your
FREE SAMPLE COPIES
of this and other
leading journals in the field!

Wiley-Blackwell is proud to publish

Biometrics

on behalf of the International Biometric Society (IBS)

I B S

Co-Editors

GEERT MOLENBERGHS, *Universiteit Hasselt, Belgium*

NAISYIN WANG, *Texas A&M University, USA*

DAVID ZUCKER, *Hebrew University of Jerusalem, Israel*

Executive Editor: **MARIE DAVIDIAN**, *North Carolina State University, USA*

Editorial Manager: **ANN HANHART**, *University of Texas at Arlington, USA*

Biometrics emphasizes the role of statistics and mathematics in the biosciences. Its objectives are to promote and extend the use of statistical and mathematical methods in the principal disciplines of biosciences by reporting on the development and application of these methods. A centerpiece of most Biometrics articles is scientific application that sets scientific or policy objectives, motivates methods development, and demonstrates the operations of new methods.

The International Biometric Society (IBS) is an international society devoted to development and application of statistical and mathematical theory and methods in the biosciences. The Society publishes two journals: ***Biometrics,*** and, jointly with the American Statistical Association, the *Journal of Agricultural, Biological, and Environmental Statistics* (JABES).

The IBS is the parent society of ENAR, the Eastern North American Region of IBS.

Discover more about the Wiley-Blackwell Biostatistics Journals Portfolio at

www.interscience.wiley.com/topics/statistics

and sign up for your FREE E-Content Alerts!

Introduction to Time Series and Forecasting

DOUGLAS C. MONTGOMERY,
CHERYL L. JENNINGS, MURAT KULAHCI

Mar 2008 • 464 pp • Cloth • $115.00
978-0-471-65397-4

Applied Survival Analysis

Regression Modeling of Time to Event Data

SECOND EDITION

DAVID W. HOSMER, STANLEY LEMESHOW, SUSANNE MAY

Feb 2008 • 420 pp • Cloth • $110.00
978-0-471-75499-2

Statistical Advances in the Biomedical Sciences

Clinical Trials, Epidemiology, Survival Analysis and Bioinformatics

ATANU BISWAS, SUJAY DATTA,
JASON P. FINE, MARK R. SEGAL

Dec 2007 • 582 pp • Cloth • $120.00
978-0-471-94753-0

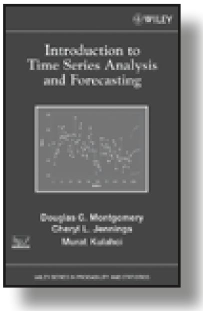
Statistical Issues in Drug Development

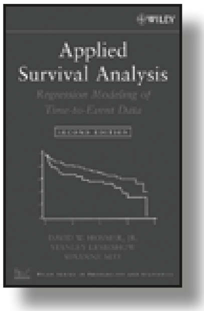
SECOND EDITION

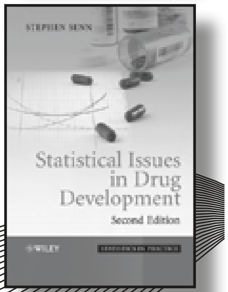
STEPHEN SENN

Feb 2008 • 524 pp • Cloth • $130.00
978-0-470-01877-4

SAVE 20% OFF BOOKS ON DISPLAY




WILEY
Applied Survival Analysis
Regression Modeling of Time-to-Event Data
SECOND EDITION



WILEY-BLACKWELL

For more titles visit our booth at ENAR, or browse & order online at **www.wiley.com/statistics.**
To order call **1-877-762-2974** *in North America* or at + 44 (0) 1243 843294 *in Rest of World*.
TO SAVE 20%: order at ENAR or use **PROMOTION CODE 87337** at checkout when you order online.