



# 2009 SPRING MEETING WITH IMS AND SECTIONS OF ASA STATISTICS: MAPPING DATA INTO DISCOVERY

MARCH 15-18, 2009 | GRAND HYATT SAN ANTONIO | SAN ANTONIO, TEXAS



INTERNATIONAL BIOMETRICAL SOCIETY  
EASTERN NORTH AMERICAN REGION  
**PROGRAM & ABSTRACTS**

# CONTENTS

<b>Section</b>	<b>Page</b>
Acknowledgements	3
Officials and Committees	4
Representatives	4
Student Awards	5
Special Thanks	6
Presidential Invited Speaker	7
IMS Medallion Lecturer	7
Short Courses	8
Tutorials	11
Roundtables	12
Program Summary	14
Scientific Program	19
Abstracts	49
Index	194
Floor Plans	208
Notes	210



# ACKNOWLEDGEMENTS

ENAR would like to acknowledge the generous support of the 2009 Local Arrangements Committee, chaired by Chen-Pin Wang, the Veterans Evidence-based Research, Dissemination, & ImplementationCenter, South Texas Veterans Health Care System, the Department of Epidemiology and Biostatistics, University Texas Health Science Center at San Antonio, the Department of Management Science and Statistics, UT San Antonio, and our student volunteers.

ENAR is grateful for the support of the National Institutes of Health (National Cancer Institute, National Heart, Lung, and Blood Institute, and National Institute of Environmental Health Sciences) and of the ENAR Junior Researchers' Workshop Coalition (Columbia University, Harvard University, The Johns Hopkins University, North Carolina State University, The Ohio State University, The University of Michigan, The University of North Carolina at Chapel Hill, and The University of Wisconsin).

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

## Sponsors

Abbott Laboratories

Amgen, Inc.

AstraZeneca/MedImmune

Biogen Idec

Boehringer Ingelheim Pharmaceuticals

Cephalon, Inc.

Eli Lilly and Company

GlaxoSmithKline

Novartis Pharmaceuticals Inc.

Pfizer, Inc.

PPD, Inc.

Proctor & Gamble

ReSearch Pharmaceutical Services, Inc.

Rho, Inc.

sanofi-aventis

SAS

Schering-Plough

Smith Hanley Associates LLC

Statistics Collaborative, Inc.

Wyeth Research

## Exhibitors

Amgen, Inc.

ASA – SIAM Series

The Cambridge Group Ltd.

Cambridge University Press

CRC Press – Taylor & Francis Group LLC

Cytel Inc.

Food and Drug Administration –  
Center for Devices and Radiological Health

Kforce Clinical Research

Oxford University Press

RPS, Inc.

SAS

SAS Institute

SAS Institute Inc. – JMP Division

Smith Hanley Associates LLC

Springer

StataCorp LP

TIBCO Software Inc.

Wiley-Blackwell

# ACKNOWLEDGEMENTS

## OFFICERS AND COMMITTEES

January – December 2009

### EXECUTIVE COMMITTEE -- OFFICERS

President	Lance Waller
Past President	Eric (Rocky) Feuer
President-Elect	Sharon-Lise Normand
Secretary (2009-2010)	Maura Stokes
Treasurer (2008-2009)	Scarlett Bellamy

### REGIONAL COMMITTEE (RECOM)

President (Chair)	Lance Waller
Eight ordinary members (elected to 3-year terms): + Amy Herring (RAB Chair)	

#### 2007-2009

Karen Bandeen-Roche  
F. Dubois Bowman  
Paul Rathouz

#### 2008-2010

Jianwen Cai  
Bradley Carlin  
Peter Macdonald

#### 2009-2011

Daniel Heitajn  
José Pinheiro  
Joanna Shih

### REGIONAL MEMBERS OF THE COUNCIL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Marie Davidian, Susan Ellenberg, Louise Ryan, Timothy G. Gregoire, Roderick Little

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

**Chair: Amy Herring**  
**Co-Chair: Hormuzd Katki**

#### 2007-2009

Christopher S. Coffey  
Hormuzd A. Katki  
Lan Kong  
Yi Li  
Lillian Lin  
Laura Meyerson  
Gene Pennello  
Tamara Pinkett  
John Preisser  
Douglas E. Schaubel

#### 2008-2010

Karla V. Ballman  
Craig Borkowf  
Avital Cnaan  
Kimberly Drews  
Matthew Gurka  
Monica Jackson  
Robert Johnson  
Robert Lyles  
Peter Song  
Ram Tawari

#### 2009-2011

Dipankar Bandyopadhyay  
Andrew Finley  
Haoda Fu  
Ronald Gangnon  
Eugene Huang  
Renée Moore  
Roger Peng  
Jennifer Schumi  
Brian Smith

### PROGRAMS

#### 2009 Spring Meeting – San Antonio, TX

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

#### 2010 Spring Meeting – New Orleans, LA

Program Chair: Michael Daniels  
Program Co-Chair: Jeffrey Morris  
Local Arrangements Chairs: Brian Marx and Julia Volaufova

#### 2009 Joint Statistical Meeting Lloyd Edwards

#### 2010 Joint Statistical Meeting Yulei He

### ENAR REGIONAL MEMBERS OF THE COUNCIL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Marie Davidian, Susan Ellenberg, Louise Ryan, Timothy G. Gregoire, Roderick Little, Linda Young

#### Biometrics Editor

#### Biometrics Co-Editors

#### Biometric Bulletin Editor

#### JABES Editor

ENAR Correspondent  
for the Biometric Bulletin

ENAR Executive Director

International Biometric  
Society Business Manager

#### Marie Davidian

Geert Molenberghs, Thomas A. Louis,  
and David Zucker

Roslyn Stone

Carl Schwarz

Lillian Lin

Kathy Hoskins

Dee Ann Walker

### REPRESENTATIVES

#### COMMITTEE OF PRESIDENTS OF STATISTICAL SOCIETIES (COPSS)

#### ENAR Representatives

Lance Waller (President)     Eric Feuer (Past-President)  
Sharon-Lise Normand (President-Elect)

#### ENAR Standing and Continuing Committees

##### Nominations Committee

Lisa LaVange, Chair (2008-2009)  
Joan Chmiel (2007-2008)  
Joel Greenhouse (2007-2008)  
Jane Pendergast (2007 Chair; 2008)  
Karen Bandeen-Roche (2008-2009)  
Elizabeth Margosches (2008-2009)

##### Sponsorship Committee

Christine Clark, Chair  
Thomas Kelleher  
Kannan Natarajan

#### ENAR Representative on the ASA Committee on Meetings

Maura Stokes, Committee Vice-Chair  
(January 2008-December 2010)

#### AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE (Joint with WNAR)

Terms were through February 22, 2011

Section E, Geology and Geography

Carol Gotway Crawford

Section N, Medical Sciences

Judy Bebbuk

Section G, Biological Sciences

Geof Givens

Section U, Statistics

Mary Foulkes

Section O, Agriculture

Mary Christman

#### NATIONAL INSTITUTE OF STATISTICAL SCIENCES

(ENAR President is also an ex-officio member) Board of Trustees

Members:

Lance Waller, ENAR President

Donna Brogan



# ACKNOWLEDGEMENTS

## Workshop for Junior Researchers

Amy Herring (Chair), University of North Carolina at Chapel Hill  
DuBois Bowman, Emory University  
Brent Coull, Harvard University  
Marie Davidian, North Carolina State University  
Francesca Dominici, The Johns Hopkins University  
Yi Li, Harvard University  
Xihong Lin, Harvard University  
Doug Schaubel, The University of Michigan  
Lance Waller, Emory University

## 2009 Fostering Diversity in Biostatistics Workshop

Reneé H. Moore (co-chair)  
Adriana Perez (co-chair)  
Mourad Atlas  
Scarlett Bellamy  
DuBois Bowman  
Amita Manatunga  
Sastry Pantula  
Dionne Price  
DeJuran Richardson  
Louise Ryan  
Nagambal Shah  
Keith Soper  
Lance Waller

## Distinguished Student Paper Awards Committee

Lisa LaVange, Ph.D. (Chair 2008)  
Sudipto Banerjee, Ph.D. (2006-2008)  
Christopher R. Bilder, Ph.D. (2007-2009)  
Nilanjan Chatterjee, Ph.D. (2008-2010)  
Elizabeth S. Garrett-Mayer, Ph.D. (2006-2008)  
Debashis Ghosh, Ph.D. (2007-2009)  
Liang Li, Ph.D. (2006-2008)  
Yi Li, Ph.D. (2007-2009)  
Robert Lyles, Ph.D. (2008-2010)  
Laura J Meyerson, Ph.D. (2008-2010)  
Jane Pendergast, Ph.D. (2008 substitute)  
Heping Zhang, Ph.D. (2007-2009)

## DISTINGUISHED STUDENT PAPER AWARD WINNERS

### Van Ryzin Award Winner

Gordana Derado, Emory University

### Award Winners

Hongyuan Cao, University of North Carolina at Chapel Hill  
Howard Chang, Johns Hopkins Bloomberg School of Public Health  
Ping Chen, University of Missouri- Columbia  
Yu-Jen Cheng, Johns Hopkins Bloomberg School of Public Health  
Yeonseung Chung, Harvard School of Public Health  
Ramon Garcia, University of North Carolina at Durham  
Jian Guo, University of Michigan  
Lei Hua, University of Iowa  
Yimei Li, University of North Carolina at Chapel Hill  
Ziyue Liu, University of Pennsylvania School of Medicine  
Jessica Myers, Johns Hopkins Bloomberg School of Public Health  
Amy Nowacki, Cleveland Clinic Foundation  
Jane Paik, Columbia University  
Nicholas Reich, Johns Hopkins School of Public Health  
Ali Shojaie, University of Michigan  
Peng Wei, University of Minnesota  
Yiyun Zhang, Penn State University  
Bingqing Zhou, University of North Carolina at Chapel Hill  
Bin Zhu, University of Michigan

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



# 2009 SPECIAL THANKS



## 2009 ENAR Program Committee

Brent Coull (Chair), *Harvard University*  
Mahlet Tadesse (Co-Chair), *Georgetown University*  
Jared Christensen, *Wyeth Pharmaceutical*  
Erin Conlon, *University of Massachusetts at Amherst*  
Abi Ekangaki, *Eli Lilly and Company*  
Jonathan French, *Pfizer*  
Misrak Gezmu, *National Institute of Allergy and Infectious Diseases*  
Joseph Hogan, *Brown University*  
Andrew Lawson, *Medical University of South Carolina*  
Loni Philip (student assistant), *Harvard University*  
Dionne Price, *Food & Drug Administration*  
Lance Waller, *Emory University*

## ASA Section Representatives

Patrick G. Arbogast (Section on Teaching Statistics in the Health Sciences), *Vanderbilt University*  
Karla Ballman (Section on Statistical Education), *Mayo Clinic*  
Jinbo Chen (Biometrics Section), *University of Pennsylvania*  
William Davis (Survey Research and Methodology Section),  
*National Cancer Institute*  
Bonnie Ghosh-Dastidar (Health Policy Statistics Section),  
*Rand Corporation*  
Myron Katzoff (Section on Statistics in Defense and National Security), *National Center for Health Statistics*  
Steve Rathbun (Section on Statistics and the Environment), *University of Georgia*  
Matilde Sanchez (Biopharmaceutical Section),  
*Arena Pharmaceuticals*  
Elizabeth R. Zell (Section on Statistics in Epidemiology),  
*Centers for Disease Control & Prevention*

## IMS Program Chair

Tianxi Cai, *Harvard University*

## ENAR Education Advisory Committee

Greg Campbell, *Food and Drug Administration*  
Marie Davidian, *North Carolina State University*  
Jose Pinheiro, *Novartis Pharmaceuticals*

## Local Arrangements Committee

Chen-Pin Wang (Chair), *University of Texas Health Science Center*

## ENAR Student Awards Committee

Lisa LaVange (Chair), *University of North Carolina at Chapel Hill*

## ENAR Diversity Workshop Committee

Renee' Moore (Co-Chair), *University of Pennsylvania*  
Adriana Perez (Co-Chair), *University of Louisville*

## ENAR Workshop for Junior Researchers Committee

Amy Herring (Chair), *University of North Carolina at Chapel Hill*



# ENAR PRESIDENTIAL INVITED SPEAKER

## Statistical Modelling for Real-time Epidemiology

Professor Peter J. Diggle

Lancaster University School of Health and Medicine  
and Johns Hopkins University School of Public Health



Large volumes of data on a range of health outcomes are now collected routinely by many health care organisations but, at least in the UK, are often not analysed other than for retrospective audit purposes. Each data-record will typically be referenced both in time and in space; for example, in the UK the temporal reference will be a date, and in some cases a time of day, whilst the spatial reference will usually

be the individual's post-code which, in urban settings, corresponds to a spatial resolution of the order of 100 metres. By real-time epidemiology, I mean the analysis of data-sets of this kind as they accrue, to inform clinical or public health decision-making. Such analyses would be triggered and the results posted automatically, for example on a web-site, by the arrival of new data.

In this talk I will review work in spatial, temporal and spatio-temporal modelling that seems especially relevant to this general task, and will describe a number of applications, including some or all of:

- real-time syndromic surveillance (Diggle, Rowlingson and Su, 2005);
- tropical disease prevalence mapping (Crainiceanu, Diggle and Rowlingson, 2008);
- early warning of incipient renal failure in primary care patients (Diggle and Sousa, 2009).

CRAINICEANU, C., DIGGLE, P.J. and ROWLINGSON, B.S. (2008). Bivariate modelling and prediction of spatial variation in Loa loa prevalence in tropical Africa (with Discussion). *Journal of the American Statistical Association*, 103, 21–43.

DIGGLE, P.J., ROWLINGSON, B. and SU, T-L. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16, 423–34.

DIGGLE, P.J. and SOUSA, I. (2009). Real-time detection of incipient renal failure in primary care patients using a dynamic time series model. (in preparation)

### Biography

Peter J. Diggle is Professor at the Lancaster University School of Health and Medicine with an adjunct appointment in the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. The development of innovative statistical methodology motivated by answering specific research questions within complicated data forms the central focus of Professor Diggle's scholarly accomplishments. His methodological research stems from close collaborations with scientists from a variety of fields ranging from ecology to microbiology to global health and provides focused and accurate inference from multiple sources of observational data containing complex temporal and spatial correlation structures. Professor Diggle is an advocate for the critical role of statistical thinking and methodology within modern science. He was an early pioneer in the use of Monte Carlo methods for statistical inference and his work provides creative solutions in the analysis of complex

data in order to address the underlying questions of primary scientific interest. He is a Fellow of the American Statistical Association, and serves as a member of the Engineering and Physical Sciences Research Council (UK) and the Medical Research Council (UK) College of Experts, as well as a Trustee for Biometrika. His research has been recognized by the statistical community with five papers read and discussed before the Royal Statistical Society, the receipt of the Guy Medal in Silver from the Royal Statistical Society in 1997, and recognition and discussion of his recent collaborative work on spatio-temporal patterns of Loa loa infection by the Journal of the American Statistical Association. Professor Diggle is a founding co-editor of the journal *Biostatistics* and a prolific author with over 160 published articles and eight books on topics such as time series, longitudinal data analysis, spatial point patterns, and model-based geostatistics.

# IMS MEDALLION LECTURER

## Statistical Challenges in Nanoscale Biophysics

Professor Samuel Kou

Department of Statistics  
Harvard University



Recent advances in nanotechnology allow scientists to follow a biological process on a single molecule basis. These advances also raise many challenging stochastic modeling and statistical inference problems. First, by zooming in on single molecules, recent nanoscale experiments reveal that many classical stochastic models derived from oversimplified assumptions are no longer valid. Second, the stochastic nature of the experimental data and the presence of latent processes significantly complicate the statistical inference. In this talk we will use the modeling of enzymatic reaction and the inference of biochemical kinetics to illustrate the statistical and probabilistic challenges in single-molecule biophysics.

### Biography

Samuel Kou is Professor of Statistics at Harvard University. He received a bachelor's degree in computational mathematics from Peking University in 1997, followed by a Ph.D. in statistics from Stanford University in 2001 under the supervision of Professor Bradley Efron. After completing his Ph.D. studies, he joined Harvard University as an Assistant Professor. He has held a visiting position at the University of Chicago. Dr. Kou is the recipient of the National Science Foundation CAREER Award and the Institute of Mathematical Statistics' Richard Tweedie Award. He is an elected Fellow of the American Statistical Association. He has served as the Associate Editor of the *Journal of Multivariate Analysis*, and currently serves as Editor of the *Chance* magazine and as Associate Editor of the *Annals of Applied Statistics* and *Statistica Sinica*.

# SHORT COURSES

**Date: Sunday, March 15, 2009**

**Full Day Fee:**

Members \$275

Nonmembers \$320

**Half Day Fee:**

Members \$185

Nonmembers \$225

## SC3: Hierarchical Modeling and Analysis of Spatial-Temporal Data: Emphasis in Forestry, Ecology, and Environmental Sciences

**Room: Texas Ballroom D, 4th Floor**

*Full Day 8:00 am-5:00 pm*

*Instructors: Andrew O. Finley, Department of Forestry and Geography, Michigan State University; and Sudipto Banerjee, Division of Biostatistics, School of Public Health, University of Minnesota*

### SC1: Bioconductor for the Analysis of Genome-Scale Data

**Room: Texas Ballroom A/B, 4th Floor**

*Full Day 8:00 am-5:00 pm*

*Instructor: Vincent Carey, Harvard University*

**Description:** Bioconductor ([www.bioconductor.org](http://www.bioconductor.org)) is a collection of workflow components, created for the R environment for statistical computing, that facilitate analysis of genome-scale data structures including expression microarrays, SNP chips, and high-throughput sequencing assays. This course will cover a selection of topics related to data capture, quality assessment, inference, machine learning, and sequence analysis. Basic approaches to software development and reproducible research methods will also be presented.

**Prerequisites:** Familiarity with R and general knowledge of microarray analysis. Participants are expected to bring reasonably modern laptops (at least 1GB RAM) with R and associated libraries installed prior to the course. Software and data will be provided prior to the meeting to illustrate all topics in hands-on lab sessions.

*NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website ([www.enar.org](http://www.enar.org)) for instructions on download prior to the meeting.*

### SC2: Regression Modeling Strategies

**Room: Texas Ballroom C, 4th Floor**

*Full Day 8:00 am-5:00 pm*

*Instructor: Frank E Harrell Jr., Vanderbilt University*

**Description:** Regression models are frequently used for hypothesis testing, estimation, and prediction in a multitude of areas. Models must be flexible enough to fit nonlinear and non-additive relationships while avoiding overfitting and the resulting failure of the model to accurately predict new observations. This shortcourse overviews elements of the presenter's book (Regression Modeling Strategies, New York: Springer; 2001). Some of the topics covered include using regression splines to relax linearity assumptions, perils of variable selection and overfitting, where to spend degrees of freedom, shrinkage, imputation of missing data, data reduction, and interaction surfaces. A default overall modeling strategy will be described. This is followed by methods for graphically understanding models (e.g., using nomograms) and using resampling to estimate a model's likely performance on new data. Two case studies will be presented. The methods covered in this tutorial apply to almost any regression model, including ordinary least squares, binary and ordinal logistic regression models, and survival models.

**Prerequisites:** A good command of ordinary multiple regression along the lines of Draper and Smith.

**Description:** Recent advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. This has encouraged collection of spatial-temporal datasets in many fields and has generated considerable interest in statistical modelling for time and location-referenced data. This accumulation of data and need for analysis is especially common in the broad fields of forestry and ecology. In these fields, spatially and temporally indexed data, typically consisting of one or more response variables and associated covariates, is used to estimate natural resource inventory, presence/absence, counts, and change. In these settings, the focus of inference is often on specific model parameters and/or subsequent prediction at a new location or time. In these modelling exercises, rarely is it safe, or even desirable, to assume that model residuals are independent and identically distributed. The propensity to violate these assumptions is especially great in environmental datasets because the data often exhibit temporal, spatial, or hierarchical structure, or all three.

This course details recent advancements in hierarchical random effects models using Markov chain Monte Carlo (MCMC) methods. The course focus is on linear and generalized linear modelling frameworks that accommodate spatial and temporal associations. Careful attention is paid to the theoretical foundations of model specification, identifiability of parameters, and computational considerations for Bayesian inference from posterior distributions. The lecture will start with a basic introduction to Bayesian hierarchical linear models and proceed to address several common challenges in environmental data, including missing data and when the number of observations is too large to efficiently fit the desired hierarchical random effects models. Diverse settings for spatial and spatial-temporal models are considered, mostly motivated by a range of studies that employ forestry and ecological monitoring datasets. The course will blend modelling, computing, and data analysis including a hands-on introduction to the R statistical environment. Special attention is given to exploration and visualization of spatial-temporal data and the practical and accessible implementation of spatial-temporal models. In particular, participants will learn how to fit a diverse class of spatial-temporal models using the spBayes R package.

**Prerequisites:** Some familiarity with classical linear models and multiple regression will be useful. A laptop with a current version of R and spBayes installed, while not required, will definitely be useful. Please visit the short course website a few weeks prior to the course for software updates and R scripts and data that will be used for illustration.

*NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website ([www.enar.org](http://www.enar.org)) for instructions on download prior to the meeting.*



## **SC4: Estimating Curves and Surfaces from Environmental Data**

**Room:** Texas Ballroom E, 4th Floor

*Full Day 8:00 am-5:00 pm*

*Instructors: Reinhard Furrer, Colorado School of Mines; Doug Nychka and Stephan Sain, National Center for Atmospheric Research*

**Description:** Determining the air quality at an unmonitored location, characterizing the mean summer temperature and precipitation over a spatial domain or relating soil properties to bulk composition are examples where a function of interest depends on irregular and limited observations. Prediction and scientific understanding of environmental data often require estimating a smooth curve or surface that describes an environmental process or summarizes complex structure. Moreover, drawing inferences from this estimate requires measures of uncertainty for the unknown function. This course will combine ideas from geostatistics, smoothing, and Bayesian inference to tackle these problems. An important component of the lectures is the use of the fields and spam contributed packages for the R statistical computing environment for hands-on experience with these methods. In addition, these R packages provide insight to the computational framework for function fitting and the facility to handle multivariate or large environmental datasets.

The first part of the course explains a common framework for spatial statistics and splines using ridge regression. This correspondence provides the common computational approach used throughout fields and leads to easy-to-use methods for Kriging and thin-plate splines. Several case studies illustrate how these methods work in practice and the class is encouraged to modify the R code to explore variations in the analysis. The second part of the course considers multivariate responses and large spatial data sets. Building from the basic methods, these lectures extend the fields functions either through multivariate covariance functions or sparse matrix methods.

**Prerequisites:** Familiarity with statistical linear models, multivariate regression and matrix algebra. Data and examples will be available for download prior to the meeting.

*NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website ([www.enar.org](http://www.enar.org)) for instructions on download prior to the meeting.*

## **SC5: Statistical Modeling and Analysis of Brain Imaging Data**

**Room:** Texas Ballroom F, 4th Floor

*Half Day Shortcourse 8:00 am-Noon*

*Instructors: F. DuBois Bowman and Ying Guo, Department of Biostatistics and Bioinformatics, Center for Biomedical Imaging Statistics (CBIS), Rollins School of Public Health of Emory University*

**Description:** Functional neuroimaging technology has played a central role in improving our understanding of normal brain function in humans and in investigating the neural basis for major psychiatric, neurologic, and substance abuse disorders. Imaging modalities, such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), capture hundreds of thousands of spatially localized measurements of brain activity at a particular point in time, typically while the subject in the scanner performs some experimental task. Thus, acquiring serial scans on a subject yields a 3-D movie of task-related brain function, encompassing both spatial and temporal correlations. Common approaches for analyzing brain imaging data target 1) functional specialization to determine localized brain regions that are associated with a given task, e.g. related to drug craving, and 2) functional integration to reveal patterns of associations between specialized brain regions.

The lecture will start with a brief introduction to some basic theories in neurophysiology that are important for understanding what neuroimaging data actually represent. The course will then proceed to discuss some basic principles of data acquisition and preprocessing steps for fMRI and PET and highlight characteristics of the data that are particularly relevant to statistical modeling. The focus of the course will be on methods for statistical analyses. Specifically, we will target statistical methods for 1) activation studies (addressing functional specialization), 2) functional connectivity (addressing functional integration), and 3) prediction. Throughout the course, we will make references to popular software for implementing many of the statistical techniques covered.

**Prerequisites:** This course is targeted to statisticians and biological scientists who are at or beyond masters-level training in statistics and who are interested in analyzing brain imaging data or developing statistical methodology for the design and analysis of neuroimaging data.

# SHORT COURSES



## SC6: Intermediate Bayesian Data Analysis Using WinBUGS and BRugs

**Room:** Texas Ballroom F, 4th Floor

*Half Day Shortcourse 1:00 pm-5:00 pm*

*Instructor: Brad Carlin, University of Minnesota*

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

**Prerequisites:** The presentation will assume familiarity with basic Bayesian methods and MCMC algorithms, at the level of, say, Chapters 2 and 3 of Carlin and Louis (2009) or Chapters 2, 3, 5 and 11 of Gelman et al. (2004). The course's goal is to make these methods come alive in the software through real data examples that the students try for themselves during the presentation. All necessary WinBUGS and BRugs code will be made available on the web, and experienced teaching assistants will also be on hand to assist those who become "stuck" for any reason.

Students will be expected to bring their own laptop computers to the session, and to have the latest versions of WinBUGS and R already installed on these computers. Both of these programs are freely available from <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml> and <http://www.r-project.org/> respectively. You may wish to arrive in the short course room 30 minutes early in order to make sure you have the correct versions of the software installed.

*NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website ([www.enar.org](http://www.enar.org)) for instructions on download prior to the meeting.*

# TUTORIALS

## T1: Methods for Reproducible Research

**Room:** Texas Ballroom F, 4th Floor

*Monday 8:30-10:15 am*

*Instructor: Roger Peng, Johns Hopkins University*

**Description:** The validity of conclusions from scientific investigations is typically strengthened by the replication of results by independent researchers. Full replication of a study's results using independent methods, data, equipment, and protocols, has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study, there are examples of scientific investigations which cannot be fully replicated, often because of a lack of time or resources. For example, epidemiologic studies which examine large populations and can potentially impact broad policy or regulatory decisions, often cannot be fully replicated in the time frame necessary for making a specific decision. In such situations, there is a need for a minimum standard which can serve as an intermediate step between full replication and nothing. This minimum standard is reproducible research, which requires that datasets and computer code be made available to others for verifying published results and conducting alternate analyses. The tutorial will provide an overview of methods for conducting reproducible research. We will focus on the R statistical computing language and will discuss other tools that can be used for producing reproducible documents. Topics that will be discussed include Sweave, literate programming, caching of large computations, and distributing reproducible research over the Web.

**Prerequisites:** Knowledge of how to use R and some familiarity with the LaTeX typesetting system.

*NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website ([www.enar.org](http://www.enar.org)) for instructions on download prior to the meeting.*

## T2: Mass Spectrometry Based Proteomics

**Room:** Texas Ballroom F, 4th Floor

*Monday 10:30 am-12:15 pm*

*Instructor: Mahlet G. Tadesse, Georgetown University*

**Description:** Mass spectrometry (MS) has become the leading technology used in proteomics. In particular, it is increasingly applied to profile complex protein mixtures and identify peptides/proteins. This tutorial will review the principles of MS, its application to proteomics, and the statistical challenges associated with analyzing MS-based proteomic data. Emphasis will be placed on the basics of the major MS technologies: matrix assisted and surface-enhanced laser desorption/ionisation time-of-flight MS (MALDI-TOF and SELDI-TOF), electrospray ionisation (ESI), tandem MS (MS/MS), and liquid chromatography coupled with mass spectrometry (LC-MS). Key statistical and computational issues related to the analysis of these data will be discussed. The topics to be covered are: (1) experimental design issues, including sources of variation and reproducibility of MS data; (2) data pre-processing steps, such as baseline correction, noise filtering, normalization, and peak detection; (3) high-level analysis steps, including sample classification, biomarker identification, and related inferential issues, such as dealing with multiple testing. Currently available methods to address these problems will be presented, and their merits and limitations will be discussed.

**Prerequisites:** General understanding of statistical principles at the level of a first year graduate student in (bio)statistics.

## T3: Genetic and Microarray Data Analysis

**Room:** Texas Ballroom F, 4th Floor

*Monday 1:45-3:30 pm*

*Instructors: Russ Wolfinger and Kelci Miclaus, SAS Institute Inc.*

**Description:** This tutorial targets biostatisticians who wish to increase their understanding of statistical issues involved with genomics data analysis. After reviewing some primary molecular biology vocabulary and concepts, we will divide our time roughly equally between analysis of genetic marker data and transcript abundance (microarray) data. The former will focus on modern genome-wide association studies and the latter on large-scale mRNA profiling methodologies. There will be a mixture of theory, applications, and practical examples emphasizing both inference and prediction as well as the critical need to blend visualization with statistical modeling.

**Prerequisites:** Course materials, including a number of illustrative JMP Genomics scripts, will be provided online. Attendees wishing to run the scripts interactively during class should download them beforehand and bring their own laptop. A link to the scripts will be made available approximately one week before class. In addition, within 30 days of class, attendees should download a free trial version of JMP from [www.jmp.com](http://www.jmp.com) .

*NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website ([www.enar.org](http://www.enar.org)) for instructions on download prior to the meeting.*



# ROUNDTABLES

**MONDAY, MARCH 16, 2009  
12:15pm-1:30 pm**

Texas Ballroom F, 4th Floor

## R1: NIH Grant Review Process for Methodology Grants

**Discussion Leader:** Charles E. McCulloch, Head, Division of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco; incoming Chair, NIH BMRD Study Section

**Description:** To facilitate the discussion, I will begin with an outline of the grant review process and a description and comparison of the usual NIH methodology grant mechanisms, namely R01, R03 and R21 grants. I will describe the characteristics of grants that lend themselves to more (or less!) favorable review, including the all-important formulation of specific aims. Other possible topics for discussion include ways in which junior researchers can successfully compete and strategies for resubmission of grants.

## R2: A Survival Kit for the Scientific Publication Jungle

**Discussion Leader:** Geert Molenberghs, I-BioStat, Hasselt University & Catholic University Leuven, Belgium

**Description:** Scholarly papers constitute a time-honored communication channel within the scientific community. Contemporarily, academic, governmental, and granting authorities place increasing importance on a researcher's scientific output, often measured via the impact factor, the citation index, the Hirsch factor, etc.

Against this background, how should the junior and the senior scholar proceed to effectively communicate via scientific papers? How should one maximize chances of acceptance? What is the optimal journal? How should one ensure a published paper draws attention, is read, and gets cited? Are there specific issues for the non-native English speaker? Should one bother about collaborative papers or rather focus exclusively on methodological manuscripts?

These and other questions will be dealt upon in this roundtable, from the perspective of the editor, the associate editor, the referee, and the author. Scholarly and career-strategic considerations will be weighed.

## R3: Surviving the Tenure Track: Navigating an Academic Career in Statistics/Biostatistics

**Discussion Leader:** Marie Davidian, Department of Statistics, North Carolina State University

**Description:** An academic career in statistics/biostatistics offers the unique opportunity for engaging in a diverse mix of activities, including research, teaching, collaboration with scientists from other disciplines, and institutional and professional service. Particularly attractive aspects are relative freedom to pursue one's own research agenda, the flexibility of schedule, and the opportunity to contribute meaningfully to one's discipline and scientific advances more generally. At the same time, the varied demands of an academic career can seem daunting to a junior scholar in his or her first few years of an academic position or to graduate students contemplating the future. Indeed, expectations for publishing, for obtaining external funding, and for balancing competing responsibilities while excelling at all of them can be challenging for both the junior and senior academic alike. Junior academicians facing the "tenure horizon" can experience considerable apprehension over whether or not they will have "done enough" to satisfy often uncertain criteria; likewise, mid-level faculty continue to be concerned over their prospects for promotion.

This roundtable will be devoted to a candid discussion of the challenges of tenure-track positions in academia at all levels. Typical reappointment, promotion, and tenure processes and their variants; strategies for time management and planning one's research agenda; and tactics for "wearing many hats" successfully will be among the topics addressed.

## R4: Hierarchical Models and Uncertainty in Ecological Analysis

**Discussion Leader:** Chris Wikle, University of Missouri

**Description:** Analyses of ecological data should account for the uncertainty in the processes that generated the data. However, accounting for these uncertainties is a difficult task, since ecology is known for its complexity. Measurement and/or process errors are often the only sources of uncertainty modeled when addressing complex ecological problems, yet analyses should also account for uncertainty in sampling design, in model specification, in parameters governing the specified model, and in initial and boundary conditions. Only then can we be confident in the scientific inferences and forecasts made from an analysis. Probability and statistics provide a framework that accounts for multiple sources of uncertainty. Given the complexities of ecological studies, the hierarchical statistical model is an invaluable tool. This approach is not new in ecology, and there are many examples (both Bayesian and non-Bayesian) in the literature illustrating the benefits of this approach. In short, hierarchical statistical modeling is a powerful way of approaching ecological analysis in the presence of inevitable but quantifiable uncertainties, even if practical issues sometimes require pragmatic compromises.



## **R5: Needs and Opportunities for Establishing the ENAR Graduate Student Council (Enrollment limited to current graduate students)**

**Discussion Leader:** Xihong Lin, Harvard School of Public Health

**Description:** This roundtable will discuss needs and opportunities for establishing the ENAR Graduate Student Council (GSC). The goal of the ENAR GSC is to build an attractive platform to help graduate students develop leadership and better prepare them with skills and knowledge necessary for a successful career, interact with each other and become successful researchers in academia, government and industry. The roundtable will also discuss the possibility of having graduate students of the ENAR GSC, with input from an advisory committee, take a lead in organizing a GSC workshop affiliated with the ENAR annual meeting by identifying topics that are of most interests to graduate students, e.g., skills to develop to get ready for academia, or government or industry, dissertation paper writing and publishing, presentation skills, communication and collaboration skills. General discussions on the graduate student needs, how the ENAR GSC can help fill these needs and the successful operation of the ENAR GSC will be made.

## **R6: Statistics in Medical Imaging: Future Directions**

**Discussion Leader:** Timothy D. Johnson, University of Michigan

**Description:** Since the early 1980's statistics has played an increasingly larger role in medical imaging. A lot of work has been done in image reconstruction, registration, object recognition and denoising, among other topics. Furthermore, much work has been done on the statistical analysis of fMRI images. More recently, several new MRI techniques have been developed by physicists, including diffusion tensor imaging and MR spectroscopy where statistics has yet to play a major role. In this roundtable discussion we will cover these facets of medical imaging and the role statistics should play in the future; including current issues and problems that we currently face.

## **R7: Bayesian Spatio-Temporal Modeling of Small Area Health Data**

**Discussion Leader:** Andrew Lawson, Medical University of South Carolina

**Description:** Bayesian hierarchical modeling of georeferenced health data is often concerned with temporal changes to maps of disease. The focus can be related to estimation of the linkage to environmental covariates or simply the relative risk estimation over time. In this roundtable, issues relating to space-time modeling will be considered in relation to environmental risk gradients. In particular, the concern that space-time interaction effects could play a significant part in both description of the risk surface as it changes and the relation with covariates displaying subtle interactions. An example of this is the possible change in risk relations over time for pollution emission sources and their risk imprint. Discussion will focus on unobserved and observed time-labeled pollution events.

## **R8: Bayesian Adaptive Clinical Trials**

**Discussion Leader:** Scott Berry, Berry Consultants

**Description:** Adaptive designs have become a popular topic in the biostatistics world. I've been involved in numerous adaptive trials—all done from a Bayesian perspective. We'll create a lively discussion about adaptive trials including the philosophical approach taken, seamless phases, adaptively selecting arms (dropping and adding), adaptive randomization, adaptive sample size selections, and confirmatory studies. We can discuss all aspects of designing and implementing adaptive trials and what it takes to get from step 1 to completion.

## **R9: Diversity Initiatives in Statistics**

**Discussion Leader:** Marcia Gumpertz, North Carolina State University

**Description:** This roundtable discussion will discuss initiatives to increase diversity in undergraduate programs, graduate statistics programs, and among university faculty. As background we will briefly look at ethnic, gender and geographic diversity in science and engineering in the U.S compared with demographic trends in the U.S. as a whole. We will describe several types of programs for promoting diversity and developing an inclusive environment. Successful initiatives involve some degree of changing the culture within the institution. The roundtable will discuss the idea of culture change and some insights from the social sciences. We welcome sharing of information about successful initiatives and hope that the discussion will spark collaborations and new initiatives.

## **R10: Informative Priors and Sensitivity Analysis for Missing Data**

**Discussion Leader:** Michael Daniels, University of Florida

**Description:** Inferences from incomplete data are not possible without unverifiable assumptions. This is true for complicated missing data models as well as standard approaches such as GEE. Assumptions about the conditional distribution of missing data given observed data are often made for convenience, but cannot be empirically checked. In that sense, they can be viewed as strong priors on the missing data distribution. Several questions will be posed for discussion: Should MAR and ignorability be the 'industry standard'? Should external sources of information, either qualitative or quantitative, be incorporated into a model? What are the characteristics of an effective sensitivity analysis, and how should they be reported? Where should we focus new research in the area of incomplete data?

# PROGRAM SUMMARY

SATURDAY, MARCH 14

**9:00 a.m. – 9:00 p.m.**

*Republic B, 4th Floor*

**4:30 – 6:30 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**Workshop for Junior Researchers**

**SUNDAY, MARCH 15**

**7:30 a.m. – 6:30 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**Conference Registration**

**8:00 a.m. – 6:00 p.m.**

*Goliad Room, 2nd Floor*

**Speaker Ready Room**

**8:00 a.m. – 12:00 p.m.**

*Texas Ballroom F, 4th Floor*

**Short Courses**

SC5: Statistical Modeling and Analysis of Brain Imaging Data

**8:00 a.m.—5:00 p.m.**

*Texas Ballroom A/B, 4th Floor*

*Texas Ballroom C, 4th Floor*

*Texas Ballroom D, 4th Floor*

*Texas Ballroom E, 4th Floor*

**Short Courses**

SC1: Bioconductor for the Analysis of Genome-Scale Data

SC2: Regression Modeling Strategies

SC3: Hierarchical Modeling and Analysis of Spatial-Temporal Data: Emphasis in Forestry, Ecology, and Environmental Sciences

SC4: Estimating Curves and Surfaces from Environmental Data

**11:00 a.m. – 4:00 p.m.**

*Travis A, 3rd Floor*

**Diversity Workshop**

**1:00 p.m. – 5:00 p.m.**

*Texas Ballroom F, 4th Floor*

**Short Courses**

SC6: Intermediate Bayesian Data Analysis Using WinBUGS and BRugs

**3:00 p.m. – 5:00 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**Exhibits Open**

**4:00 p.m. – 7:00 p.m.**

*Republic A, 4th Floor*

**ENAR Executive Committee**

**4:30 p.m. – 6:30 p.m.**

*Republic B, 4th Floor*

**Placement Service**

**7:30 p.m. – 8:00 p.m.**

*Texas Ballroom, 4th Floor*

**New Member Reception**

**8:00 p.m. – 11:00 p.m.**

*Texas Ballroom, 4th Floor*

**Social Mixer and Poster Session**

1. Posters: Clinical Trials
2. Posters: Power/Sample Size
3. Posters: Microarray Analysis
4. Posters: Statistical Genetics/Genomics
5. Posters: Causal Inference
6. Posters: Imaging
7. Posters: Survival Analysis
8. Posters: Missing Data
9. Posters: Spatial/Temporal Modeling and Environmental/Ecological Applications
10. Posters: Categorical Data Analysis and Survey Research
11. Posters: Variable/Model Selection
12. Posters: Diagnostic Tests
13. Posters: Nonparametric Methods
14. Posters: Statistical Models and Methods



## MONDAY, MARCH 16

**7:30 a.m. – 8:30 a.m.**

*Texas Ballroom D, 4th Floor*

**Student Breakfast**

**7:30 a.m. – 5:00 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**Conference Registration**

**7:30 a.m. – 5:00 p.m.**

*Goliad Room, 2nd Floor*

**Speaker Ready Room**

**8:30 a.m. – 5:00 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**Exhibits Open**

**8:30 a.m. – 10:15 a.m.**

*Texas Ballroom A, 4th Floor*

**Tutorial 1: Methods for Reproducible Research**

**8:30 a.m. – 10:15 a.m.**

*Crockett C, 4th Floor*

*Crockett A/B, 4th Floor*

*Independence, 3rd Floor*

*Presidio A, 3rd Floor*

*Texas Ballroom E, 4th Floor*

*Bonham B, 3rd Floor*

*Travis A, 3rd Floor*

*Travis B, 3rd Floor*

*Travis C, 3rd Floor*

*Travis D, 3rd Floor*

*Republic A, 4th Floor*

*Texas Ballroom C, 4th Floor*

**Scientific Program**

15. Statistical Analysis of Metabolomics Data
16. Advanced Statistical Methods for Health Services Research
17. Model Specification and Uncertainty in Ecological Analyses
18. Analysis Challenges of Modern Longitudinal Biomedical Data
19. Recent Advances on Feature Selection and Its Applications
20. Contributed Papers: Analysis of Genome-wide SNP Arrays
21. Contributed Papers: Biomarkers and Diagnostic Tests
22. Contributed Papers: Causal Inference
23. Contributed Papers: An EM Approach for Partial Correlation and Missing Data
24. Contributed Papers: Power/Sample Size
25. Contributed Papers: Multivariate Survival
26. Panel Discussion: Bayesian Methods in Clinical Trials: Leveraging Industry-Academic Partnerships

**9:30 a.m. – 5:00 p.m.**

*Republic B, 4th Floor*

**Placement Service**

**10:15 a.m. – 10:30 a.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**Refreshment Break and Visit the Exhibitors**

**10:30 a.m. – 12:15 p.m.**

*Texas Ballroom A, 4th Floor*

**Tutorial 2: Mass Spectrometry Based Proteomics**

**10:30 a.m. – 12:15 p.m.**

*Texas Ballroom C, 4th Floor*

*Texas Ballroom E, 4th Floor*

*Presidio A, 3rd Floor*

*Crockett A/B, 4th Floor*

*Crockett C, 4th Floor*

*Travis A, 3rd Floor*

*Travis B, 3rd Floor*

*Bonham B, 3rd Floor*

*Travis C, 3rd Floor*

*Independence, 3rd Floor*

*Travis D, 3rd Floor*

**Scientific Program**

27. Recent Advancements in Longitudinal Analysis
28. Adaptive Designs in Practice: Benefits, Risks and Challenges
29. Outcome Dependent Sampling
30. New Statistical Methods in Detecting Epistasis Interactions in Genome-wide Association Studies
31. Analysis of Medical Cost Data: Joint Venture of Health Economists and Statisticians
32. Contributed Papers: Genetic Diversity, Mutations and Natural Selection
33. Contributed Papers: Estimation Methods
34. Contributed Papers: Spatial Models
35. Contributed Papers: Toxicology/Dose-response Models
36. Contributed Papers: Classification/Machine Learning
37. Contributed Papers: Clustered Survival Data

**12:15 p.m. – 1:30 p.m.**

*Texas Ballroom F, 4th Floor*

**Roundtable Luncheons**

# PROGRAM SUMMARY

**12:30 p.m. – 4:30 p.m.**  
*Presidio C, 3rd Floor*

## Regional Advisory Board (RAB) Luncheon Meeting

**1:45 p.m. – 3:30 p.m.**  
*Texas Ballroom A, 4th Floor*

## Tutorial 3: Genetic and Microarray Data Analysis

**1:45 p.m. – 3:30 p.m.**  
*Texas Ballroom E, 4th Floor*  
*Texas Ballroom C, 4th Floor*  
*Crockett A/B, 4th Floor*  
*Crockett C, 4th Floor*  
*Presidio A, 3rd Floor*

## Scientific Program

38. Margins and Monitoring of Non-inferiority Clinical Trials
39. Statistical Analysis of Informative Missing Data
40. Model-based Clustering of High-dimensional Genomic Data
41. Issues in Complicated Designs and Survival Analysis
42. Statistical Inference for Forest Inventory and Monitoring Using Remotely Sensed Data
43. Contributed Papers: Pre-processing and Quality Control for High-throughput Genomic Technologies
44. Contributed Papers: Assessing Gene and Environment Interactions in Genome-wide Studies
45. Contributed Papers: Hypothesis Testing
46. Contributed Papers: Variable Selection Methods
47. Contributed Papers: Longitudinal Data Analysis
48. Contributed Papers: Multiple Testing in High-dimensional Data

*Travis A, 3rd Floor*

*Independence, 3rd Floor*

*Travis B, 3rd Floor*

*Travis C, 3rd Floor*

*Travis D, 3rd Floor*

*Bonham B, 3rd Floor*

**3:30 p.m. – 3:45 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

## Refreshment Break and Visit the Exhibitors

**3:45 p.m. – 5:30 p.m.**

*Texas Ballroom E, 4th Floor*  
*Texas Ballroom C, 4th Floor*  
*Crockett C, 4th Floor*  
*Crockett A/B, 4th Floor*  
*Presidio A, 3rd Floor*  
*Independence, 3rd Floor*  
*Bonham B, 3rd Floor*  
*Travis A, 3rd Floor*  
*Travis B, 3rd Floor*  
*Travis C, 3rd Floor*  
*Travis D, 3rd Floor*

## Scientific Program

49. Role of Meta-Analysis in Drug Development
50. Analysis of High-dimensional Data with Biological Applications
51. Analysis of Longitudinal Data with Informative Observation and Dropout Processes
52. Addressing Key Statistical Issues in Environmental Epidemiology
53. Recent Development of Quantile Regression Methods for Survival Data
54. Contributed Papers: Adaptive Design in Clinical Trials
55. Contributed Papers: Gene Selection in DNA Microarray Studies
56. Contributed Papers: Image Analysis
57. Contributed Papers: Survey Research
58. Contributed Papers: Measurement Error Models
59. Contributed Papers: Mixture Models

**6:30 p.m. – 7:30 p.m.**

*Presidio B, 3rd Floor*

## President's Reception (By Invitation Only)

## TUESDAY, MARCH 17

**7:30 a.m. – 5:00 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

## Conference Registration

**8:00 a.m. – 5:00 p.m.**

*Goliad Room, 2nd Floor*

## Speaker Ready Room

**9:30 a.m. – 3:30 p.m.**

*Republic B, 4th Floor*

## Placement Service

**8:30 a.m. – 5:00 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

## Exhibits Open

**8:30 a.m. – 10:15 a.m.**

*Texas Ballroom A, 4th Floor*

*Presidio B, 3rd Floor*

*Crockett C, 4th Floor*

*Texas Ballroom B, 4th Floor*

*Texas Ballroom C, 4th Floor*

*Crockett D, 4th Floor*

*Travis A, 3rd Floor*

*Independence, 3rd Floor*

*Travis B, 3rd Floor*

*Travis C, 3rd Floor*

*Travis D, 3rd Floor*

**10:15 a.m. – 10:30 a.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**10:30 a.m. – 12:15 p.m.**

*Texas Ballroom, 4th Floor*

**12:30 p.m. – 4:30 p.m.**

*Presidio C, 3rd Floor*

**1:45 p.m. – 3:30 p.m.**

*Texas Ballroom A, 4th Floor*

*Texas Ballroom B, 4th Floor*

*Texas Ballroom C, 4th Floor*

*Presidio B, 3rd Floor*

*Crockett B, 4th Floor*

*Crockett C, 4th Floor*

*Independence, 3rd Floor*

*Travis A, 3rd Floor*

*Bonham B, 3rd Floor*

*Travis B, 3rd Floor*

*Travis C, 3rd Floor*

**3:30 p.m. – 3:45 p.m.**

*Texas Ballroom Pre-Function Area, 4th Floor*

**3:45 p.m. – 5:30 p.m.**

*Texas Ballroom C, 4th Floor*

*Presidio B, 3rd Floor*

*Texas Ballroom B, 4th Floor*

*Crockett B, 4th Floor*

*Texas Ballroom A, 4th Floor*

*Travis A, 3rd Floor*

*Crockett C, 4th Floor*

*Travis B, 3rd Floor*

*Travis C, 3rd Floor*

### **Scientific Program**

60. Statistical Issues in Brain Imaging
61. Causal Inference with Instrumental Variables Methods
62. From Theory to Practice: Examples and Discussion of Software Development for Wide Dissemination of Statistical Methodology
63. Bayesian Methods in Protein Bioinformatics
64. Adaptive Designs for Early-phase Clinical Trials
65. Memorial Session: The Life of David Beatty Duncan: Biostatistician, Mentor, and Gentleman
66. Contributed Papers: Spatial Analyses of Alcohol, Illegal Drugs, Violence and Race
67. Contributed Papers: Incorporating External Knowledge in the Analysis of Genomic Data
68. Contributed Papers: ROC Analysis
69. Contributed Papers: Model Selection/Assessment
70. Contributed Papers: Joint Models for Longitudinal and Survival Data

### **Refreshment Break and Visit the Exhibitors**

### **71. ENAR Presidential Invited Address**

### **Regional Committee Meeting (*By Invitation Only*)**

### **Scientific Program**

72. Prediction and Cure Modeling in Modern Medical Data Analysis
73. Network Analysis Models, Methods and Applications
74. Statistical Methods in Genome-wide Gene Regulation Studies
75. Experimental Designs in Drug Discovery & Clinical Trials
76. Statistical Methods for Flow Cytometry Data
77. Contributed Papers: Multiple Testing in Genome-wide Association Studies
78. Contributed Papers: Environmental and Ecological Applications
79. Contributed Papers: Categorical Data Analysis
80. Contributed Papers: Infectious Diseases
81. Contributed Papers: Rater agreement and Screening Tests
82. Contributed Papers: Applied Data Analysis, Graphical Displays, and Biostatistical Literacy

### **Refreshment Break and Visit the Exhibitors**

### **Scientific Program**

83. IMS Medallion Lecture
84. Mediation and Causal Inference
85. Challenges in the Bayesian Spatio-temporal Analysis of Large and Heterogeneous Datasets
86. Advances in Meta-analysis
87. Population Stratification Evaluation and Adjustment in Genome Wide Association Studies
88. Contributed Papers: Inference for Clinical Trials
89. Contributed Papers: Statistical Genetics
90. Contributed Papers: Health Services Research
91. Contributed Papers: Experimental Design

# PROGRAM SUMMARY

<i>Travis D, 3rd Floor</i>	92. Contributed Papers: Survival Analysis
<i>Crockett D, 4th Floor</i>	93. Contributed Papers: Functional Data Analysis
<b>5:30 p.m. – 6:00 p.m.</b> <i>Crockett B, 4th Floor</i>	<b>ENAR Business Meeting</b>
<b>6:15 p.m. – 9:30 p.m.</b>	<b>Tuesday Night Event – Dinner at Boudros (Registration Required)</b> <i>(Meet in the lobby of the hotel – shuttle buses begin to depart between 6:15 and 6:30 p.m.)</i>

## WEDNESDAY, MARCH 18

<b>7:30 a.m. – 9:00 a.m.</b> <i>Seguin A, 4th Floor</i>	<b>Planning Committee Breakfast Meeting (By Invitation Only)</b>
<b>8:00 a.m. – 12:30 p.m.</b> <i>Texas Ballroom Pre-Function Area, 4th Floor</i>	<b>Conference Registration</b>
<b>8:00 a.m. – 12:30 p.m.</b> <i>Crockett D, 4th Floor</i>	<b>Speaker Ready Room</b>
<b>8:00 a.m. – 12:00 p.m.</b> <i>Texas Ballroom Pre-Function Area, 4th Floor</i>	<b>Exhibits Open</b>
<b>8:30 a.m. – 10:15 a.m.</b> <i>Presidio B, 3rd Floor</i> <i>Presidio A, 3rd Floor</i> <i>Crockett A/B, 4th Floor</i> <i>Texas Ballroom C, 4th Floor</i> <i>Independence, 3rd Floor</i> <i>Bonham C, 3rd Floor</i> <i>Bonham D, 3rd Floor</i> <i>Travis A, 3rd Floor</i> <i>Travis B, 3rd Floor</i> <i>Travis C, 3rd Floor</i> <i>Travis D, 3rd Floor</i>	<b>Scientific Program</b> 94. Evaluating Markers for Risk Prediction 95. Advances in Functional Data Analysis 96. New Statistical Challenges and Advancements in Genome-wide Association Studies 97. Response-Adaptive Designs for Clinical Trials 98. Development of Bayesian Survival and Risk Analysis 99. Contributed Papers: Proteomics / Metabolomics 100. Contributed Papers: Detecting Gene Dependencies and Co-expression 101. Contributed Papers: Nonparametric Methods 102. Contributed Papers: Bayesian Spatial/Temporal Modeling 103. Contributed Papers: Missing Values in Survival and/or Longitudinal Data 104. Contributed Papers: Meta-Analysis
<b>10:15 a.m. – 10:30 a.m.</b> <i>Texas Ballroom Pre-Function Area, 4th Floor</i>	<b>Refreshment Break and Visit the Exhibitors</b>
<b>10:30 a.m.—12:15 p.m.</b> <i>Texas Ballroom C, 4th Floor</i> <i>Presidio B, 3rd Floor</i> <i>Crockett A/B, 4th Floor</i> <i>Presidio A, 3rd Floor</i> <i>Bonham C, 3rd Floor</i> <i>Travis A, 3rd Floor</i> <i>Bonham D, 3rd Floor</i> <i>Independence, 3rd Floor</i> <i>Travis B, 3rd Floor</i> <i>Travis C, 3rd Floor</i> <i>Travis D, 3rd Floor</i>	<b>Scientific Program</b> 105. Integrating Genomic and/or Genetics Data 106. Application of Dynamic Treatment Regimes 107. Data Sharing: An Example of Conflicting Incentive 108. Mapping Spatial Data into the Future 109. Strategies for Successful Statistical Consulting/Collaboration in Drug and Vaccine Discovery and Development 110. Contributed Papers: Clinical Trial Design 111. Contributed Papers: Genetic Studies with Related Individuals 112. Contributed Papers: Variable Selection for High-dimensional Data 113. Contributed Papers: Clustered Data Methods 114. Contributed Papers: Estimation in Survival Models 115. Contributed Papers: Biologics, Pharmaceuticals, Medical Devices

# SCIENTIFIC PROGRAM

## SUNDAY, MARCH 15

7:30 - 8:00 p.m. New Member Reception  
Texas Ballroom, 4th Floor

8:00 – 11:00 p.m. Opening Mixer and Poster Presentations Reception  
Texas Ballroom, 4th Floor

### POSTER PRESENTATIONS

#### 1. POSTERS: CLINICAL TRIALS

Sponsor: ASA Biopharmaceutical Section

**1a. Comparison of Three Adaptive Dose-finding Models for Combination Therapy in Phase I Clinical Trials**

Rui Qin\*, Mayo Clinic, Yufen Zhang, University of Minnesota, Sumithra J. Mandrekar, Mayo Clinic, Wei Zhang, Boehringer Ingelheim, Daniel J. Sargent, Mayo Clinic

**1b. Incorporating Patient Heterogeneity in Adaptive Phase I Trial Designs**

Thomas M. Braun\*, University of Michigan School of Public Health

**1c. A Phase 2 Clinical Trial with Adaptation on 2 Factors**

Richard J. McNally\* and David McKenzie, Celgene Corporation

**1d. Estimating Percentiles in Dose-response Curves for Delayed Responses using an Adaptive Compound Urn Design**

Rameela Chandrasekhar\* and Gregory E. Wilding, University at Buffalo/Roswell Park Cancer Institute

**1e. Proportional Odds Model for Design of Dose-finding Clinical Trials with Ordinal Toxicity Grading**

Emily M. Van Meter\*, Elizabeth Garrett-Mayer and Dipankar Bandyopadhyay, Medical University of South Carolina

**1f. Simulation of the Optimal Timing of a Multiple Regime**

Yining Du\*, Ning Wang and Clyde Martin, Texas Tech University

#### 2. POSTERS: POWER/SAMPLE SIZE

Sponsor: ASA Biopharmaceutical Section

**2a. Determination of Sample Size for Demonstrating Efficacy of Radiation Countermeasures**

Ralph L. Kodell, Shelly Y. Lensing\* and Reid D. Landes, University of Arkansas for Medical Sciences, K. Sree, Kumar Armed Forces Radiobiology Research Institute, Martin Hauer-Jensen, University of Arkansas for Medical Sciences

**2b. When ICCs go AWRY: A Case Study from a School-based Smoking Prevention Study in South Africa**

Ken Resnicow and Nanhua Zhang\*, School of Public Health, University of Michigan, Roger D. Vaughan, Columbia University, Sasiraghla P. Reddy, Medical Research Council of South Africa, Cape Town, South Africa

**2c. Sample Size Determination for a 5 Year Longitudinal Clinical Trial In Children: Using Simulation**

Yahya A. Daoud\*, Sunni A. Barnes and Dunlei Cheng, Baylor Health Care System, Ed DeVol and C. Richard Boland, Baylor University Medical Center

**2d. Sample Size Calculation for Clustered Binary Outcomes with Sign Tests**

Fan Hu\* and William Schucany, Southern Methodist University, Chul Ahn, University of Texas Southwestern Medical Center

**2e. Sample Size Calculation for Two-Stage Randomization Designs with Censored Data**

Zhiguo Li\* and Susan Murphy, Institute for Social Research, University of Michigan

**2f. Compare Sample Size Adjustment Methods for Cluster Randomized Trials**

Dhuly Chowdhury\* and Hrishikesh Chakraborty, RTI International

**2g. Performance of the Hochberg Multiple Testing Procedure in Cluster Randomized Designs**

Jeffrey J. Wing\*, Brisa N. Sánchez and Cathie Spino, University of Michigan

**2h. Comparison of Simon's Two-Stage Design, Sequential Probability Ratio Test, and Triangular Test in Phase II Clinical Trials**

Leigh A. Morton\* and David T. Redden, Ph.D., University of Alabama-Birmingham

#### 3. POSTERS: MICROARRAY ANALYSIS

Sponsor: ASA Biometrics Section

**3a. Differential DNA Methylation: Methodology and Study Design**

Richard E. Kennedy\* and Xiangqin Cui, Section on Statistical Genetics, University of Alabama-Birmingham

**3b. Use of the Item Response Theory to Evaluate Gene Expression in Congenic Rat Strains**

Julia P. Soler\* and Carlos E. Neves, University of Sao Paulo, Brazil, Suely R. Giolo, Federal University of Parana, Brazil, Dalton F. Andrade, Federal University of Santa Catarina, Brazil, Mariza de Andrade, Mayo Clinic, Ayumi A. Miyakawa and Jose E. Krieger, Heart Institute, University of Sao Paulo, Brazil

**3c. Differentiating mRNA Expression Levels of Tumor Versus Non-Tumor Cells in a Cancer Tissue**

Li-yu D. Liu\*, National Taiwan University

**3d. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression using Java Visualization**

Rui Ding\*, University of Minnesota-Morris, Jong-Min Kim, University of Minnesota-Morris, Deukwoo Kwon, Division of Cancer Epidemiology and Genetics, National Cancer Institute

**3e. Biomarker Detection Methods When Combining Multiple Multi-class Microarray Studies**

Shuya Lu\*, Jia Li and George C. Tseng, University of Pittsburgh

\* Presenter

# SCIENTIFIC PROGRAM

## 3f. Strategies for Applying Gene Signatures to Prospective Clinical Studies

William T. Barry\* and Michael Datto, Duke University Medical Center

## 4. POSTERS: STATISTICAL GENETICS/GENOMICS

Sponsor: ASA Biometrics Section

### 4a. Haplotype Analysis of Quantitative Traits in Outcrossing Plant Populations

Wei Hou\*, Department of Epidemiology and Health Policy Research, University of Florida, Rongling Wu, Department of Statistics, University of Florida and Department of Public Health Sciences, Penn State University

### 4b. A Multi-step Approach to Genetic Association for Asthma Characteristics in the Isle of Wight Birth Cohort

Marianne Huebner\*, Mayo Clinic, Hasan Arshad, University of Southampton, UK, Eric Schauberger and Karen Friderici, Michigan State University, Marsha Wills-Karp, Cincinnati Childrens Hospital, Wilfried Karmaus, University of South Carolina, Susan Ewart, Michigan State University

### 4c. Value of SNPs in Models that Predict Breast Cancer Risk

Mitchell H. Gail\* and Ruth M. Pfeiffer, Division of Cancer Epidemiology and Genetics, National Cancer Institute

### 4d. An Approach to Detect Gene-Gene Interactions in SNP Data Based on Probabilistic Measures

Ramon Casanova\*, Josh D. Grab, Miranda C. Marion, Paula S. Ramos, Jasmin Divers and Carl D. Langefeld, Wake Forest University Health

### 4e. Change-Point Identification in Hidden Markov Models for DNA Sequence Segmentation Modeling

Darfiana Nur\*, University of Newcastle, Australia, Kerrie L. Mengersen, Queensland University of Technology, Australia

## 5. POSTERS: CAUSAL INFERENCE

Sponsor: ASA Section on Statistics in Epidemiology

### 5a. Causal Inference for Intervention Effects on Nicotine Withdrawal Symptoms

Brian L. Egleston\*, Fox Chase Cancer Center, Karen L. Cropsey, University of Alabama School of Medicine, Amy B. Lazev and Carolyn J. Heckman, Fox Chase Cancer Center

### 5b. Estimation of Marginal Structural Survival Models in the Presence of Competing Risks

Maarten Bekaert\* and Stijn Vansteelandt, Ghent University, Ghent, Belgium, Karl Mertens, Scientific Institute of Public Health, Brussels, Belgium

### 5c. A Markov Compliance Class and Outcome Model for Causal Analysis in the Longitudinal Setting

Xin Gao\* and Michael R. Elliott, School of Public Health, University of Michigan

## 6. POSTERS: IMAGING

Sponsor: ASA Biometrics Section

### 6a. A Bayesian Generalized Non-Linear Predictive Model of Treatment Efficacy Using qMRI

Jincao Wu\* and Timothy D. Johnson, University of Michigan

### 6b. Spatial Point Process Modeling of Group fMRI Data

Timothy D. Johnson\*, University of Michigan, Thomas E. Nichols, GlaxoSmithKline; University of Oxford, FMRIB; University of Michigan, Lei Xu, Vanderbilt University, Tor D. Wager, Columbia University

### 6c. Meta-Analysis of fMRI Data Via a Bayesian Cox Cluster Process

Jian Kang\* and Timothy D. Johnson, University of Michigan, Thomas E. Nichols, GlaxoSmithKline; University of Oxford, FMRIB; University of Michigan, Tor D. Wager, Columbia University

### 6d. Extraction of the Hemodynamic Response Function and Parameter Estimation for the Two Gamma Difference Model

Joel C. O'Hair\*, Richard F. Gunst, William R. Schucany and Wayne A. Woodward, Southern Methodist University

### 6e. Wavelet Packet Resampling for fMRI Experiments

Ohn Jo Koh\*, William R. Schucany, Richard F. Gunst and Wayne A. Woodward, Southern Methodist University

### 6f. Dirichlet Process Models for Changes in fMRI Visual Field

Raymond G. Hoffmann\*, Pippa Simpson, Shun-Hwa Li and Ke Yan, Edgar A. DeYoe and Daniel B. Rowe, Medical College of Wisconsin

## 7. POSTERS: SURVIVAL ANALYSIS

Sponsor: ASA Biometrics Section

### 7a. Analyzing Patient Survival after Deceased-donor Kidney Transplants: The Novel Use of Time-varying Covariates

Arwin M. Thomasson\*, Peter P. Reese and Justine Shults, University of Pennsylvania

### 7b. Survomatic: A User-Friendly Package for Analysis of Survival and Mortality Data

Alex F. Bokov\*, University of Texas Health Science Center at San Antonio, Scott D. Pletcher, Baylor College of Medicine, Department of Molecular and Human Genetics and Huffington Center on Aging, Jonathan A.L. Gelfond, University of Texas Health Science Center at San Antonio

### 7c. Association Between Progression-free and Overall Survival in Randomized Clinical Trials

Kristine Broglio\* and Donald Berry, U.T. M.D. Anderson Cancer Center

### 7d. On an Empirical Method for a Generalised Version of the Yang and Prentice Model

Carl M. DiCasoli\*, Sujit K. Ghosh and Subhashis Ghosal, North Carolina State University



#### **7e. Bayesian Hazard Rate Estimation and Sufficient Dimension Reduction**

Shraddha S. Mehta\*, Purdue University, Surya T. Tokdar, Carnegie Mellon University, Jayanta K. Ghosh, Purdue University and Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, Kolkata, India, Bruce A. Craig, Purdue University

#### **7f. Estimating Mediation Effects in Survival Analysis with Censored Data**

Yanhui Sun\*, Chichi Aban, Gary R. Cutter and David L. Roth, University of Alabama at Birmingham

#### **7g. Joint Modeling of Survival and Binomial Data Based on Generalized Self-consistency with Application to Prostate Cancer Stage-specific Incidence**

Chen Hu\* and Alex Tsodikov, School of Public Health, University of Michigan

#### **7h. Survival Analysis with Error Prone Time-varying Covariates: A Risk Set Calibration Approach**

Xiaomei Liao\*, Harvard School of Public Health, David Zucker, Hebrew University, Jerusalem, Israel, Yi Li and Donna Speigelman, Harvard School of Public Health

#### **7i. Longitudinal Changes in Carotid IMT and Risk of MI, Stroke and CHD: The Cardiovascular Health Study**

David Yanez, Michal Juraska\* and Bruce M. Psaty, University of Washington, Mary Cushman, University of Vermont, Cam Solomon, CHSCC, University of Washington, Joseph F. Polak and Daniel O'Leary, Tufts University

### **8. POSTERS: MISSING DATA**

Sponsor: ASA Health Policy Statistics Section

#### **8a. A Hot-deck Multiple Imputation Procedure for Gaps in Longitudinal Event Histories**

Chia-Ning Wang\*, Roderick Little, Bin Nan and Sioban Harlow, University of Michigan

#### **8b. A Multiple Imputation Approach to the Analysis of Interval-censored Failure Time Data with the Additive Hazards Model**

Ling Chen\* and Jianguo Sun, University of Missouri-Columbia

#### **8c. Assessing the Convergence of Multiple Imputation Algorithms Using a Sequence of Regression Models**

Jian Zhu\* and Trivellore E. Raghunathan, University of Michigan

#### **8d. A Comparison of Missing Data Methods for Quality of Life Measures in a Clinical Trial with Long-term Follow-up**

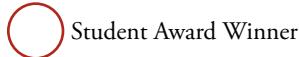
Paul Kolm\* and Wei Zhang, Christiana Care Health System, John A. Spertus, Mid America Heart Institute, David J. Maron, Vanderbilt University, William E. Boden, Buffalo General Hospital, William S. Weintraub, Christiana Care Health System

#### **8e. Analysis of Non-ignorable Missing and Left-Censored Longitudinal Biomarker Data Using Weighted Pseudo Likelihood Method**

Abdus Sattar\* and Lisa Weissfeld, University of Pittsburgh

#### **8f. Estimation in Hierarchical Models with Incomplete Data**

Yong Zhang\* and Trivellore E. Raghunathan, University of Michigan



#### **8g. An Approach to Sensitivity Analysis of Nested Case-control Studies with Outcome-dependent Follow-up**

Kenneth J. Wilkins\*, Infectious Disease Clinical Research Program, Department of Preventive Medicine & Biometrics, Uniformed Services University of the Health Sciences

#### **8h. Missing Animals in Toxic Treatment Studies**

Pippa M. Simpson\*, Shun H. Li and Ke Yan, Medical College of Wisconsin, Bevan E. Huang, CSIRO, Calvin Williams, Dipeca Haribhai, and Raymond G. Hoffmann, Medical College of Wisconsin

#### **8i. Pseudolikelihood Ratio Tests with Biased Observations**

Bin Zhang\*, Boston University, Joan X. Hu, Simon Fraser University

### **9. POSTERS: SPATIAL/TEMPORAL MODELING AND ENVIRONMENTAL/ECOLOGICAL APPLICATIONS**

Sponsors: ASA Section on Statistics and the Environment; ASA Section on Statistics in Defense and National Security

#### **9a. Estimating the Maximum Growth Rate of Harmful Algal Blooms Using a Combined Model Method**

Margaret A. Cohen\*, University of North Carolina at Wilmington

#### **9b. A Spatial Scan Statistic for Multinomial Data**

Inkyung Jung\*, University of Texas Health Science Center at San Antonio, Martin Kulldorff, Harvard Medical School, Otukei J. Richard, Makerere University, Kampala, Uganda

#### **9c. Longitudinal Spatial Point Processes for Residential Histories**

Patrick E. Brown\*, Cancer Care Ontario, Peter Henrys, Lancaster University

#### **9d. Hierarchical Dynamic Modeling of Spatial-Temporal Binary Data**

Yanbing Zheng\*, University of Kentucky, Jun Zhu, University of Wisconsin – Madison, Brian Aukema, Natural Resources Canada, Canadian Forest Service and Ecosystem Science and Management Program, University of Northern British Columbia

#### **9e. Two-stage Generalized Method of Moments Estimation with Applications in Spatio-temporal Models**

Yun Bai\*, Peter X.K. Song and Trivellore Raghunathan, University of Michigan

#### **9f. Spatio-Temporal Modelling for Lupus Incidence in Toronto, Canada since 1965**

Ye Li\*, University of Toronto, Patrick E. Brown, Cancer Care Ontario

#### **9g. Spatial Modeling of Air Pollution Exposure, Measurement Error and Adverse Birth Outcomes**

Simone Gray\* and Alan Gelfand, Duke University, Marie Lynn Miranda and Sharon Edwards, Nicholas School of the Environment

#### **9h. Statistical Analysis of The Effects of Air Pollution on Children's Health**

Elizabeth A. Stanwyck\* and Bimal Sinha, University of Maryland Baltimore County

#### **9i. The Effect of Rainfall on Visits to Pediatric Emergency Rooms for Diarrhea**

Shun H. Li\*, Pippa M. Simpson, Stephen Stanhope, Ke Yan, Marc Gorelick and Medical College of Wisconsin, Bevan E. Huang, CSIRO, Raymond G. Hoffmann, Medical College of Wisconsin

\* Presenter

# SCIENTIFIC PROGRAM

## 9j. Stochastic Models of Flow Through a Random Graph

Nicholas M. Murray\* and Clyde Martin, Texas Tech University, Dorothy Wallace, Dartmouth College

## 9k. Bayesian Alignment of Continuous Molecular Shapes

Irina Czogiel\*, Ian L. Dryden and Christopher J. Brignell, University of Nottingham, UK

## 9l. Modelling Spatio-temporal Trends of Forest Health Monitoring Data

Nicole H. Augustin\*, University of Bath Mathematical Sciences, Bath, UK, Monica Musio, University of Cagliari, Italy, Klaus von Wilpert and Edgar Kublin, Forest Research Centre Baden-Wurttemberg, Freiburg, Germany, Simon N. Wood, University of Bath Mathematical Sciences, Bath, UK, Martin Schumacher, University Hospital Freiburg University, Freiburg, Germany

## 10. POSTERS: CATEGORICAL DATA ANALYSIS AND SURVEY RESEARCH

Sponsor: ASA Survey Research and Methodology Section

### 10a. Synthesizing Categorical Datasets to Enhance Inference

Veronica J. Berrocal\* and Alan E. Gelfand, Duke University, Sourab Bhattacharya, Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute, Marie L. Miranda, Duke University, Nicholas School of the Environment, Geeta Swamy, Duke University, Department of Obstetrics and Gynecology

### 10b. Use of Secondary Data Analysis and Instantaneous States in a Discrete-State Model of Diabetic Heart Disease

Jacob Barhak, Deanna JM Isaman and Wen Ye\*, University of Michigan

### 10c. Sampling Tables Given a Set of Conditionals

Juyoun Lee\* and Aleksandra Slavkovic, Penn State University

### 10d. Examining the Robustness of Fully Synthetic Data

Techniques for Data with Binary Variables  
Gregory Matthews\*, University of Connecticut

### 10e. Bayesian Model-based Estimates of Diabetes Incidence by State

Theodore J. Thompson, Betsy L. Cadwell, James P. Boyle and Lawrence Barker, Centers for Disease Control and Prevention

### 10f. Health Disparity Indices - Simulations of Underlying Dependencies

Stuart A. Gansky\* and Nancy F. Cheng, University of California, San Francisco, Center to Address Disparities in Children's Oral Health

### 10g. New Development of Optimal Coefficients for a Best Linear Unbiased Estimator of the Total for Simple Random Sampling with Replacement Using Godambe's General Linear Estimator

Shuli Yu \* and Edward J. Stanek III, School of Public Health, University of Massachusetts-Amherst

## 11. POSTERS: VARIABLE/MODEL SELECTION

Sponsor: ASA Biometrics Section

### 11a. Nonparametric Bayes Conditional Distribution Modeling with Variable Selection

Yeonseung Chung\*, Harvard School of Public Health and David Dunson, Duke University

### 11b. A New Approach to High Dimensional Variable Selection

Xingye Qiao\*, Yufeng Liu and J.S. Marron, University of North Carolina-Chapel Hill

### 11c. Testing for Conditional Independence Via Multiple Models: An Orthopedic Application for the Six Segment Foot Model

Sergey Tarima\*, Xue-Cheng Liu, Roger Lyon, John Thometz and Channing Tassone, Medical College of Wisconsin

### 11d. Model Checking for Bayesian Estimation of State Diabetes Incidence Rates

James P. Boyle\*, Betsy L. Cadwell, Theodore J. Thompson and Lawrence Barker, Centers for Disease Control

### 11e. Checking Transformation Models With Censored Data

Li Chen\*, University of North Carolina at Chapel Hill

### 11f. Classification of Functional Data: A Segmentation Approach

Bin Li\* and Qingzhao Yu, Louisiana State University

### 11g. Network Exploration Via the Adaptive LASSO and SCAD Penalties

Jianqing Fan and Yang Feng\*, Princeton University, Yichao Wu, North Carolina State University

## 12. POSTERS: DIAGNOSTIC TESTS

Sponsor: ENAR

### 12a. Determining Presence of GB Virus Type C in HIV Positive Subjects

Carmen J. Smith\* and Kathryn Chaloner, University of Iowa

### 12b. Modeling Sensitivity and Specificity with a Time-varying Reference Standard within a Longitudinal Setting

Qin Yu\* and Wan Tang, University of Rochester, Sue Marcus, Department of Psychiatry, Mount Sinai School of Medicine, Yan Ma and Xin M. Tu, University of Rochester

### 12c. Comparison of Correlated Correlation Coefficients using Bootstrapping

Juhee Song\*, Scott & White Hospital, Jeffrey D. Hart, Texas A&M University

### 12d. Methods for Calibrating Bivariate Laboratory Data

Ke Yan\*, Raymond G. Hoffmann, Shi-Hwan Li, Robert Montgomery and Pippa Simpson, Medical College of Wisconsin"

## 13. POSTERS: NONPARAMETRIC METHODS

Sponsor: ENAR

### 13a. Estimating the Variance of BJE under Discrete Assumption

Yishi Wang\* and Cuixian Chen, University of North Carolina-Wilmington

### 13b. The GMLE Based Buckley-James Estimator with Modified Case-cohort Data

Cuixian Chen\*, University of North Carolina-Wilmington



**13c. Bootstrap Confidence Intervals for the Predictors of Treatment Means in a One Factor Experimental Design with A Finite Population**

Bo Xu\* and Edward Stanek, School of Public Health, University of Massachusetts-Amherst

**13d. Generalized ANOVA for Currently Modeling Mean and Variance within a Longitudinal Data Setting**

Hui Zhang\* and Xin Tu, University of Rochester Medical Center

**13e. Practical Estimation and Discussion of Neuronal Phase-response (Phase-resetting) Curves**

Daniel G. Polhamus\*, Charles J. Wilson and Carlos A. Paladini, University of Texas, San Antonio

**14. POSTERS: STATISTICAL MODELS AND METHODS**

Sponsor: ENAR

**14a. Further Development of Semi-parametric Methods in Bayesian Beta Regression**

Christopher J. Swearingen\* and Dipankar Bandyopadhyay, Medical University of South Carolina

**14b. Penalized Maximum Likelihood Estimation in Logistic Dose Response Models**

Amy E. Wagler\*, University of Texas at El Paso

**14c. Achieving Covariance Robustness in the Linear Mixed Model**

Matthew J. Gurka, University of Virginia and Keith E. Muller\*, University of Florida

**14d. Bayesian Mixtures for Modeling the Correlation of Longitudinal Data**

Lei Qian\* and Robert Weiss, University of California at Los Angeles

**14e. Estimation of Probability Distributions using Control Theoretic Splines**

Janelle K. Charles\* and Clyde F. Martin, Texas Tech University

**14f. Application of the Kalman Filter Algorithm to Estimate a Functional Mixed Model**

Meihua Wu\* and Brisa N. Sánchez, University of Michigan

**MONDAY, MARCH 16**

**8:30 a.m.—10:15 a.m.**

**15. STATISTICAL ANALYSIS OF METABOLOMICS DATA**

CROCKETT C, 4TH FLOOR

Sponsor: ENAR

Organizer: David L. Banks, Duke University

Chair: Leanna House, Virginia Tech University

**8:30 Statistical Issues in Metabolomics**

David L. Banks\*, Duke University

**9:00 Statistical Ways to Choose a Distance Measure for Metabolomic Data**

Philip M. Dixon\*, Iowa State University

**9:30 Incorporating Interactive Graphics into Metabolomics Data Pre-processing**

Dianne Cook\*, Iowa State University, Michael Lawrence, Suh-yeon Choi, Heike Hofmann, Eve Wurtele

**10:00 Floor Discussion**

**16. ADVANCED STATISTICAL METHODS FOR HEALTH SERVICES RESEARCH**

CROCKETT A/B, 4TH FLOOR

Sponsors: ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section

Organizer: Joel Greenhouse, Carnegie Mellon University

Chair: Joel Greenhouse, Carnegie Mellon University

**8:30 Characterizing Patterns of Treatment Utilization for Youth with ADHD**

Gary Klein\* and Joel Greenhouse, Carnegie Mellon University, Abigail Schlesinger, Western Psychiatric Institute and Clinic, University of Pittsburgh, Bradley Stein, Western Psychiatric Institute and Clinic, University of Pittsburgh, Community Care Behavioral Health Organization, RAND Corporation

**9:00 Statistical Strategies for PostMarket Surveillance of Medical Devices**

Sharon-Lise T. Normand\*, Department of Health Care Policy, Harvard Medical School

**9:30 The Role of Health and Health Behaviors in the Formation and Dissolution of Friendship Ties**

A James O'Malley\* and Nicholas A. Christakis, Harvard Medical School

**10:00 Floor Discussion**

**17. MODEL SPECIFICATION AND UNCERTAINTY IN ECOLOGICAL ANALYSES**

INDEPENDENCE, 3RD FLOOR

Sponsor: ASA Section on Statistics and the Environment

Organizer: Ali Arab Georgetown University

Chair: Mevin Hooten, Utah State University

**8:30 Data-model Integration for Understanding Belowground Ecosystems**

Kiona Ogle\*, University of Wyoming

**9:00 A Bayesian Bioclimate Model for the Lower Trophic Ecosystem in the North Pacific Ocean**

Christopher K. Wikle\*, University of Missouri

**9:30 Modeling and Inference of Animal Movement in Response to Landscapes**

Jun Zhu\*, University of Wisconsin – Madison, Jeff Tracey and Kevin Crooks, Colorado State University

**10:00 Floor Discussion**

**18. ANALYSIS CHALLENGES OF MODERN LONGITUDINAL BIOMEDICAL DATA**

PRESIDIO A, 3RD FLOOR

Sponsor: ASA Biopharmaceutical Section, IMS

Organizer: Yingye Zheng, Fred Hutchinson Cancer Research Center

Chair: Tianxi Cai, Harvard University

# SCIENTIFIC PROGRAM

- 8:30 Incorporating Correlation for Multivariate Failure Time Data When Cluster Size Is Large**  
Li Wang and Lan Xue, Oregon State University, Annie Qu\*, University of Illinois at Urbana-Champaign
- 8:55 Variable Selection in Additive Mixed Models for Longitudinal Data**  
Daowen Zhang\*, North Carolina State University
- 9:20 Individualized Prediction in Prostate Cancer Studies Using a Joint Longitudinal-Survival-Cure Model**  
Menggang Yu\*, Indiana University, School of Medicine
- 9:45 Longitudinal Analysis of Surgical Trials with Non-compliance**  
Patrick J. Heagerty\* and Colleen Sitrani, University of Washington
- 10:10 Floor Discussion**

## 19. RECENT ADVANCES ON FEATURE SELECTION AND ITS APPLICATIONS

TEXAS BALLROOM E, 4TH FLOOR

Sponsor: IMS

Organizer: Runze Li, The Pennsylvania State University

Chair: Runze Li, The Pennsylvania State University

- 8:30 Feature Selection in GLM with Large Model Spaces**  
Jiahua Chen\*, University of British Columbia, Zehua Chen, The National University of Singapore
- 8:55 Higher Criticism Thresholding: Optimal Feature Selection when Features are Rare and Weak**  
Jiashun Jin\*, Carnegie Mellon University, David L. Donoho, Stanford University
- 9:20 Weighted Wilcoxon-type Smoothly Clipped Absolute Deviation Method**  
Lan Wang\*, University of Minnesota, Runze Li, The Pennsylvania State University
- 9:45 Ultrahigh Dimensional Variable Selection: Beyond the Linear Model**  
Jianqing Fan\*, Princeton University, Richard Samworth, University of Cambridge, Yichao Wu, North Carolina State University
- 10:10 Floor Discussion**

## 20. CONTRIBUTED PAPERS: ANALYSIS OF GENOME-WIDE SNP ARRAYS

BONHAM B, 3RD FLOOR

Sponsor: ASA Biometrics Section

Chair: Xiaobo Li, University of Florida

- 8:30 Pharmacogenomics Meta-analysis - A Simulation Study**  
Mei Yang\*, Boston University, Meng Chen, Pfizer Global Research & Development
- 8:45 Simultaneous Bayesian Multiple Shrinkage Inference for Genetic Association Studies Allowing for Mode of Inheritance Uncertainty**  
Nicholas M. Pajewski\*, University of Alabama at Birmingham, Purushottam W. Laud, Medical College of Wisconsin
- 9:00 Using Cases from Genome-Wide Association Studies to Strengthen Inference on the Association between Single Nucleotide Polymorphisms and a Secondary Phenotype**  
Hulin Li\* and Mitchell H. Gail, National Cancer Institute

- 9:15 An Algorithm for Constructing an Imprinted Map of the Cancer Genome**  
Louie R. Wu\*, Buchholz High School, Yao Li, Rongling Wu and Arthur Berg, University of Florida
- 9:30 A Mixture Model for the Analysis of Allelic Expression Imbalance**  
Rui Xiao\*, Michael Boehnke and Laura Scott, University of Michigan
- 9:45 Mapping Imprinted Quantitative Trait Loci Underlying Endosperm Trait in Flowering Plant: A Variance Component Approach**  
Gengxin Li\* and Yuehua Cui, Michigan State University
- 10:00 A Bayesian Change-point Algorithm for Detecting Copy Number Alteration**  
Fridtjof Thomas\*, University of Tennessee Health Science Center, Stanley Pounds, St. Jude Children's Research Hospital

## 21. CONTRIBUTED PAPERS: BIOMARKERS AND DIAGNOSTIC TESTS

TRAVIS A, 3RD FLOOR

Sponsor: ASA Biometrics Section

Chair: Eunhee Kim, University of North Carolina at Chapel Hill

- 8:30 Information Theoretic Approach to Surrogate Markers Evaluation for Time-to-Event Clinical Endpoints**  
Pryseley N. Assam\*, Abel E. Tilahun, Ariel Alonso and Geert, Molenberghs, Hasselt University-Belgium
- 8:45 Median Regression for Longitudinal Biomarker Measurements Subject to Detection Limit**  
Kong Lan\* and Minjae Lee, School of Public Health, University of Pittsburgh
- 9:00 A Multiple Imputation Approach for Left-censored Biomarkers with Limits of Detection**  
Minjae Lee\* and Lan Kong, School of Public Health, University of Pittsburgh
- 9:15 Estimation and Comparison of the Predictiveness Curve for Repeated Measures Design**  
Kwonho Jeong\*, Abdus M. Sattar and Lisa Weissfeld, University of Pittsburgh
- 9:30 An Evaluation of Logic Forest for Identification of Disease Biomarkers**  
Bethany J. Wolf\*, Elizabeth H. Slate and Elizabeth G. Hill, Medical University of South Carolina
- 9:45 Estimates of Observed Sensitivity and Specificity Must be Corrected when Reporting the Results of the Second Test in a Screening Trial Conducted in Series**  
Brandy M. Ringham\* and Deborah H. Glueck, University of Colorado
- 10:00 Bayesian Hierarchical Modeling of Probabilities from Repeated Binary Diagnostic Tests**  
Daniel P. Beavers\*, James D. Stamey and John W. Seaman III, Baylor University

## 22. CONTRIBUTED PAPERS: CAUSAL INFERENCE

TRAVIS B, 3RD FLOOR

Sponsors: ASA Health Policy Statistics Section, ASA Section on Statistics in Epidemiology

Chair: Brian L. Egleston, Fox Chase Cancer Center

- 8:30 Causal Inference with Longitudinal Subpopulation Indicators**  
Booil Jo\*, Stanford University



8:45	<b>A Causal Model of Baseline and Post-treatment Confounding for Observational Studies</b> Chen-pin Wang*, University of Texas Health Science Center-San Antonio	<b>24. CONTRIBUTED PAPERS: POWER/SAMPLE SIZE</b> TRAVIS D, 3RD FLOOR Sponsor: ASA Health Policy Statistics Section Chair: Michelle Shardell, University of Maryland, School of Medicine
9:00	<b>Accounting for Unmeasured Confounders with Latent Variable</b> Haiqun Lin*, Yale University School of Public Health	<b>8:30 Power Analysis for Mediation Effect in Longitudinal Studies</b> Cuiling Wang*, Albert Einstein College of Medicine
9:15	<b>A Causal Selection Model to Compare Treatment Groups in a Subset Selected Post-Randomization with Application to an HIV Antiretroviral Immunotherapy Trial</b> Robin Mogg* and Marshall M. Joffe, University of Pennsylvania School of Medicine, Devan V. Mehrotra, Merck Research Laboratories, Thomas R. Ten Have, University of Pennsylvania School of Medicine	<b>8:45 Power and Type I Error Rates in Repeated Measures Experiments as the Number of Time Points in a Fixed Length Time Interval Increase and Under Several Covariance Structures for the Repeated Measures</b> John D. Keighley*, University of Kansas Medical Center, Dallas E. Johnson, Kansas State University
9:30	<b>Estimating Drug Effects in the Presence of Placebo Response</b> Bengt O. Muthén*, UCLA, Hendricks C. Brown, University of South Florida	<b>9:00 Intracluster Correlation Adjustments to Maintain Power in Cluster Trials for Binary Variables</b> Hrishikesh Chakraborty*, Janet Moore and Tyler D. Hartwell, RTI International.
9:45	<b>Inference on Treatment Effects from a Randomized Clinical Trial in the Presence of Premature Treatment Discontinuation: The SYNERGY Trial</b> Min Zhang*, University of Michigan, Anastasios A. Tsiatis and Marie Davidian, North Carolina State University, Karen S. Pieper and Kenneth Mahaffey, Duke Clinical Research Institute	<b>9:15 A Two-stage Adaptive Design can Increase Power for Multiple Comparisons</b> Deborah H. Glueck* and Anis Karimpour-Fard, University of Colorado, Keith E. Muller, University of Florida
10:00	<b>Detection of Surrogates using a Potential Outcomes Framework</b> Andreas G. Klein*, University of Western Ontario	<b>9:30 Long Term Survivor Models and Two Component Mixture Models</b> Wonkuk Kim*, University of South Florida"
<b>23. CONTRIBUTED PAPERS: AN EM APPROACH FOR PARTIAL CORRELATION AND MISSING DATA</b> TRAVIS C, 3RD FLOOR Sponsor: ASA Health Policy Statistics Section Chair: Mohamed Alosif, U.S. Food and Drug Administration, HHS		
8:30	<b>An EM Approach for Partial Correlation and Missing Data</b> Gina D'Angelo* Washington University School of Medicine	<b>9:45 Determination of Sample Size for Validation Study in Pharmacogenomics</b> Youlan Rao*Yoonkyung Lee and Jason C. Hsu, The Ohio State University
8:45	<b>Latent Variable Regression for Multiple Outcomes with some Predictors Missing Non-randomly</b> Chengjie Xiong, Washington University School of Medicine, Division of Biostatistics	<b>10:00 R Programs for Calculating Sample Size and Power in Bioequivalence Trials</b> Qinfang Xiang*, Endo Pharmaceuticals, Inc.
9:00	<b>Outfluence -- The Impact of Missing Values</b> Yvette I. Sheline, Washington University School of Medicine, Department of Psychiatry	
9:15	<b>A Double Robust Local Multiple Imputation</b> Chiu-Hsieh Hsu*, University of Arizona, Qi Long, Emory University	
9:30	<b>Avoid Ecological Fallacy: Using BART to Impute Missing Ordinal Data</b> Song Zhang*, University of Texas Southwestern Medical Center, Tina Shih and Peter Muller, University of Texas M.D. Anderson Cancer Center	
9:45	<b>Meta-Analysis of Studies with Missing Data</b> Ying Yuan*, University of Texas M.D. Anderson Cancer, Roderick Little, University of Michigan	
10:00	<b>Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data</b> Weihsia Cao*, Anastasios A. Tsiatis and Marie Davidian, North Carolina State University	
<b>25. CONTRIBUTED PAPERS: MULTIVARIATE SURVIVAL</b> REPUBLIC A, 4TH FLOOR Sponsor: ASA Biometrics Section Chair: Paul S Albert, National Institutes of Health, National Cancer Institute		
8:30		<b>8:30 Nonparametric Quantile Estimation for Successive Events Subject to Censoring</b> Adin-Cristian Andrei*, University of Wisconsin-Madison
8:45		<b>8:45 Nonparametric and Semiparametric Estimations for Bivariate Failure Time Distribution with Interval Sampling</b> Hong Zhu* and Mei-Cheng Wang, Johns Hopkins University
9:00		<b>9:00 Partially Monotone Spline Estimation with Bivariate Current Status Data</b> Yuan Wu*and Ying Zhang, University of Iowa
9:15		<b>9:15 Comparison of State Occupation, Entry, Exit and Waiting Times in K Independent Multistate Models under Current Status Data</b> Ling Lan*, Medical College of Georgia, Somnath Datta, University of Louisville
9:30		<b>9:30 Additive Hazards Model for Case-cohort Studies with Multiple Disease Outcomes</b> Sangwook Kang*, University of Georgia, Jianwe Cai, University of North Carolina at Chapel Hill

# SCIENTIFIC PROGRAM

- 9:45 **Modelling Cumulative Incidences of Dementia and Dementia-free Death Using a Novel Three-parameter Logistic Function**  
Yu Cheng\*, Department of University of Pittsburgh  
10:00 **Generalized t-test for Censored Data**  
Mi-Ok Kim\*, Cincinnati Children's Hospital Medical Center

## 26. PANEL DISCUSSION: BAYESIAN METHODS IN CLINICAL TRIALS: LEVERAGING INDUSTRY-ACADEMIC PARTNERSHIPS

TEXAS BALLROOM C, 4TH FLOOR  
Sponsor: ASA Biopharmaceutical Section  
Organizer and Chair: Brad Carlin, University of Minnesota, School of Public Health  
Panelists: Amy Xia, Amgen Corporation  
Stacy Lindborg, Eli Lilly and Company  
Gary Rosner, MD Anderson Cancer Center  
Gene Pennello, Center for Devices and Radiological Health – U.S. Food and Drug Administration

- 8:30 **Introduction by the chair**  
8:35 **Individual presentations by the panelists**  
9:15 **Panel discussion**  
10:00 **Floor discussion**

## MONDAY, MARCH 16

10:15—10:30 a.m.  
**REFRESHMENT BREAK AND VISIT THE EXHIBITORS**  
TEXAS BALLROOM PRE-FUNCTION AREA, 4TH FLOOR

10:30 a.m.—12:15 p.m.

## 27. RECENT ADVANCEMENTS IN LONGITUDINAL ANALYSIS

TEXAS BALLROOM C, 4TH FLOOR  
Sponsor: ASA Biopharmaceutical Section  
Organizer: Erning Li, Texas A&M University  
Chair: Naisiyin Wang, Texas A&M University

- 10:30 **Joint Modeling of Longitudinal Categorical Data and Survival Data**  
Jianwen Cai\*, Jaeun Choi and Donglin Zeng, University of North Carolina at Chapel Hill  
10:55 **Dropout in Longitudinal Clinical Trials with Binary Outcome**  
Mike G. Kenward\* and Rhian M. Daniel, London School of Hygiene and Tropical Medicine  
11:20 **Variable Selection in Longitudinal Data using Regularized Likelihoods**  
Xihong Lin\*, Harvard School of Public Health  
11:45 **Functional Latent Feature Models for Data with Longitudinal Covariate Processes**  
Erning Li\*, Texas A&M University, Yehua Li, University of Georgia, Nae-Yuh Wang, The Johns Hopkins University School of Medicine, Naisiyin Wang, Texas A&M University  
12:10 **Floor Discussion**

## 28. ADAPTIVE DESIGNS IN PRACTICE: BENEFITS, RISKS AND CHALLENGES

TEXAS BALLROOM E, 4TH FLOOR  
Sponsor: ASA Biopharmaceutical Section  
Organizer: Judith Quinlan, Cytel, Inc.  
Chair: José Pinheiro, Novartis Pharmaceuticals

- 10:30 **Bayesian Adaptive Designs in Medical Device Trials**  
Scott M. Berry, Berry Consultants  
10:55 **Improved Dose Ranging through Adaptive Dose Allocation**  
Judith A. Quinlan, Cytel Inc.  
11:20 **Developing an Adaptive Phase 2/3 Design through Trial Simulation**  
Brenda L. Gaydos, Eli Lilly and Company  
11:45 **Discussant:**  
Telba Irony, Center for Devices and Radiological Health – U.S. Food and Drug Administration  
12:10 **Floor Discussion**

## 29. OUTCOME DEPENDENT SAMPLING

PRESIDIO A, 3RD FLOOR  
Sponsors: ASA Survey Research and Methodology Section, ASA Section on Statistics in Epidemiology, ASA Biometrics Section  
Organizer: Sébastien Haneuse, Group Health Center for Health Studies  
Chair: Sébastien Haneuse, Group Health Center for Health Studies

- 10:30 **On Planning a Retrospective, Outcome Dependent Sampling Study for Longitudinal Binary Response Data**  
Jonathan S. Schildcrout\*, Vanderbilt University, Patrick J. Heagerty, University of Washington  
10:55 **Partial Linear Model for Data from an Outcome Dependent Sampling Design**  
Haibo Zhou\* and Guoyou Qin, University of North Carolina at Chapel Hill  
11:20 **Longitudinal Studies of Binary Response Data Following Case-control and Stratified Case-control Sampling: Design and Analysis**  
Jonathan S. Schildcrout, Vanderbilt University School of Medicine, Paul J. Rathouz\*, Department of Health Studies, University of Chicago  
11:45 **The Analysis of Retrospective Family Studies**  
John Neuhaus\*, University of California-San Francisco, Alastair , Scott and Chris Wild, University of Auckland  
12:10 **Floor Discussion**

## 30. NEW STATISTICAL METHODS IN DETECTING EPISTASIS INTERACTIONS IN GENOME-WIDE ASSOCIATION STUDIES

CROCKETT A/B, 4TH FLOOR  
Sponsor: ASA Biometrics Section  
Organizer: Hongzhe Li, University of Pennsylvania  
Chair: Hongzhe Li, University of Pennsylvania

- 10:30 **Bayesian Detection of Gene-Gene Interactions Associated with Type 1 Diabetes within MHC Region**  
Yu Zhang, Penn State University, Jing Zhang, and Jun S. Liu\*, Harvard University  
11:00 **Genome-wide Strategies for Gene-gene Interaction**  
Dan L. Nicolae\*, The University of Chicago



# SCIENTIFIC PROGRAM

- 11:30 **Bayesian Association Mapping**  
Anders Albrechtsen, University of Copenhagen, Rasmus Nielsen\*, University of California-Berkeley  
12:00 **Floor Discussion**

## 31. ANALYSIS OF MEDICAL COST DATA: JOINT VENTURE OF HEALTH ECONOMISTS AND STATISTICIANS

CROCKETT C, 4TH FLOOR

Sponsors: ASA Biopharmaceutical Section, ASA Health Policy Statistics Section, ASA Section on Statistics in Epidemiology  
Organizer: Lei Liu, University of Virginia  
Chair: Daniel Heitjan , University of Pennsylvania

- 10:30 **A Decomposition of Changes in Medical Care Expenditure Distribution in the US Households: Do We Fare Better Twenty Years After?**  
Ya-Chen T. Shih\*, University of Texas M.D. Anderson Cancer Center"
- 10:55 **Stochastic Models in Cost-Effectiveness Analysis**  
Joseph C. Gardiner\*, Michigan State University, Zhehui Luo, RTI International
- 11:20 **Semi-parametric Models for Longitudinal Cost Data Subject to Incomplete Observation**  
Eleanor M. Pullenayegum\*, McMaster University, Andrew R. Willan, University of Toronto
- 11:45 **A Flexible Two-Part Random Effects Model for Correlated Medical Costs**  
Lei Liu\*, University of Virginia, Mark Cowen, Quality Institute, St. Joseph Mercy Health System, Robert Strawderman, Cornell University, Tina Shih, M. D. Anderson Cancer Center, University of Texas
- 12:10 **Floor Discussion**

## 32. CONTRIBUTED PAPERS: GENETIC DIVERSITY, MUTATIONS AND NATURAL SELECTION

TRAVIS A, 3RD FLOOR

Sponsor: ENAR  
Chair: Yuehua Cui, Michigan State University

- 10:30 **Detecting Natural Selection Across Dependent Populations**  
Eleanne Solorzano\*, University of New Hampshire, Hongyu Zhao, Yale University
- 10:45 **Hierarchical Bayesian Analysis of Genetic Diversity in Geographically Structured Populations**  
Seongho Song\*, University of Cincinnati, Dipak K. Dey and Kent E. Holsinger, University of Connecticut
- 11:00 **DNA Barcoding: Bayesian Discrete Ordered Classification**  
Michael P. Anderson\* and Suzanne R. Dubnicka, Kansas State University
- 11:15 **Joint Bayesian Estimation of Phylogeny and Sequence Alignment**  
Heejung Shim\* and Bret Larget, University of Wisconsin-Madison
- 11:30 **A Statistical Perspective of DNA-protein Cross-links (DPX) Data**  
Martin Klein\* and Bimal Sinha, University of Maryland, Baltimore County

- 11:45 **Characterization of mRNA Secondary Structure using Weibull Random Variable**  
Fisseha Abebe\* and William Seffens, Clark Atlanta University
- 12:00 **Comparing Bayesian and Frequentist Approaches to Estimating Mutation Rates**  
Qi Zheng\*, School of Rural Public Health, Texas A&M Health Science Center

## 33. CONTRIBUTED PAPERS: ESTIMATION METHODS

TRAVIS B, 3RD FLOOR

Sponsor: ENAR  
Chair: Jichun Xie, University of Pennsylvania

- 10:30 **Weighted Likelihood Method for a Linear Mixed Model**  
Tianyue Zhou\*, Sanofi-aventis
- 10:45 **Estimator of the Intensity of a Modulated Poisson Process with a Gamma Prior**  
Benjamin B. Neustifter\* and Stephen L. Rathbun, University of Georgia
- 11:00 **The Use of Extreme Order Statistics to Estimate Standard Deviation**  
Chand K. Chauhan\* and Yvonne M. Zubovic, Indiana Purdue University Fort Wayne Indiana
- 11:15 **The Biased-bootstrap For GMM Models**  
Mihai C. Giurcanu\*, University of Louisiana at Lafayette, Brett D. Presnell, University of Florida
- 11:30 **Biomedical Applications of Convolutions of Mixed Distributions**  
Calvin L. Williams and Charity N. Watson\*, Clemson University
- 11:45 **Sieve Type Deconvolution Estimation of Mixture Distributions with Boundary Effects**  
Mihee Lee\*, University of North Carolina at Chapel Hill, Peter Hall, The University of Melbourne, Haipeng Shen, Christina Burch, Jon Tolle and J. S. Marron, University of North Carolina at Chapel Hill
- 12:00 **A Method for Accelerating the Quadratic Lower-bound Algorithm**  
Aiyi Liu and Chunling Liu\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Man Lai Tang, Hong Kong Baptist University, Guo-liang Tian, University of Hong Kong

## 34. CONTRIBUTED PAPERS: SPATIAL MODELS

BONHAM B, 3RD FLOOR

Sponsor: ASA Health Policy Statistics Section  
Chair: Xuanyao He, University of North Carolina at Chapel Hill

- 10:30 **A Kronecker Product Linear Exponent AR(1) Family Of Correlation Structures**  
Sean L. Simpson\*, Wake Forest University School of Medicine, Lloyd J. Edwards, University of North Carolina at Chapel Hill, Keith E. Muller, University of Florida
- 10:45 **On the Voronoi Estimator for Intensity of an Inhomogeneous Planar Poisson Process**  
Christopher D. Barr\*, Johns Hopkins Univeristy, Frederic P. Schoenberg, University of California, Los Angeles

# SCIENTIFIC PROGRAM

- 11:00 Zero-Inflated Binomial Spatial Models, With Applications To Colon Carcinogenesis**  
Tatiyana V. Apanasovich\*, Thomas Jefferson University, Marc G. Genton and Raymond J. Carroll, Texas A&M University
- 11:15 Spatial Modeling of Air Pollution and Mortality Time Trends in the United States**  
Sonja Greven\*, Francesca Dominici and Scott Zeger, Johns Hopkins University
- 11:30 Spatial Cluster Detection for Weighted Outcomes using Cumulative Geographic Residuals**  
Andrea J. Cook\*, Group Health Center for Health Studies, Yi Li, Harvard School of Public Health and The Dana Farber Cancer Institute, David Arterburn, Group Health Center for Health Studies, Ram C. Tiwari, CDR, U.S. Food and Drug Administration
- 11:45 Adjustments for Local Multiplicity with Scan Statistics**  
Ronald E. Gangnon\*, University of Wisconsin
- 12:00 Improving Disease Surveillance by Incorporating Residential History**  
Justin Manjourides\* and Marcello Pagano, Harvard School of Public Health
- 35. CONTRIBUTED PAPERS: TOXICOLOGY/DOSE-RESPONSE MODELS**  
TRAVIS C, 3RD FLOOR  
Sponsor: ASA Section on Statistics and the Environment  
Chair: Hui Zhang, University of Rochester Medical Center
- 10:30 Robust Statistical Theory and Methodology for Nonlinear Models with Application to Toxicology**  
Changwon Lim\*, University of North Carolina at Chapel Hill, Biostatistics Branch, NIEHS, NIH, Pranab K. Sen, University of North Carolina at Chapel Hill, Shyamal D. Peddada, Biostatistics Branch, NIEHS, NIH, RTP, NC
- 10:45 Incorporating Historical Control Information into Quantal Bioassay with Bayesian Approach**  
Din Chen\*, South Dakota State University
- 11:00 A Comparative Study on Constructing Confidence Bands for Effective Doses**  
Gemechis D. Djira\* and Din Chen, South Dakota State University
- 11:15 Semiparametric Bayes Multiple Testing: Applications to Tumor Data**  
Lianming Wang\*, University of South Carolina, David B. Dunson, Duke University
- 11:30 Testing for Sufficient Similarity in Dose-Response in Complex Chemical Mixtures: Do Interaction and Dose Scale Matter?**  
LeAnna G. Stork\*, Monsanto Co., Scott L. Marshall and Chris Gennings, Virginia Commonwealth University, Linda K. Teuschler and John Lipscomb, National Center for Environmental Assessment, U.S. Environmental Protection Agency, Mike Devito and Kevin Crofton, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency
- 11:45 Investigating Statistical Distance as a Similarity Measure for Determining Sufficient Similarity in Dose-Response in Chemical Mixtures**  
Scott Marshall\* and Chris Gennings, Virginia Commonwealth University LeAnna G. Stork, Monsanto Co., Linda Teuschler, Glenn Rice and John Lipscomb, National Center for Environmental Assessment, U.S. Environmental Protection Agency, Mike DeVito and Kevin Crofton, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency
- 12:00 Proportion of Similar Response (PSR) and Receiver Operating Characteristics (ROC) Methodologies in Assessing Correlates of Protection For Vaccine Efficacy**  
Katherine E.D. Giacoletti\* and Joseph F. Heyse, Merck Research Labs
- 36. CONTRIBUTED PAPERS: CLASSIFICATION/MACHINE LEARNING**  
INDEPENDENCE, 3RD FLOOR  
Sponsor: ENAR  
Chair: Bin Li, Louisiana State University
- 10:30 Classification of Data under Autoregressive Circulant Covariance Structure**  
Christopher L. Louden\* and Anuradha Roy, The University of Texas at San Antonio
- 10:45 A Support Vector Machine Approach for Genome-wide Copy-number-variation Study**  
Shyh-Huei Chen\*, National Yunlin University of Science and Technology, Yunlin, Taiwan, Fang-Chi Hsu, Wake Forest University School of Medicine
- 11:00 On Margin-based Classification Methods**  
Lingsong Zhang\*, Harvard School of Public Health
- 11:15 Grouped LASSO-Patternsearch Algorithm**  
Weiliang Shi\*, GlaxoSmithKline, Grace Wahba, University of Wisconsin-Madison
- 11:30 Performance Guarantee for Individualized Treatment Rules**  
Min Qian\* and Susan A. Murphy, University of Michigan
- 11:45 Nonparametric Classifications of Tumors Using Gene Expression Data Based on the Triangle Data Depth**  
Zhenyu Liu\* and Reza Modarres, The George Washington University
- 12:00 Classification of Self-Modeling Regressions with Unknown Shape Functions**  
Rhonda D. VanDyke\*, Cincinnati Children's Hospital and Department of Pediatrics, University of Cincinnati, Kert Viele and Robin L. Cooper, University of Kentucky
- 37. CONTRIBUTED PAPERS: CLUSTERED SURVIVAL DATA**  
TRAVIS D, 3RD FLOOR  
Sponsor: ENAR  
Chair: Joshua M. Tebbs, University of South Carolina
- 10:30 Marginal Models for Clustered Time to Event Data with Competing Risks using Pseudo-values**  
Brent R. Logan\*, Mei-Jie Zhang and John P. Klein, Medical College of Wisconsin
- 10:45 Competing Risks Regression for Stratified Data**  
Bingqiang Zhou\*, University of North Carolina at Chapel Hill



# SCIENTIFIC PROGRAM

- 11:00** **Statistical Analysis of Clustered Current Status Data**  
Ping Chen\*, University of Missouri-Columbia, Junshan Shen, Beijing University-China, Jianguo Sun, University of Missouri-Columbia
- 11:15** **Parametric Analysis of Interval Censored Data with Informative Cluster Size**  
Xinyan Zhang\* and Jianguo Sun, University of Missouri
- 11:30** **Modeling Survival Data with Alternating States and a Cure Fraction using Frailty Models**  
Yimei Li\*, Paul E. Wileyto and Daniel F. Heitjan, University of Pennsylvania
- 11:45** **Models with Multiple Event Types and their Predictions**  
Kent R. Bailey\*, Mayo Clinic
- 12:00** **Constrained Survival Analysis**  
Yong Seok Park\*, University of Michigan

## MONDAY, MARCH 16

**12:15—1:30 p.m.**  
**ROUNDTABLE LUNCHEONS (REGISTRATION REQUIRED)**  
TEXAS BALLROOM F, 4TH FLOOR

**1:45—3:30 p.m.**

### 38. MARGINS AND MONITORING OF NON-INFERIORITY CLINICAL TRIALS

TEXAS BALLROOM E, 4TH FLOOR  
Sponsor: ASA Biopharmaceutical Section  
Organizer: Craig B. Borkowf, Centers for Disease Control and Prevention  
Chair: Laura Lee Johnson, National Center for Complementary and Alternative Medicine

- 1:45** **Non-inferiority in Orphan Diseases: Can we Improve upon Existing Therapies?**  
Janet Wittes\*, Statistics Collaborative
- 2:15** **Non-inferiority Margin in the Presence of Constancy Violation or Different Patient Populations**  
Sue-Jane Wang\*, H.M. James Hung and Robert T. O'Neill, U.S. Food and Drug Administration
- 2:45** **Monitoring Non-inferiority Trials with Recurrent Event Outcomes over Multiple Treatment Periods**  
Richard Cook\* and Grace Yi, University of Waterloo
- 3:15** **Discussant:**  
Lisa LaVange, University of North Carolina-Chapel Hill

### 39. STATISTICAL ANALYSIS OF INFORMATIVE MISSING DATA

TEXAS BALLROOM C, 4TH FLOOR  
Sponsors: ASA Biopharmaceutical Section, ASA Health Policy Statistics Section  
Organizer and Chair: Ying Yuan, The University of Texas M.D. Anderson Cancer Center

- 1:45** **Every Missing not at Random Model for Incomplete Data has got a Missing at Random Counterpart with Equal Fit**  
Geert Molenberghs\*, Universiteit Hasselt and Katholieke Universiteit Leuven, Michael G. Kenward, London School of Hygiene and Tropical Medicine, United Kingdom  
Geert Verbeke, Caroline Beunckens and Cristina Sotto, Universiteit Hasselt and Katholieke Universiteit Leuven

- 2:15** **Bayesian Semiparametric Selection Models with Application to a Breast Cancer Prevention Trial**  
Chenguang Wang and Michael Daniels\*, University of Florida, Daniel Scharfstein, Johns Hopkins University
- 2:45** **Constructing and Calibrating Informative Priors for Nonidentified Parameters in Models Fit to Incomplete Data**  
Joseph W. Hogan\*, Brown University
- 3:15** **Floor Discussion**

## 40. MODEL-BASED CLUSTERING OF HIGH-DIMENSIONAL GENOMIC DATA

CROCKETT A/B, 4TH FLOOR  
Sponsor: ASA Section on Statistics in Epidemiology  
Organizer: Andres Houseman, Harvard University  
Chair: Michael Wu, Harvard University

- 1:45** **Identifying Cluster Structure and Relevant Variables in High-dimensional Data Sets**  
Mahlet G. Tadesse\*, Georgetown University
- 2:10** **Recursively Partitioned Mixture Models with Applications to DNA Methylation Array Data**  
E. Andres Houseman\* and Brock C. Christensen, Brown University, Ru-Fang Yeh, University of California San Francisco, Carmen J. Marsit, Brown University, Margaret R. Karagas, Dartmouth-Hitchcock Medical Center, Margaret Wrensch, University of California San Francisco, Heather H. Nelson, University of Minnesota School of Public Health, Joseph Wiemels, University of California San Francisco, John K. Wiencke, University of California San Francisco, Karl T. Kelsey, Brown University
- 2:35** **A Latent Class Model With Hidden Markov Dependence for Array CGH Data**  
Stacia M. DeSantis\*, Medical University of South Carolina, E. Andres Houseman, Brown University, Brent A. Coull, Harvard School of Public Health, David N. Louis and MA Gayatri Mohapatra, Massachusetts General Hospital, Rebecca A. Betensky, Harvard School of Public Health
- 3:00** **Transposable Regularized Covariance Models with an Application to High-dimensional Missing Data Imputation**  
Genevera Allen\* and Rob Tibshirani, Stanford University
- 3:25** **Floor Discussion**

## 41. ISSUES IN COMPLICATED DESIGNS AND SURVIVAL ANALYSIS

CROCKETT C, 4TH FLOOR  
Sponsors: ASA Biopharmaceutical Section, IMS  
Organizer: Bin Nan, University of Michigan  
Chair: Bin Nan, University of Michigan

- 1:45** **Statistical Identifiability and the Surrogate Endpoint Problem with Application to Vaccine Trials**  
Julian Wolfson\*, University of Washington and Peter Gilbert, University of Washington and Fred Hutchinson Cancer Research Center
- 2:10** **Multiphase Case-control Sampling Designs**  
Bryan Langholz\* and Ly Thomas, University of Southern California, Rakowski Cyril, Chapman University
- 2:35** **Semiparametric Efficient Estimation in Case-Cohort Study**  
Donglin Zeng\* and Danyu Lin, University of North Carolina

\* Presenter



# SCIENTIFIC PROGRAM

3:00	<b>Estimating the Effect of a Time-dependent Therapy on Restricted Mean Lifetime using Observational Data</b> Douglas E. Schaubel* and John D. Kalbfleisch, University of Michigan	3:15	<b>Optimal Shrinkage Variance Estimation and Outlier Detection in Microarray Data Analysis</b> Nysisia I. George*, U.S. Food and Drug Administration and Naisyin Wang, Texas A&M University
3:25	<b>Floor Discussion</b>		
<b>42. STATISTICAL INFERENCE FOR FOREST INVENTORY AND MONITORING USING REMOTELY SENSED DATA</b> PRESIDIO A, 3RD FLOOR Sponsor: ASA Section on Statistics and the Environment Organizer: Ronald E. McRoberts, U.S. Forest Service Chair: Sudipto Banerjee, University of Minnesota			
<b>1:45 Hierarchical Spatial Models with Remotely Sensed Predictors for Mapping Tree Species Assemblages across Large Domains</b> Andrew O. Finley*, Michigan State University, Sudipto Banerjee, University of Minnesota, Ronald E. McRoberts, Northern Research Station, U.S. Forest Service			
2:15	<b>A Combined Design and Model-based Derivation of the MSE of Estimated Aboveground Biomass from Profiling Airborne Laser System</b> Timothy G. Gregoire*, Yale University, Ross F. Nelson, NASA-Goddard Space Flight Center, Erik Naesset, Norwegian University of Life Sciences, Goran Stahl, Swedish University of Agricultural Sciences, Terje Gobakken, Norwegian University of Life Sciences	2:00	<b>A Likelihood-based Approach for Detecting Gene-Gene Interaction in a Case-control Study</b> Saonli Basu*, University of Minnesota
2:45	<b>Model-based Inference for Natural Resource Inventories</b> Ronald E. McRoberts*, Northern Research Station, U.S. Forest Service	2:15	<b>High-Resolution QTL Mapping via Simultaneous Analysis of Dense Markers</b> Nengjun Yi*, University of Alabama at Birmingham
3:15	<b>Floor Discussion</b>	2:30	<b>Testing for Genetic Main Effects in Presence of Gene-Gene and Gene-Environment Interactions in Genome-Wide Association Studies</b> Arnab Maity* and Xihong Lin, Harvard School of Public Health
<b>43. CONTRIBUTED PAPERS: PRE-PROCESSING AND QUALITY CONTROL FOR HIGH-THROUGHPUT GENOMIC TECHNOLOGIES</b> TRAVIS A, 3RD FLOOR Sponsor: ENAR Chair: Julia Sharp, Clemson University			
1:45	<b>Background Correction Based on the Box-Cox Transformation of Noises for Illumina Bead Array Data</b> Min Chen*, Yale University and Yang Xie, University of Texas Southwestern Medical Center	2:45	<b>Locate Complex Disease Loci by Investigating Gene and Environment Interaction for Genome-Wide Association Studies</b> Jin Zheng* and Goncalo R. Abecasis, University of Michigan
2:00	<b>Background Adjustment for DNA Microarrays using a Database of Microarray Experiments</b> Yunxia Sui* and Zhijin Wu, Brown University, Xiaoyue Zhao, Bionovo Inc.	3:00	<b>A General Framework for Estimating Genetic Effects and Gene-Environment Interactions with Missing Data</b> Yijuan Hu*, Danyu Lin and Donglin Zeng, University of North Carolina-Chapel Hill
2:15	<b>Statistical Metrics for Quality Assessment of High Density Tiling Array Data</b> Hui Tang*, Mayo Clinic and Terence Speed, University of California at Berkeley	3:15	<b>Risk Effect Estimation for Multiple Phenotypes and Gene-environment Interaction: A Conditional Likelihood Approach</b> Arpita Ghosh*, Fei Zou and Fred A. Wright, University of North Carolina-Chapel Hill
2:30	<b>The Effects of Missing Imputation on Various Down-stream Analyses in Microarray Experiments</b> Sunghee Oh* and George C. Tseng, University of Pittsburgh		
2:45	<b>A Novel Test for Quality Control in Family-based Genome-wide Association Studies</b> David Fardo*, University of Kentucky, Juliana Ionita-Laza and Christoph Lange, Harvard School of Public Health		
3:00	<b>Statistical Inference for Pooled Samples in Next Generation Sequencing</b> Justin W. Davis*, University of Missouri		
<b>44. CONTRIBUTED PAPERS: ASSESSING GENE AND ENVIRONMENT INTERACTIONS IN GENOME-WIDE STUDIES</b> INDEPENDENCE, 3RD FLOOR Sponsor: ASA Section on Statistics and the Environment Chair: Hemant K. Tiwari, University of Alabama at Birmingham			
1:45	<b>Detecting Gene-Gene Interaction via Optimally Weighted Markers</b> Jing He* and Mingyao Li, University of Pennsylvania School of Medicine		
2:00	<b>A Likelihood-based Approach for Detecting Gene-Gene Interaction in a Case-control Study</b> Saonli Basu*, University of Minnesota		
2:15	<b>High-Resolution QTL Mapping via Simultaneous Analysis of Dense Markers</b> Nengjun Yi*, University of Alabama at Birmingham		
2:30	<b>Testing for Genetic Main Effects in Presence of Gene-Gene and Gene-Environment Interactions in Genome-Wide Association Studies</b> Arnab Maity* and Xihong Lin, Harvard School of Public Health		
2:45	<b>Locate Complex Disease Loci by Investigating Gene and Environment Interaction for Genome-Wide Association Studies</b> Jin Zheng* and Goncalo R. Abecasis, University of Michigan		
3:00	<b>A General Framework for Estimating Genetic Effects and Gene-Environment Interactions with Missing Data</b> Yijuan Hu*, Danyu Lin and Donglin Zeng, University of North Carolina-Chapel Hill		
3:15	<b>Risk Effect Estimation for Multiple Phenotypes and Gene-environment Interaction: A Conditional Likelihood Approach</b> Arpita Ghosh*, Fei Zou and Fred A. Wright, University of North Carolina-Chapel Hill		
<b>45. CONTRIBUTED PAPERS: HYPOTHESIS TESTING</b> TRAVIS B, 3RD FLOOR Sponsor: ENAR Chair: Wonkuk Kim, University of South Florida			
1:45	<b>Penalized Likelihood Ratio Test When Some Parameters are Present Only Under the Alternative</b> Chongzhi Di*, Kung-Yee Liang and Ciprian M. Crainiceanu, Johns Hopkins University		
2:00	<b>Likelihood Ratio Test for Qualitative Interactions</b> Qing Pan*, George Washington University		
2:15	<b>Using Multiple Control Groups as Evidence About Unobserved Biases in an Observational Study of Treatments for Melanoma</b> Frank B. Yoon, Phyllis A. Gimotty, DuPont Guerry and Paul R. Rosenbaum, University of Pennsylvania		



# SCIENTIFIC PROGRAM

- 2:30 **An Efficient Rank-based Test for the Generalized Nonparametric Behrens-Fisher Problem**  
Kai F. Yu\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Qizhai Li, National Cancer Institute, Aiyyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Kai Yu, National Cancer Institute
- 2:45 **Nonparametric Procedures for Comparing Correlated Multiple Endpoints with Applications to Oxidative Stress Biomarkers**  
Aiyyi Liu\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Chunling Liu, Enrique Schisterman, Harvard School of Public Health
- 3:00 **On Global P-value Calculation in Multi-stage Designs**  
Shanhong Guan\*, Merck & Co.
- 3:15 **Insights into p-values and Bayes Factors via False Positive and False Negative Bayes Factors**  
Hormuzd A. Katki\*, National Cancer Institute

## 46. CONTRIBUTED PAPERS: VARIABLE SELECTION METHODS

TRAVIS C, 3RD FLOOR  
Sponsor: ENAR  
Chair: Michael Swartz, University of Texas M.D. Anderson Cancer Center

- 1:45 **Variable Selection for Identifying Environmental Contaminants Associated with Human Fecundity**  
Sungduk Kim\*, Rajeshwari Sundaram and Germaine M. Louis, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
- 2:00 **Bayesian Variable Selection for Latent Class Models**  
Joyee Ghosh\* and Amy H. Herring, University of North Carolina at Chapel Hill
- 2:15 **Regularization Parameter Selections via Generalized Information Criterion**  
Yiyun Zhang\* and Runze Li, Penn State University, Chih-Ling Tsai, University of California, Davis
- 2:30 **Penalized Estimating Equations for Semiparametric Linear Transformation Models**  
Hao Zhang and Wenbin Lu\*, North Carolina State University, Hansheng Wang, Peking University
- 2:45 **Regularized Estimation in AFT Models with High-dimensional Covariates**  
Liping Huang\*, Mai Zhou and Arne C. Bathke, University of Kentucky
- 3:00 **Variable Selection for the Cox Regression Model with Covariates Missing at Random**  
Ramon Garcia\*, University of North Carolina
- 3:15 **Bayesian Semiparametric Frailty Selection in Multivariate Event Time Data**  
Bo Cai\*, University of South Carolina

## 47. CONTRIBUTED PAPERS: LONGITUDINAL DATA ANALYSIS

TRAVIS D, 3RD FLOOR  
Sponsor: ASA Health Policy Statistics Section  
Chair: Abdus Sattar, University of Pittsburgh

- 1:45 **Three-level Mixed Effects Location Scale Model**  
Eisuke Segawa\*, Donald Hedeker and Robin J. Mermelstein, University of Illinois at Chicago
- 2:00 **Nonparametric Modeling of Semi-continuous Data with Application to Medical Cost Data Analysis**  
Pang Du\*, Virginia Tech, Lei Liu, University of Virginia, Anna Liu, University of Massachusetts-Amherst
- 2:15 **Semiparametric Analysis of Multivariate Recurrent and Terminal Events**  
Liang Zhu\*, St. Jude Children's Research Hospital and Jianguo Sun, University of Missouri-Columbia
- 2:30 **Determining When Time Response Curves Differ in the Presence of Censorship with Application to a Rheumatoid Arthritis Biomarker Study**  
Ann A. Lazar\*, Harvard School of Public Health & Dana-Farber Cancer Institute and Gary O. Zerbe, University of Colorado, Denver
- 2:45 **SAS/IML for Parameter Estimation of Logistic Regression for Transition, Reverse Transition and Repeated Transition from Follow-up Data**  
Rafiqul I. Chowdhury, Kuwait University, Kuwait, M. A. Islam, Dhaka University, Bangladesh, Shahriar S. Huda, Kuwait University, Kuwait
- 3:00 **A Composite Likelihood Approach to the Analysis of Longitudinal Clonal Data on Multitype Cellular Systems under an Age-dependent Branching Process**  
Rui Chen and Ollivier Hyrien, University of Rochester Medical Center
- 3:15 **The Univariate Approach to Repeated Measures ANOVA for High Dimension, Low Sample Size**  
Yueh-Yun Chi\* and Keith E. Muller, University of Florida

## 48. CONTRIBUTED PAPERS: MULTIPLE TESTING IN HIGH-DIMENSIONAL DATA

BONHAM B, 3RD FLOOR  
Sponsor: ENAR  
Chair: Stanley Pounds, St. Jude Children's Research Hospital

- 1:45 **Estimation of False Discovery Rate Using Permutation P-Values with Different Discrete Distributions**  
Tim Bancroft\* and Dan Nettleton, Iowa State University
- 2:00 **Controlling False Discoveries in Multidimensional Directional Decisions, with Applications to Gene Expression Data on Ordered Categories**  
Wenge Guo\*, National Institute of Environmental Health Sciences, Sanat K. Sarkar, Temple University, Shyamal D. Peddada, National Institute of Environmental Health Sciences
- 2:15 **Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks**  
Wenguang Sun\*, North Carolina State University and Tony Cai, University of Pennsylvania
- 2:30 **Finding Critical Value for t-Tests in Very High Dimensions**  
Hongyuan Cao\* and Michael R. Kosorok, University of North Carolina-Chapel Hill

# SCIENTIFIC PROGRAM

- 2:45 **Multiplicity-Calibrated Bayesian Hypothesis Tests**  
Mengye Guo\* and Daniel F. Heitjan, University of Pennsylvania
- 3:00 **Spike and Slab Dirichlet Prior for Bayesian Multiple Testing in Random Effects Models**  
Sinae Kim\*, University of Michigan, David B. Dahl, Texas A&M University, Marina Vannucci, Rice University
- 3:15 **Computation of Exact p-values for Nonparametric Test**  
Yuanhui Xiao\*, Georgia State University

## MONDAY, MARCH 16

3:30—3:45 p.m.  
**REFRESHMENT BREAK AND VISIT THE EXHIBITORS**  
TEXAS BALLROOM F, 4TH FLOOR

3:45—5:30 p.m.

### 49. ROLE OF META-ANALYSIS IN DRUG DEVELOPMENT

TEXAS BALLROOM E, 4TH FLOOR

Sponsor: ASA Biopharmaceutical Section  
Organizer: Sue-Jane Wang, U.S. Food and Drug Administration  
Chair: Sue-Jane Wang, U.S. Food and Drug Administration

- 3:45 **The Use of Cumulative Meta Analysis in Drug Development**  
Kuang-Kuo G. Lan\*, Johnson & Johnson PRD
- 4:15 **Utility and Pitfalls of Meta Analysis for Designing Non-Inferiority Trial**  
H.M. James Hung\*, U.S. Food and Drug Administration
- 4:45 **Are Things Really As Un-Rosi As They Appear?**  
Michael A. Proschan\*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
- 5:15 **Discussant:**  
David L. DeMets, University of Wisconsin

### 50. ANALYSIS OF HIGH-DIMENSIONAL DATA WITH BIOLOGICAL APPLICATIONS

TEXAS BALLROOM C, 4TH FLOOR

Sponsors: IMS  
Organizer: Jianqing Fan, Princeton University  
Chair: Jianqing Fan, Princeton University

- 3:45 **Maximum Likelihood Estimation of a Multidimensional Log-concave Density**  
Richard Samworth, Madeleine Cule and Robert Gramacy, University of Cambridge, Michael Stewart, University of Sydney
- 4:10 **Forward-Lasso Adaptive Shrinkage**  
Gareth James\* and Peter Radchenko, University of Southern California
- 4:35 **Partial Correlation Estimation by Joint Sparse Regression Models**  
Ji Zhu\*, University of Michigan, Jie Peng, University of California, Davis, Pei Wang, Fred Hutchinson Cancer Center, Nengfeng Zhou, University of Michigan
- 5:00 **Estimation in Additive Models with Highly Correlated Covariates**  
Jiancheng Jiang, University of North Carolina at Charlotte, Yingying Fan\*, University of Southern California, Jianqing Fan, Princeton University

### 51. ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE OBSERVATION AND DROPOUT PROCESSES

CROCKETT C, 4TH FLOOR  
Sponsors: ASA Biopharmaceutical Section, ASA Health Policy Statistics Section  
Organizer: Xin He, The Ohio State University  
Chair: Mei-Ling Ting Lee, The Ohio State University

- 3:45 **Regression Analysis of Longitudinal Data with Dependent Observation Process**  
(Tony) Jianguo Sun\*, University of Missouri-Columbia
- 4:10 **Analysis of Longitudinal Data with Informative Dropout Time from an Extended Hazards Model**  
Yi-Kuan Tseng, National Central University, Taiwan, Meng Mao and Jane-Ling Wang\*, University of California, Davis
- 4:35 **Marginal Analysis of Longitudinal Data with both Response and Covariates Subject to Missingness**  
Grace Y. Yi\*, Baojiang Chen and Richard Cook, University of Waterloo
- 5:00 **Semiparametric Regression Analysis of Longitudinal Data with Informative Observation Times**  
Jianguo (Tony) Sun, University of Missouri-Columbia, Do-Hwan Park\*, University of Maryland-Baltimore County, Liuquan Sun, Chinese Academy of Sciences, Xingqiu Zhao, Hong Kong Polytechnic University

### 52. ADDRESSING KEY STATISTICAL ISSUES IN ENVIRONMENTAL EPIDEMIOLOGY

CROCKETT A/B, 4TH FLOOR  
Sponsors: Section on Statistics and the Environment, ASA Section on Statistics in Epidemiology  
Organizer: Chris Paciorek, Harvard University  
Chair: Brent Coull, Harvard University

- 3:45 **Multistage Sampling for Latent Variable Models in Environmental and Genetic Epidemiology**  
Duncan C. Thomas\*, University of Southern California
- 4:15 **Adjustment Uncertainty in Effect Estimation**  
Ciprian M. Crainiceanu\*, Johns Hopkins University
- 4:45 **Bias and Spatial Scale in Models with Spatial Confounding**  
Christopher J. Paciorek\*, Harvard School of Public Health
- 5:15 **Floor Discussion**

### 53. RECENT DEVELOPMENT OF QUANTILE REGRESSION METHODS FOR SURVIVAL DATA

PRESIDIO A, 3RD FLOOR  
Sponsor: ENAR  
Organizers: Limin Peng and Yijian Huang, Emory University  
Chair: Wenbin Lu, North Carolina State University

- 3:45 **Quantile Regression for Doubly Censored Data**  
Guixian Lin, Xuming He\* and Stephen Portnoy, University of Illinois
- 4:10 **Locally Weighted Censored Quantile Regression**  
Huixia Judy Wang\*, North Carolina State University and Lan Wang, University of Minnesota
- 4:35 **Quantile Regression with Censored Data**  
Yijian Huang\*, Emory University



# SCIENTIFIC PROGRAM

5:00	<b>Competing Risks Quantile Regression</b> Limin Peng*, Rollins School of Public Health, Emory University and Jason P. Fine, University of North Carolina at Chapel Hill	4:45	<b>DualKS: Defining Gene Sets with Tissue Set Enrichment Analysis</b> Eric J. Kort, Van Andel Research Institute, Yarong Yang*, Northern Illinois University, Zhongfa Zhang and Bin Teh, Van Andel Research Institute, Nader Ebrahimi, Northern Illinois University
<b>54. CONTRIBUTED PAPERS: ADAPTIVE DESIGN IN CLINICAL TRIALS</b> INDEPENDENCE, 3RD FLOOR Sponsor: ASA Biopharmaceutical Section Chair: Tamekia Jones, University of Florida			
3:45	<b>Adaptive Group Sequential Design in Clinical Trials with Changing Patient Populations</b> Huabao Feng* and Qing Liu, Johnson & Johnson, Jun Shao, University of Wisconsin-Madison	5:00	<b>Evaluation of a Classifier Performance at Various Cutoffs of Gene Selection in Microarray Data with Time-to-event Endpoint</b> Dung-Tsa Chen*, Moffitt Cancer Center & Research Institute, University of South Florida, Ying-Lin Hsu and Tzu-Hsin Liu, National Chung Hsing University, Taichung, Taiwan, James J. Chen, National Center for Toxicological Research, U.S. Food and Drug Administration, Timothy Yeatman, Moffitt Cancer Center & Research Institute, University of South Florida
4:00	<b>Adaptive Penalized D-optimal Designs for Dose Finding for Continuous Bivariate Outcomes</b> Krishna Padmanabhan*, Wyeth Research, Francis Hsuan, Temple University, Vladimir Dragalin, Wyeth Research	5:15	<b>Does a Gene Expression Classifier have a Clinical Value?</b> Samir Lababidi*, U.S. Food and Drug Administration
4:15	<b>Dose finding by Jointly Modeling Toxicity and Efficacy as Time-to-Event Outcomes</b> Ying Yuan and Guosheng Yin*, M. D. Anderson Cancer Center	<b>56. CONTRIBUTED PAPERS: IMAGE ANALYSIS</b> TRAVIS A, 3RD FLOOR Sponsor: ASA Biometrics Section Chair: Ray Hoffmann, Medical College of Wisconsin	
4:30	<b>Bayesian Adaptive Randomization Designs versus Frequentist Designs for Targeted Agent Development</b> Xuemin Gu*, Suyu Liu and Jack J. Lee, M.D. Anderson Cancer Center	3:45	<b>On the Merits of Voxel-based Morphometric Path-analysis for Investigating Volumetric Mediation of a Toxicant's Influence on Cognitive Function</b> Shu-chih Su* Merck & Co., Inc. and Brian Caffo, Johns Hopkins Bloomberg School of Public Health
4:45	<b>Stopping Boundaries of Flexible Sample Size Design with Flexible Trial Monitoring - A Unified Approach</b> Yi He and Zhenming Shun, Sanofi-aventis, Yijia Feng*, Penn State University	4:00	<b>Intrinsic Regression Models for Medial Representation of Subcortical Structures</b> Xiaoyan Shi*, Hongtu Zhu and Joseph G. Ibrahim, University of North Carolina at Chapel Hill, Faming Liang, Texas A&M University, Martin Styner, University of North Carolina at Chapel Hill
5:00	<b>A Surrogate: Primary Replacement Algorithm for Response-adaptive Randomization in Stroke Clinical Trials</b> Amy Nowacki*, Cleveland Clinic Foundation	4:15	<b>Multiscale Adaptive Regression Models for Imaging Data</b> Yimei Li*, Hongtu Zhu, Joseph G. Ibrahim and Dinggang Shen, University of North Carolina-Chapel Hill
5:15	<b>Bayesian Nonparametric Emax Model</b> Haoda Fu*, Eli Lilly and Company	4:30	<b>Connectivity Analysis Based on fMRI and DTI Brain Imaging Data</b> Shuo Chen*, DuBois Bowman and Gordana Derado, Rollins School of Public Health, Emory University
<b>55. CONTRIBUTED PAPERS: GENE SELECTION IN DNA MICROARRAY STUDIES</b> BONHAM B, 3RD FLOOR Sponsor: ASA Biometrics Section Chair: Sinae Kim, University of Michigan			
3:45	<b>An Integrative Analysis Approach for Identification of Genes Associated with Multiple Cancers</b> Shuangge Ma*, Yale University	4:45	<b>Modeling the Spatial and Temporal Dependence in fMRI Data</b> Gordana Derado*, Emory University
4:00	<b>Association Pattern Testing: A Powerful Statistical Tool to Identify Biologically Interesting Genomic Variables</b> Stanley B. Pounds*, Cheng Cheng, Xueyuan Cao, James R. Downing, Raul C. Ribeiro and Kristine R. Crews, St. Jude Children's Research Hospital, Jatinder Lamba, University of Minnesota	5:00	<b>Approximation of the Geisser-Greenhouse Sphericity Estimator and its Application to Analyzing Diffusion Tensor Imaging Data</b> Meagan E. Clement*, Rho, Inc., David Couper, University of North Carolina-Chapel Hill, Keith E. Muller, University of Florida, Hongtu Zhu, University of North Carolina-Chapel Hill
4:15	<b>Incorporating Gene Effects into Parametric Empirical Bayes Methods for Microarrays</b> Steven P. Lund* and Dr. Dan Nettleton, Iowa State University	<b>57. CONTRIBUTED PAPERS: SURVEY RESEARCH</b> TRAVIS B, 3RD FLOOR Sponsor: ASA Survey Research and Methodology Section Chair: Stuart Gansky, University of California at San Francisco	
4:30	<b>Network-based Support Vector Machine for Classification of Microarray Samples</b> Yanni Zhu*, Xiaotong Shen and Wei Pan, University of Minnesota	3:45	<b>Bayesian Inference of Finite Population Distribution Functions and Quantiles from Unequal Probability Samples</b> Qixuan Chen*, Michael R. Elliott and Roderick J.A. Little, University of Michigan School of Public Health

 Student Award Winner

\* Presenter

# SCIENTIFIC PROGRAM

- 4:00 Application of Nonparametric Percentile Regression to Body Mass Index Percentile Curves from Survey Data**  
Yan Li\*, Barry I. Graubard and Edward L. Korn, National Cancer Institute
- 4:15 Optimal Coefficients for Simple Random Sampling Without Replacement using Godambe's General Linear Estimator**  
Ruitao Zhang\* and Ed Stanek, Department of Public Health, University of Massachusetts-Amherst
- 4:30 Should Auxiliary Variables with Measurement Error be used in the Estimation of Population Mean Based on Survey Samples?**  
Wenjun Li\*, University of Massachusetts Medical School and Edward J. Stanek III, Department of Public Health, University of Massachusetts-Amherst
- 4:45 Proxy Pattern-Mixture Analysis for Survey Nonresponse**  
Rebecca R. Andridge\* and Roderick J. Little, University of Michigan
- 5:00 Multiple Imputation Methods for Disclosure Limitation in Longitudinal Data**  
Di An\*, Merck Research Laboratories, Merck & Co., Inc. Roderick J.A. Little and James W. McNally, University of Michigan
- 5:15 Impact of Multi-level Measurement Errors in Survey Data**  
Jianjun Gan\* and Hongmei Zhang, School of Public Health, University of South Carolina
- 58. CONTRIBUTED PAPERS: MEASUREMENT ERROR MODELS**  
TRAVIS C, 3RD FLOOR  
Sponsor: ASA Section on Statistics and the Environment  
Chair: Song Zhang, University of Texas, Southwestern Medical Center
- 4:45 Dichotomized Mismeasured Predictors in Regression Models**  
Loki Natarajan\*, University of California at San Diego
- 4:00 Augmenting Instrumental Variables Estimators in a Two-Stage Design**  
Tor D. Tosteson\*, Dartmouth Medical School
- 4:15 Regression Analysis on a Covariate with Heteroscedastic Measurement Error**  
Ying Guo\* and Roderick Little, University of Michigan
- 4:30 Cox Models with Smooth Functional Effect of Covariates Measured with Error**  
Yu-Jen Cheng\* and Ciprian Crainiceanu, Johns Hopkins University
- 4:45 Underreporting in the Generalized Poisson Regression Model**  
Mavis Pararai\*, Indiana University of Pennsylvania
- 5:00 Measurement Error in Longitudinal Data without Validation Samples**  
Ruifeng Xu\*, Merck & Co., Inc., Jun Shao and Mari Palta, University of Wisconsin – Madison, Zhiguo Xiao, School of Management, Fudan University
- 5:15 Modeling Heaping in Longitudinal Self-reported Cigarette Counts**  
Hao Wang\* and Daniel F. Heitjan, University of Pennsylvania
- 59. CONTRIBUTED PAPERS: MIXTURE MODELS**  
TRAVIS D, 3RD FLOOR  
Sponsor: ASA Health Policy Statistics Section  
Chair: Michael LaValley, Boston University
- 3:45 Confounding and Bias from Intermediate Variables, and a Joint Model for Birthweight and Gestational Age Addressing Them**  
Scott L. Schwartz\* and Alan E. Gelfand, Duke University, Marie L. Miranda, Duke University
- 4:00 Using Finite Multivariate Mixtures to Model Adverse Birth Outcomes**  
Matthew W. Wheeler\* and Amy Herring, University of North Carolina-Chapel Hill, Eric Kalendra, Montse Fuentes and Brian Reich, North Carolina State University
- 4:15 A Mixture Model for the Analysis of Correlated Binomial Data**  
N. Rao Chaganty\*, Old Dominion University, Yihao Deng, Indiana University-Purdue University Fort Wayne, Roy Sabo, Virginia Commonwealth University
- 4:30 Latent Transition Models to Study Change in Dietary Patterns over Time**  
Daniela Sotres-Alvarez\*, Amy H. Herring and Anna Maria Siega-Riz, University of North Carolina-Chapel Hill
- 4:45 Conditional Assessment of Zero-inflated Mixture Models**  
Yan Yang\*, Arizona State University and Doug G. Simpson, University of Illinois at Urbana-Champaign
- 5:00 Bayesian Estimation of Multilevel Mixture Models**  
Tihomir Asparouhov\*, Mplus and Bengt Muthén, UCLA
- 5:15 Smooth Density Estimation with Moment Constraints Using Mixture Densities**  
Ani Eloyan\* and Sujit K. Ghosh, North Carolina State University

## TUESDAY, MARCH 17

8:30—10:15 a.m.

## 60. STATISTICAL ISSUES IN BRAIN IMAGING

TEXAS BALLROOM A, 4TH FLOOR

Sponsor: ENAR

Organizer: Hongtu Zhu, University of North Carolina, Chapel Hill  
Chair: Ying Guo, Emory University

- 8:30 On Combining and Contrasting Brains**  
Nicole A. Lazar\*, University of Georgia
- 8:55 Analyzing fMRI Data with Unknown Brain Activation Profiles**  
Martin A. Lindquist\*, Lucy F. Robinson and Tor D. Wager, Columbia University
- 9:20 Statistical Analysis of Brain Morphometric Measures on Riemannian Manifold**  
Hongtu Zhu\*, Joseph G. Ibrahim, Yimei Li, Weili Lin and Yasheng Cheng, University of North Carolina-Chapel Hill
- 9:45 Tiling Manifolds with Orthonormal Basis**  
Moo K. Chung\*, University of Wisconsin-Madison
- 10:10 Floor Discussion**



# SCIENTIFIC PROGRAM

## 61. CAUSAL INFERENCE WITH INSTRUMENTAL VARIABLES METHODS

PRESIDIO B, 3RD FLOOR

Sponsors: ASA Section on Statistics in Epidemiology, ASA Biometrics Section, ASA Health Policy Statistics Section  
Organizer: Jing Cheng, University of Florida  
Chair: Dylan Small, University of Pennsylvania

### 8:30 A Nonparametric Approach to Instrumental Variables Analysis with Binary Outcomes

Michael Baiocchi\*, Paul Rosenbaum and Dylan Small, University of Pennsylvania

### 9:00 Semiparametric Estimation and Inference for Distributional and General Treatment Causal Effects

Jing Cheng\*, University of Florida College of Medicine, Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Biao Zhang, University of Toledo

### 9:30 Extended Instrumental Variables Estimation for Overall Effects

Marshall M. Joffe\*, Dylan Small, Thomas Ten Have, Steven Brunelli and Harold I. Feldman, University of Pennsylvania

### 10:00 Floor Discussion

## 62. FROM THEORY TO PRACTICE: EXAMPLES AND DISCUSSION OF SOFTWARE DEVELOPMENT FOR WIDE DISSEMINATION OF STATISTICAL METHODOLOGY

CROCKETT C, 4TH FLOOR

Sponsors: ASA Section on Teaching Statistics in the Health Sciences, ASA Section on Statistical Education  
Organizer: Justine Shults, University of Pennsylvania  
Chair: Justine Shults, University of Pennsylvania

### 8:30 Software Development for GEE and ORTH

John S. Preisser\* and Bahjat F. Qaqish, University of North Carolina at Chapel Hill

### 9:00 SAS MACRO QIF: Transition of Methodology Research to Application

Peter X. Song\*, University of Michigan School of Public Health

### 9:30 Derivation and Software Implementation of the Canonical Negative Binomial Model

Joseph M. Hilbe\*, Arizona State University

### 10:00 Discussant:

Anthony Rossini, Novartis Pharma AG

## 63. BAYESIAN METHODS IN PROTEIN BIOINFORMATICS

TEXAS BALLROOM B, 4TH FLOOR

Sponsor: ENAR

Organizer: Abel Rodriguez, University of California Santa Cruz  
Chair: Abel Rodriguez, University of California Santa Cruz

### 8:30 Modeling the Joint Distribution of Pairs of Dihedral Angles for Protein Structure Prediction

David B. Dahl\*, Texas A&M University, Ryan Day, Jerry W. Tsai, University of the Pacific

### 8:55 Bayesian Nonparametric Analysis of Site-specific Selection Effects in Serially DNA Sequences

Daniel Merl\*, Duke University, Raquel Prado and Athanasios Kottas, University of California, Santa Cruz

## 9:20 Model-based Validation of Protein-Protein Interactions in Large-Scale Affinity Purification-Mass Spectrometry Experiments

Hyungwon Choi\*, University of Michigan, Ashton Breitkreutz, Brett Larsen and Anne-Claude Gingras, Samuel Lunenfeld Research Institute, Mount Sinai Hospital Toronto, Mike Tyers, Wellcome Trust Centre for Cell Biology, University of Edinburgh, UK, Zhaohui S. Qin and Alexey I. Nesvizhskii, University of Michigan

## 9:45 Bayesian Analysis of Molecular Forcefields

Scott C. Schmidler\*, Duke University

## 10:10 Floor Discussion

## 64. ADAPTIVE DESIGNS FOR EARLY-PHASE CLINICAL TRIALS

TEXAS BALLROOM C, 4TH FLOOR

Sponsor: ASA Biopharmaceutical Section

Organizer: Yuan Ji, The University of Texas M. D. Anderson Cancer Center

Chair: Yisheng Li, The University of Texas M. D. Anderson Cancer Center

### 8:30 Sequential Implementation of Stepwise Procedures for Identifying the Maximum Tolerated Dose

Ken Cheung\*, Columbia University

### 9:00 Adaptive Randomization for Multi-arm Comparative Clinical Trials Based on Joint Efficacy/Toxicity Outcome

B. Nebiyou Bekele\* and Yuan Ji, M. D. Anderson Cancer Center

### 9:30 Finding the Dose with the Best Efficacy/Tolerability Profile

Anastasia Ivanova\*, University of North Carolina at Chapel Hill

## 10:00 Floor Discussion

## 65. MEMORIAL SESSION: THE LIFE OF DAVID BEATTY DUNCAN: BIOSTATISTICIAN, MENTOR, AND GENTLEMAN

CROCKETT D, 4TH FLOOR

Sponsor: ENAR

Organizers: Gene A. Pennello, Food and Drug Administration and Dennis O. Dixon, National Institute of Allergy and Infectious Diseases

Chair: Karen Bandeen-Roche, Johns Hopkins University

Speakers: Jay Herson, Independent Consultant and Johns Hopkins University; Gene A. Pennello, Food and Drug Administration; Dennis O. Dixon, National Institute of Allergy and Infectious Diseases; Kenneth M. Rice, University of Washington

## 66. CONTRIBUTED PAPERS: SPATIAL ANALYSES OF ALCOHOL, ILLEGAL DRUGS, VIOLENCE AND RACE

TRAVIS A, 3RD FLOOR

Sponsors: ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section

Chair: Margaret Short, University of Alaska-Fairbanks

### 8:30 Geospatial Models of Alcohol, Drugs and Violence

Dennis M. Gorman\*, School of Rural Public Health, Texas A&M Health Science Center

# SCIENTIFIC PROGRAM

- 8:45 Modeling the Methamphetamine Epidemic in California**  
Paul J. Gruenewald\* and William R. Ponicki, Prevention Research Center Pacific Institute for Research and Evaluation
- 9:00 Spatial Relationships between the Substance Use Environment and Child Maltreatment**  
Bridget J. Freisthler\*, UCLA
- 9:15 Solving the Misalignment Problem in the Analysis of Drug Issues**  
Li Zhu\*, Texas A&M Health Science Center, Lance Waller, Emory University, Paul Gruenewald, Prevention Research Center
- 9:30 Varying Parameter Models in Space and Time**  
William R. Ponicki\*, Prevention Research Center, Pacific Institute for Research and Evaluation and Lance A. Waller, Rollins School of Public Health, Emory University
- 9:45 Hierarchical Additive Modeling of Nonlinear Association with Spatial Correlations - An Application to Relate Alcohol Outlet Density and Neighborhood Assault Rates**  
Qingzhao Yu\*, Bin Li and Richard Scribner, Louisiana State University
- 10:00 Spatial Association Between Racial and Social Factors in Determining Low Birth Weight and Preterm Birth, New York, 1995-2005**  
Eric Kalendra\*, Montserrat Fuentes and Brian Reich, North Carolina State University, Amy Herring and Matthew Wheeler, University of North Carolina

## 67. CONTRIBUTED PAPERS: INCORPORATING EXTERNAL KNOWLEDGE IN THE ANALYSIS OF GENOMIC DATA

INDEPENDENCE, 3RD FLOOR

Sponsor: ASA Biometrics Section

Chair: Laila M Poisson, University of Michigan

- 8:30 Incorporation of Prior Knowledge from Linkage Studies into Genetic Association Studies**  
Brooke L. Fridley\*, Daniel Serie, Kristin L. White, Gregory Jenkins, William Bamlet and Ellen L. Goode, Mayo Clinic
- 8:45 Testing Untyped SNPs in Case-Control Association Studies with Related Individuals**  
Zuoheng Wang\* and Mary Sara McPeek, University of Chicago
- 9:00 Multiple SNP-based Approach for Genome-wide Case-control Association Study**  
Min Zhang\*, Yanzhu Lin, Libo Wang, Vitara Pungpapong, James C. Fleet and Dabao Zhang, Purdue University
- 9:15 Mining Pathway-based SNP Sets in GWAS Study with Sparse Logistic Regression**  
Lin S. Chen\*, Ulrike Peters and Li Hsu, Fred Hutchinson Cancer Research Center
- 9:30 Bayesian Modeling of Pharmacogenetics Data**  
Donatello Telesca\*, Gary L. Rosner and Peter Muller, University of Texas M.D. Anderson Cancer Center
- 9:45 Network Based Gene Set Analysis Under Temporal Correlation**  
Ali Shojaie\* and George Michailidis, University of Michigan
- 10:00 Network-based Genomic Discovery: Application and Comparison of Markov Random Field Models**  
Peng Wei\* and Wei Pan, School of Public Health, University of Minnesota

## 68. CONTRIBUTED PAPERS: ROC ANALYSIS

TRAVIS B, 3RD FLOOR

Sponsor: ENAR

Chair: Stephen L Hillis, University of Iowa

- 8:30 The Case for FROC Analysis**  
Dev P. Chakraborty\*, University of Pittsburgh
- 8:45 Bayesian Nonparametric Combination of Multiple Diagnostic Measurements**  
Lorenzo Trippa\*, MD Anderson Cancer Center
- 9:00 Optimal Cutpoint Estimation with Censored Data**  
Mithat Gonen and Cami Sima\*, Memorial Sloan-Kettering Cancer Center
- 9:15 Semiparametric ROC Models with Multiple Biomarkers**  
Eunhee Kim\* and Donglin Zeng, University of North Carolina at Chapel Hill
- 9:30 Comparison of two Binary Diagnostic Tests: Circumventing an ROC Study**  
Andriy I. Bandos\*, Howard E. Rockette and David Gur, University of Pittsburgh
- 9:45 Time-dependent Predictive Accuracy in the Presence of Competing Risks**  
Paramita Saha\* and Patrick J. Heagerty, University of Washington

## 69. CONTRIBUTED PAPERS: MODEL SELECTION/ASSESSMENT

TRAVIS C, 3RD FLOOR

Sponsor: ENAR

Chair: Fengrong Wei, University of Iowa

- 8:30 The Gradient Function for Checking Goodness-of-fit of the Random-effects Distribution in Mixed Models**  
Geert Verbeke\* and Geert Molenberghs, Katholieke Universiteit Leuven and Universiteit Hasselt, Belgium
- 8:45 The Model Selection of Zero-inflated Mixture Poisson Regression**  
Huaiye Zhang\* and Inyoung Kim, Virginia Tech
- 9:00 CLAN: A Novel, Practical Method of Curvature Assessment in Nonlinear Regression Models**  
Jieru Xie\* and Linda Jane Goldsmith, University of Louisville
- 9:15 Bayesian Case Influence Measures and Applications**  
Hyunsoon Cho\*, Hongtu Zhu and Joseph G. Ibrahim, University of North Carolina at Chapel Hill
- 9:30 A Bayesian Chi-Squared Goodness-of-Fit Test for Censored Data Models**  
Jing Cao\*, Southern Methodist University, Ann Moosman, Patrick Air Force Base, Valen Johnson, University of Texas M.D. Anderson Cancer Center
- 9:45 Bayes Factor Consistency in Linear Models**  
Ruixin Guu\* and Paul L. Speckman, University of Missouri
- 10:00 Surrogate Decision Rule in Q-learning**  
Peng Zhang\*, Bin Nan and Susan A. Murphy, University of Michigan



# SCIENTIFIC PROGRAM

## 70. CONTRIBUTED PAPERS: JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

TRAVIS D, 3RD FLOOR

Sponsor: ASA Biometrics Section

Chair: Adin-Cristian Andrei, University of Wisconsin

- 8:30 Joint Modeling of Multivariate Longitudinal Data for Mixed Responses and Survival with Application to Multiple Sclerosis Data**  
Pulak Ghosh\*, Emory University, Anneke Neuhaus and Martin Daumer, Sylvia Lawry Centre for Multiple Sclerosis Research, Sanjib Basu, Northern Illinois University
- 8:45 Evaluating Predictions of Event Probabilities from a Joint Model for Longitudinal and Event Time Data**  
Nicholas J. Salkowski\* and Melanie M. Wall, University of Minnesota
- 9:00 Semiparametric Estimation of Treatment Effect with Time-Lagged Response in the Presence of Informative Censoring**  
Xiaomin Lu\*, University of Florida and Anastasios Tsiatis, North Carolina State University
- 9:15 Some Results on Length-Biased and Current Duration Sampling**  
Broderick O. Oluyede\* Georgia Southern University
- 9:30 Diagnostic Tools for Joint Models for Longitudinal and Time-to-Event Data**  
Dimitris Rizopoulos\*, Erasmus MC, Netherlands
- 9:45 On Estimating the Relationship Between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-stage Procedure**  
Paul S. Albert\* and Joanna H. Shih, Biometric Research Branch National Cancer Institute
- 10:00 Adjusting for Measurement Error When Subject-specific Variance Estimates Are Used as Covariates in a Primary Outcome Model**  
Laine E. Thomas\*, Marie Davidian and Stefanski A. Leonard, North Carolina State University

## TUESDAY, MARCH 17

**10:15—10:30 a.m.**

**REFRESHMENT BREAK AND VISIT THE EXHIBITORS**  
TEXAS BALLROOM PRE-FUNCTION AREA, 4TH FLOOR

**10:30 a.m.—12:15 p.m.**

## 71. PRESIDENTIAL INVITED ADDRESS

TEXAS BALLROOM DEF, 4TH FLOOR

Sponsor: ENAR

Organizer/Chair: Lance Waller, Emory University

**10:30 Introduction:**

Lance Waller, Emory University

**10:35 Distinguished Student Paper Awards**

**10:45 Statistical Modeling for Real-time Epidemiology**

Professor Peter J. Diggle, Lancaster University School of Health and Medicine and Johns Hopkins University School of Public Health

**1:45—3:30 p.m.**

## 72. PREDICTION AND CURE MODELING IN MODERN MEDICAL DATA ANALYSIS

TEXAS BALLROOM A, 4TH FLOOR

Sponsors: ASA Biopharmaceutical Section, ASA Health Policy Statistics Section

Organizers: Yi Li, Dana Farber Cancer Institute and Harvard University and Megan Othus, Harvard University  
Chair: Megan Othus, Harvard University

- 1:45 Transformation Models with Gamma-Frailty for Multivariate Failure Times**  
Joseph G. Ibrahim\* and Donglin Zeng, University of North Carolina, Qingxia Chen, Vanderbilt University
- 2:10 Prediction of U.S. Mortality Counts Using Semiparametric Bayesian Techniques**  
Ram Tiwari\*, U.S. Food and Drug Administration
- 2:35 A Family of Cure Models**  
Jeremy MG Taylor\* and Ning Smith, University of Michigan
- 3:00 Analysis of Cure Rate Survival Data Under Proportional Odds Model**  
Debjyoti Sinha\*, Florida State University, Sudipto Banerjee, University of Minnesota, Yu Gu, Florida State University
- 3:25 Floor Discussion**

## 73. NETWORK ANALYSIS MODELS, METHODS AND APPLICATIONS

TEXAS BALLROOM B, 4TH FLOOR

Sponsor: ASA Section on Statistics in Defense and National Security

Organizer: Hernando Ombao, Brown University

Chair: Hernando Ombao, Brown University

- 1:45 Neural Functional Connectivity Networks**  
Crystal Linkletter\*, Hernando Ombao and Mark Fiecas, Brown University
- 2:15 Network Filtering with Application to Detection of Gene Drug Targets**  
Eric D. Kolaczyk\*, Boston University
- 2:45 Exponential-Family Random Graph Models for Biological Networks**  
David Hunter\*, Penn State University
- 3:15 Floor Discussion**

## 74. STATISTICAL METHODS IN GENOME-WIDE GENE REGULATION STUDIES

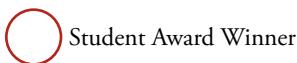
TEXAS BALLROOM C, 4TH FLOOR

Sponsor: ASA Biometrics Section

Organizer: Hongkai Ji, Johns Hopkins University

Chair: Hongkai Ji, Johns Hopkins University

- 1:45 Integrative Modeling of Transcription and Epigenetic Regulation**  
Xiaole Shirley Liu\*, Harvard University
- 2:10 Learning Gene Regulatory Network Profile across Multiple Experimental Conditions**  
Qing Zhou\* and Michael J. Mason, UCLA
- 2:35 A Hierarchical Semi-Markov Model for Detecting Enrichment with Application to ChIP-Seq Experiments**  
Sunduz Keles\* and Pei Fen Kuan, University of Wisconsin, Madison



Student Award Winner

\* Presenter

# SCIENTIFIC PROGRAM

3:00	<b>A Correlation Motif Based Hidden Markov Model for Pooling Information from Multiple ChIP-chip Experiments</b> Hongkai Ji* and Hao Wu, Johns Hopkins Bloomberg School of Public Health	2:15	<b>Bayesian Association Testing of SNP Markers and Wood Chemistry Traits in Clonal Trials of Loblolly Pine</b> Xiaobo Li*, Dudley A. Huber and George Casella, University of Florida, David B. Neale, University of California-Davis, Gary F. Peter, University of Florida
3:25	<b>Floor Discussion</b>	2:30	<b>Within-Cluster Resampling (Multiple Outputation) for Analysis of Family Data: Ready for Prime-Time?</b> Hemant K. Tiwari*, Amit Patki and David B. Allison, University of Alabama at Birmingham
<b>75. EXPERIMENTAL DESIGNS IN DRUG DISCOVERY &amp; CLINICAL TRIALS</b> PRESIDIO B, 3RD FLOOR Sponsor: ASA Biopharmaceutical Section Organizer: Sourish Saha, GlaxoSmithKline Chair: Sourish Saha, GlaxoSmithKline		2:45	<b>Estimation of the Contribution of Rare Disease-predisposing Variants to Complex Diseases</b> Weihua Guan*, Laura J. Scott and Michael Boehnke, University of Michigan
1:45	<b>Penalized Designs of Multi-response Experiments</b> Valerii V. Fedorov* and Rongmei Zhang, GlaxoSmithKline	3:00	<b>Genomics of Complex Diseases</b> Li Luo*, The University of Texas School of Public Health, Gang Peng, School of Life Science, Fudan University, Eric Boerwinkle and Momiao Xiong, The University of Texas School of Public Health
2:15	<b>Aspects of Optimal Dose Response Design</b> Randall D. Tobias*, SAS Institute Inc. and Alexander N. Donev, University of Manchester	3:15	<b>Reducing Costs of Two Stage Genome Wide Association Studies</b> Michael D. Swartz* and Sanjay Shete, University of Texas M. D. Anderson Cancer Center
2:45	<b>An Adaptive Optimal Design for the Emax Model and Its Application in Clinical Trials</b> Sergei Leonov* and Sam Miller, GlaxoSmithKline	<b>78. CONTRIBUTED PAPERS: ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS</b> INDEPENDENCE, 3RD FLOOR Sponsor: ASA Section on Statistics and the Environment Chair: Andrea J Cook, Group Health Center for Health Studies	
3:15	<b>Floor Discussion</b>		
<b>76. STATISTICAL METHODS FOR FLOW CYTOMETRY DATA</b> CROCKETT B, 4TH FLOOR Sponsor: ENAR Organizer: George Luta, Georgetown University Chair: George Luta, Georgetown University		1:45	<b>Median Polish Algorithms for Automated Anomaly Detection in Environmental Sensor Networks</b> Ernst Linder*, University of New Hampshire, Zoe Cardon, Woods Hole, Marine Biology Laboratory, Jared Murray, University of New Hampshire
1:45	<b>Automated Feature Extraction for Flow Cytometry</b> Errol Strain and Perry Haaland, BD Technologies	2:00	<b>Data Assimilation for Prediction of Fine Particulate Matter</b> Ana G. Rappold*, U.S. Environmental Protection Agency and Marco A. Ferreira, University of Missouri - Columbia
2:10	<b>Bioconductor Tools for High-throughput Flow-cytometry Data Analysis</b> Florian M. Hahne*, Fred Hutchinson Cancer Research Center	2:15	<b>Bayesian Model Averaging Approach in Health Effects Studies</b> Ya-Hsiu Chuang* and Sati Mazumdar, University of Pittsburgh
2:35	<b>Characterizing Immune Responses via Flow Cytometry</b> Martha Nason*, National Institute of Allergy and Infectious Diseases, National Institutes of Health	2:30	<b>Effect Modification of Prenatal Mercury Exposure</b> Association with Developmental Outcomes by Social and Environmental Factors Tanzu M. Love* and Sally Thurston, University of Rochester Medical Center
3:00	<b>Automated Gating of Flow Cytometry Data via Robust Model-based Clustering</b> Kenneth Lo, University of British Columbia, Ryan Brinkman, BC Cancer Agency and Raphael Gottardo*, Clinical Research Institute of Montreal	2:45	<b>Latent Spatial Modelling for Species Abundance</b> Avishek Chakraborty* and Alan E. Gelfand, Duke University
3:25	<b>Floor Discussion</b>	3:00	<b>A Flexible Regression Modeling Framework for Analyzing Seagrass Areal Coverage as Characterized by Braun-Blanquet (BB) Vegetation Cover Scores</b> Paul S. Kubilis* and Mary C. Christman, University of Florida, Penny Hall, Florida Fish and Wildlife Research Institute
<b>77. CONTRIBUTED PAPERS: MULTIPLE TESTING IN GENOME-WIDE ASSOCIATION STUDIES</b> CROCKETT C, 4TH FLOOR Sponsor: ENAR Chair: Nicholas M. Pajewski, University of Alabama at Birmingham		3:15	<b>On Shortest Prediction Intervals in Log-Gaussian Random Fields</b> Victor De Oliveira*, University of Texas at San Antonio and Changxiang Rui, University of Arkansas
1:45	<b>Winner's Curse and Bias Correction in Genome-wide Association and Candidate Gene Studies</b> Lei Shen*, Eli Lilly and Company		
2:00	<b>Extended Homozygosity Score Tests to Detect Positive Selection in Genome-wide Scans</b> Ming Zhong*, Texas A&M University, Kenneth Lange and Jeanette C. Papp, UCLA, Ruzong Fan, Texas A&M University		



# SCIENTIFIC PROGRAM

## 79. CONTRIBUTED PAPERS: CATEGORICAL DATA ANALYSIS

TRAVIS A, 3RD FLOOR

Sponsor: ASA Health Policy Statistics Section  
Chair: Veronica J Berrocal, Duke University

- 1:45 **On Tests of Homogeneity for Partially Matched-pair Data**  
Hani Samawi, Robert L. Vogel\* and Wilson A. Koech, Jiann-Ping Hsu College of Public Health, Georgia Southern University
- 2:00 **A New Method for Estimating the Odds Ratio from Incomplete Matched Data**  
Kelly Miller and Stephen Looney\*, Medical College of Georgia
- 2:15 **Latent Variable Model for the Analysis of Binary Data Collected on Nuclear Families**  
Yihao Deng\*, Indiana University Purdue University - Fort Wayne, Roy Sabo, Virginia Commonwealth University, N. Rao Chaganty, Old Dominion University
- 2:30 **A Self-consistent Approach to Multinomial Logit Model with Repeated Measures**  
Shufang Wang\* and Alex Tsodikov, University of Michigan
- 2:45 **A Full Exchangeable Negative Binomial Likelihood Procedure For Modeling Correlated Overdispersed Count Data**  
Xiaodong Wang\*, Sanofi-aventis and Hanxiang Peng, Indiana University Purdue University - Indianapolis
- 3:00 **Analysis of Multivariate Longitudinal Binary Data using Marginalized Random Effects Models**  
Keunbaik Lee\*, Louisiana State University Health Science Center, Yongsung Joo, Dongguk University, South Korea, Jae Keun Yoo, University of Louisville, JungBok Lee, Korea University
- 3:15 **Robust Inference for Sparse Clustered Count Data**  
John J. Hanfelt, Ruosha Li\* and Yi Pan, Rollins School of Public Health, Emory University, Pierre Payment, Institute Armand-Frappier, Canada

## 80. CONTRIBUTED PAPERS: INFECTIOUS DISEASES

BONHAM B, 3RD FLOOR

Sponsor: ASA Section on Statistics in Defense and National Security  
Chair: Qi Zheng, Texas A&M Health Science Center

- 1:45 **Statistical Analysis of HIV-1 env Sequences and their Role in Selection Process of Viral Variants in MTCT**  
Rongheng Lin\*, University of Massachusetts Amherst, Mohan Somasundaran and Michael Kishko, University of Massachusetts Medical School
- 2:00 **An Epidemiological Model for Genetic Mapping of Viral Pathogenesis**  
Yao Li\*, Arthur Berg and Maryon M. Chang, University of Florida, Rongling Wu, Penn State University
- 2:15 **Statistical Model for a Dual-color Tag System for Investigating Virus-virus Interactions**  
Jing Zhang, Douglas A. Noe\*, Stephen E. Wright and John Bailer, Miami University

- 2:30 **Pooled Nucleic Acid Testing to Identify Antiretroviral Treatment Failure during HIV Infection**  
Susanne May\*, University of Washington, Anthony Gamst, University of California-San Diego, Richard Haubrich and Constance Benson, University of California-San Diego Medical Center, Davey M. Smith, University of California-San Diego, School of Medicine
- 2:45 **Regression Analysis of Clustered Interval Censored Data with Informative Cluster Size**  
Yang-Jin Kim\*, Ewha Womans University, Korea
- 3:00 **Estimating Incubation Period Distributions with Coarse Data**  
Nicholas Reich\*, Johns Hopkins School of Public Health
- 3:15 **Recursive Partitioning for Longitudinal Markers Based on a U-Statistic**  
Shannon Stock\* and Victor DeGruttola, Harvard University, Chengcheng Hu, University of Arizona

## 81. CONTRIBUTED PAPERS: RATER AGREEMENT AND SCREENING TESTS

TRAVIS B, 3RD FLOOR

Sponsor: ENAR  
Chair: Bo Lu, The Ohio State University College of Public Health

- 1:45 **Evaluation of Individual Observer Agreement from Data with Repeated Measurements**  
Jingjing Gao\* and Michael Haber, Emory University, Huiman Barnhart, Duke University
- 2:00 **A Missing Data Approach for Adjusting Diagnoses of Post-Traumatic Stress Disorder that are Subject to Rater Bias**  
Juned Siddique\*, Northwestern University, Bonnie L. Green, Georgetown University, Robert D. Gibbons, University of Illinois at Chicago
- 2:15 **Multivariate Concordance Correlation Coefficient**  
Sasiprappa Hiriote\*, Eberly College of Science, Penn State University, Vernon M. Chinchilli, Penn State College of Medicine
- 2:30 **Bayesian Performance Assessment for Radiologists Interpreting Mammography**  
Dawn Woodard\*, Cornell University
- 2:45 **Modeling the Cumulative Risk of a False Positive Screening Mammogram**  
Rebecca A. Hubbard\* and Diana L. Miglioretti, Group Health Center for Health Studies
- 3:00 **Reexamination and Further Development of the Roe and Metz Simulation Model for Multiple Reader ROC Decision Data**  
Stephen L. Hillis\*, VA Iowa City Medical Center
- 3:15 **Bayesian Inference for True-Benefit and Over-diagnosis in Periodic Cancer Screening**  
Dongfeng Wu\*, School of Public Health, University of Louisville and Gary L. Rosner, University of Texas, MD Anderson Cancer Center

# SCIENTIFIC PROGRAM

## 82. CONTRIBUTED PAPERS: APPLIED DATA ANALYSIS, GRAPHICAL DISPLAYS, AND BIOSTATISTICAL LITERACY

TRAVIS C, 3RD FLOOR

Sponsors: ASA Section on Teaching Statistics in the Health Sciences, ASA Section on Statistical Education  
Chair: Xingye Qiao, University of North Carolina, Chapel Hill

**1:45 Analysis of Variance on a Categorized Continuous Variable**

Wenyaw Chan\*, School of Public Health, University of Texas-Health Science Center at Houston, Lin-An Chen, Institute of Statistics, National Chiao Tung University, Hsin Chu, Taiwan, Younghun Han, Division of Epidemiology, University of Texas- M. D. Anderson Cancer Center

**2:00 Bayesian Cancer Trend Analysis**

Pulak Ghosh, Emory University, Kaushik Ghosh\*, University of Nevada Las Vegas, Ram C. Tiwari, U.S. Food and Drug Administration

**2:15 Mortality Model for Prostate Cancer**

Shih-Yuan Lee\* and Alex Tsodikov, University of Michigan

**2:30 A Class of Distributions with Normal Shape Densities on Finite Intervals**

Ahmad Reza Soltani\*, Kuwait University

**2:45 Simulation based Visualization of Inference Functions**

Daeyoung Kim\*, University of Massachusetts Amherst and Bruce G. Lindsay Department of Statistics, Penn State University

**3:00 Animated Graphics and Visual Metaphors Help Explain Complex Mathematical Relationships.**

John T. Brinton\*, University of Colorado Health Science Center and Deborah H. Glueck, University of Colorado at Denver and Health Sciences Center

**3:15 Assessing Biostatistical Literacy and Statistical Thinking**

Felicity B. Enders, Mayo Clinic

## TUESDAY, MARCH 17

**3:30—3:45 p.m.**

**REFRESHMENT BREAK AND VISIT THE EXHIBITORS**  
TEXAS BALLROOM PRE-FUNCTION AREA, 4TH FLOOR

**3:45—5:30 p.m.**

## 83. IMS MEDALLION LECTURE

TEXAS BALLROOMC, 4TH FLOOR

Sponsor: IMS

Organizer: Tianxi Cai, Harvard University  
Chair: Tianxi Cai, Harvard University

**3:45 Statistical Challenges in Nanoscale Biophysics**

Professor Samuel Kou, Harvard University, Department of Statistics

## 84. MEDIATION AND CAUSAL INFERENCE

PRESIDIO B, 3RD FLOOR

Sponsors: ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section

Organizer: James M. Robins, Harvard University  
Chair: James M. Robins, Harvard University

**3:45 Bayesian Inference for Mediation Effects Using Principal Stratification**

Michael R. Elliott\* and Trivellore E. Raghunathan, University of Michigan

**4:15 Controlled Direct and Mediated Effects: Definition, Identification and Bounds**

Tyler J. VanderWeele\*, University of Chicago

**4:45 Estimating Controlled Direct Effects in Random and Outcome-dependent Sampling Designs**

Stijn Vansteelandt\*, Ghent University, Belgium  
**5:15 Floor Discussion**

## 85. CHALLENGES IN THE BAYESIAN SPATIO-TEMPORAL ANALYSIS OF LARGE AND HETEROGENEOUS DATASETS

TEXAS BALLROOM B, 4TH FLOOR

Sponsors: ASA Section on Statistics and the Environment, ASA

Section on Statistics in Defense and National Security

Organizer: Michele Guindani, University of New Mexico

Chair: Gabriel Huerta, University of New Mexico

**3:45 Spatial Misalignment in Time Series Studies of Air Pollution and Health Data**

Roger D. Peng\*, Johns Hopkins Bloomberg School of Public Health and Michelle L. Bell, Yale University

**4:15 Bayesian Variable Selection for Multivariate Spatially-Varying Coefficient Regression: Application to Physical Activity During Pregnancy**

Montserrat Fuentes\* and Brian Reich, North Carolina State University, Amy Herring, University of North Carolina-Chapel Hill

**4:45 Hierarchical Spatial Modeling of Genetic Variance for Large Spatial Trial Datasets**

Sudipto Banerjee\*, University of Minnesota, Andrew O. Finley, Michigan State University, Patrik Waldmann and Tore Ericsson, Swedish University of Agricultural Sciences, Sweden

**5:15 Floor Discussion**

## 86. ADVANCES IN META-ANALYSIS

CROCKETT B, 4TH FLOOR

Sponsors: ASA Biometrics Section, ASA Biopharmaceutical Section

Organizer: Kenneth Rice, University of Washington

Chair: David Dunson, Duke University

**3:45 Generalizing Data from Randomized Trials**

Eloise Kaizar\*, The Ohio State University

**4:10 Multivariate Meta-analysis: Modelling Correlation Structures**

Robert W. Platt\*, McGill University and Khajak Ishak, United BioSource Corporation

**4:35 Non-parametric ROC Curve Meta-analysis with Varying Number of Thresholds**

Vanja M. Dukic\*, University of Chicago

**5:00 Discussant:**

Dalene Stangl, Duke University



# SCIENTIFIC PROGRAM

## 87. STATISTICAL METHODS FOR GENOME-WIDE ASSOCIATION STUDIES

TEXAS BALLROOM A, 4TH FLOOR

Sponsor IMS

Organizer: Jinbo Chen, University of Pennsylvania

Chair: Jinbo Chen, University of Pennsylvania

- 3:45 **Estimating Genetic Effects and Gene-Environment Interactions with Missing Data**  
Danyu Lin\*, University of North Carolina
- 4:10 **Detecting Gene-Gene Interactions Using Genome-Wide Association Studies (GWAS) in the Presence of Population Stratification**  
Nilanjan Chatterjee and Samsiddhi Bhattacharjee, National Cancer Institute
- 4:35 **Predictive Models for Genome-wide Association Studies**  
Charles Kooperberg\*, Michael LeBlanc and Valerie Obenchain, Fred Hutchinson Cancer Research Center
- 5:00 **Population Stratification Evaluation and Adjustment in Genome Wide Association Studies**  
Kai Yu\*, Qizhai Li, National Cancer Institute
- 5:25 **Floor Discussion**

## 88. CONTRIBUTED PAPERS: INFERENCE FOR CLINICAL TRIALS

TRAVIS A, 3RD FLOOR

Sponsor: ASA Biopharmaceutical Section

Chair: Karen Messer, University of California San Diego Medical Center

- 3:45 **A Comparative Simulation Study between ETRANK® Methodology and Constraint Longitudinal Data Analysis (cLDA)**  
Lian Liu\* and Richard Entsuah, Merck & Co., Inc.
- 4:00 **Randomization Tests of Multi-arm Randomized Clinical Trials**  
Youngsook Jeon\* and Feifang Hu, University of Virginia
- 4:15 **Comparison of Variations of the Duffy-Santner Confidence Intervals for the One-sample Proportion based on Multistage Designs**  
Haihong Li\*, Vertex Pharmaceuticals
- 4:30 **Weighted Kaplan-Meier Estimator for Two-stage Treatment Regimes**  
Sachiko Miyahara\* and Abdus S. Wahed, University of Pittsburgh
- 4:45 **Borrowing Strength with Non-exchangeable Priors over Subpopulations**  
Peter Mueller and Benjamin N. Bekeley, M.D. Anderson Cancer Center, Luis G. Leon Novelo\*, Rice University, Kyle Wathen, M.D. Anderson Cancer Center, Fernando A. Quintana, Pontificia Universidad Católica de Chile
- 5:00 **Inference for Nonregular Parameters in Optimal Dynamic Treatment Regimes**  
Bibhas Chakraborty\* and Susan Murphy, University of Michigan
- 5:15 **Reverse Regression in Randomized Clinical Trials**  
Zhiwei Zhang\*, U.S. Food and Drug Administration

## 89. CONTRIBUTED PAPERS: STATISTICAL GENETICS

CROCKETT C, 4th Floor

Sponsor: ENAR

Chair: Tracy L Bergemann, University of Minnesota

- 3:45 **Haplotype-Based Regression Analysis and Inference of Case-Control Studies with Unphased Genotypes and Measurement Errors in Environmental Exposures**  
Iryna V. Lobach\*, New York University School of Medicine, Raymond J. Carroll, Texas A&M University, Christine Spinka, University of Missouri, Mitchell Gail and Nilanjan Chatterjee, Division of Cancer Epidemiology and Genetics, National Cancer Institute
- 4:00 **Modeling Haplotype-Haplotype Interactions in Case-Control Genetic Association Studies**  
Li Zhang\*, Cleveland Clinic Foundation and Rongling Wu, Penn State University
- 4:15 **Nucleotide Mapping Complex Disease and the Limiting Distribution of the Likelihood Ratio Test**  
Yuehua Cui\*, Michigan State University and Dong-Yun Kim, Virginia Tech
- 4:30 **Modeling SNP Genotype Data with Informative Missingness in Samples of Unrelated Individuals**  
Nianjun Liu\*, University of Alabama at Birmingham
- 4:45 **Contribution of Genetic Effects to Genetic Variance Components with Epistasis and Linkage Disequilibrium**  
Tao Wang\*, Department of Population Health, Medical College of Wisconsin and Zhao-Bang Zeng, Bioinformatics Research Center, North Carolina State University
- 5:00 **Association Study of G Protein-Coupled Receptor Kinase 4 Gene Variants with Essential Hypertension in Northern Han Chinese**  
Yaping Wang\*, Emory University, Biao Li, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China, Weiyan Zhao, Division of Population Genetics and Prevention, Cardiovascular Institute and Fu Wai Hospital, Pei Liu, The School of Public Health Southeast University, Nanjing, China, Qi Zhao, Tulane University, Shufeng Chen, Hongfan Li and Dongfeng Gu, Division of Population Genetics and Prevention, Cardiovascular Institute and Fu Wai Hospital
- 5:15 **Composite Likelihood: Issues in Efficiency**  
Jianping Sun\*, Penn State University

## 90. CONTRIBUTED PAPERS: HEALTH SERVICES RESEARCH

TRAVIS B, 3RD FLOOR

Sponsors: ASA Section on Statistics in Epidemiology, ASA Health Policy Statistics Section

Chair: Lingling Li, Harvard Medical School

- 3:45 **Statistical Methods in Healthcare Quality Improvement**  
Claudia Pedroza\*, University of Texas School of Public Health
- 4:00 **Structural Equation Modeling for Quality of Life Data in Cardiovascular Disease**  
Zugui Zhang\*, Paul Kolm and William S. Weintraub, Christiana Care Health System

# SCIENTIFIC PROGRAM

4:15

## Bayesian Hierarchical Models for Extracting Useful Information from Medication Error Reports

Jessica A. Myers\*, Francesca Dominici and Laura Morlock, Bloomberg School of Public Health, Johns Hopkins University

4:30

## Analysis of Duration Times with Unobserved Heterogeneity through Finite Mixtures

Xiaoqin Tang\*, Hwan Chung and Joseph Gardiner, Michigan State University

4:45

## A Study on Confidence Intervals for Incremental Cost-effectiveness

Hongkun Wang\*, University of Virginia and Hongwei Zhao, Texas A&M Health Science Center

5:00

## Joint Modeling of Zero-inflated Data using Copulas

Joanne K. Dagg\*<sup>1</sup>, Purdue University and Bruce A. Craig, University of Wisconsin-Madison

5:15

## Estimating a Volume-outcome Association from Aggregate Longitudinal Data

Benjamin French\*, University of Pennsylvania, Farhood Farjah, David R. Flum and Patrick J. Heagerty, University of Washington

## 91. CONTRIBUTED PAPERS: EXPERIMENTAL DESIGN

TRAVIS C, 3RD FLOOR

Sponsor: ENAR

Chair: John Keighley, University of Kansas Medical Center

3:45

## Counterintuitive Results when Calculating Sample Size in ANOVA

Yolanda Munoz Maldonado\*, Michigan Technological University

4:00

## Bayesian Experimental Design for Stability Studies

Harry Yang\* and Lanju Zhang, MedImmune

4:15

## Calculating Sample Size for Studies with Expected All-or-none Nonadherence and Selection Bias

Michelle D. Shardell\* and Samer S. El-Kamary, University of Maryland School of Medicine

4:30

## Comparison of Different Sample Size Designs - Group Sequential versus Re-estimation

Xiaoru Wu\*, Columbia University and Lu Cui, Eisai Medical Research Inc.

4:45

## On the Role of Baseline Measurements for Crossover Designs under the Self and Mixed Carryover Effects Model

Yuanyuan Liang\*, University of Texas Health Science Center at San Antonio and Keumhee Chough Carriere, University of Alberta

5:00

## Efficiency of Study Designs in Diagnostic Randomized Clinical Trials

Bo Lu\*, The Ohio State University and Constantine Gatsonis, Brown University

5:15

## Statistical Validity and Power for Testing for Heterogeneous Effects with Quantitative Traits and its Application to Pharmacogenetic Studies

Todd G. Nick\*, Mi-ok Kim, Chunyan Liu and Yu Wang, Cincinnati Children's Hospital Medical Center

## 92. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS

TRAVIS D, 3RD FLOOR

Sponsor: ENAR

Chair: Ying Ding, University of Michigan

3:45

## Prediction and Misclassification in Right Censored Time-to-Event Data

Keith A. Betts\*, Harvard School of Public Health and David P. Harrington, Harvard School of Public Health and Dana-Farber Cancer Institute

4:00

## Incorporating Rate of Change into Tree Structured Models with Time Varying Covariates

Meredith J. Lotz\*, Stewart J. Anderson and Sati Mazumdar, University of Pittsburgh

4:15

## Exponential Tilt Models in the presence of Censoring

Chi Wang\*, Johns Hopkins University, Zhiqiang Tan, Rutgers University, Thomas A. Louis, Johns Hopkins University

4:30

## A Risk-Adjusted O-E CUSUM with V-mask in a Continuous Time Setting

Jie (Rena) Sun and John D. Kalbfleisch, University of Michigan

4:45

## Bias-corrected Logrank Test with Dependent Censoring

Yabing Mai\*, Merck & Co., Inc. and Eric V. Slud, University of Maryland

5:00

## The Comparison of Alternative Smoothing Methods for Fitting Non-linear Exposure-response Relationships with Cox Models in a Simulation Study

Usha S. Govindarajulu\*, Harvard Medical School, Betty J. Malloy, American University, Bhawati Ganguli, University of Calcutta, Donna Spiegelman, Harvard School of Public Health, Ellen A. Eisen, University of California, Berkeley and Harvard School of Public Health

5:15

## Utilizing Biostatistical Methods in the Analysis of Data in Discrimination Cases

Joseph L. Gastwirth\* and Qing Pan, George Washington University

## 93. CONTRIBUTED PAPERS: FUNCTIONAL DATA ANALYSIS

CROCKETT D, 4TH FLOOR

Sponsor: ENAR

Chair: Hong-Bin Fang, University of Maryland

3:45

## Generalized Multilevel Functional Regression

Ciprian M. Crainiceanu, Johns Hopkins University, Ana-Maria Staicu\*, University of Bristol, UK, ChongZhi Di, Johns Hopkins University

4:00

## Stochastic Functional Data Analysis: A Diffusion Model-based Approach

Bin Zhu\*, Peter X.-K. Song and Jeremy M.G. Taylor, University of Michigan

4:15

## Wavelet-based Functional Mixed Models via DPM

Alejandro Villagran\*, Rice University, Sang Han Lee, New York University, Marina Vannucci, Rice University

4:30

## Data Driven Adaptive Spline Smoothing with Applications to Epileptic EEG Data

Ziyue Liu\* and Wensheng Guo, University of Pennsylvania School of Medicine



# SCIENTIFIC PROGRAM

- 4:45 **Modelling Labor Curves in Women Attempting a Vaginal Birth After a Cesarean Using a B-splines Based Semiparametric Nonlinear Mixed Effects Model**  
Angelo Elmí\*, Sarah Ratcliffe, Sam Parry and Wensheng Guo, University of Pennsylvania
- 5:00 **A Bayesian Regression Model for Multivariate Functional Data**  
Ori Rosen\*, University of Texas at El Paso and Wesley K. Thompson, University of California San Diego
- 5:15 **Analysis of Long Period Variable Stars with a Nonparametric Significance Test of No Trend**  
Woncheol Jang\*, Cheolwoo Park and Jeongyoun Ahn, University of Georgia, Martin Hendry, University of Glasgow

## WEDNESDAY, MARCH 18

8:30—10:15 a.m.

### 94. EVALUATING MARKERS FOR RISK PREDICTION

PRESIDIO B, 3RD FLOOR

Sponsors: ASA Health Policy Statistics Section, IMS  
Organizer: Holly Janes, Fred Hutchinson Cancer Research Center  
Chair: Tianxi Cai, Harvard University

- 8:30 **Decision Curve Analysis: A Simple, Novel Method for the Evaluation of Prediction Models, Diagnostic Tests and Molecular Markers**  
Andrew J. Vickers\*, Memorial Sloan-Kettering Cancer Center
- 8:55 **On Incorporating Biomarkers into Models for Absolute Risk Prediction Models**  
Ruth Pfeiffer\* and Mitchell Gail, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health
- 9:20 **Estimating the Capacity for Improvement in Risk Prediction with a Marker**  
Wen Gu\* and Margaret S. Pepe, University of Washington and Fred Hutchinson Cancer Research Center
- 9:45 **Calculating Disease Risk: Evaluating Risk Prediction Models using Risk Stratification Tables**  
Holly Janes\*, Margaret S. Pepe and Jessie Gu, University of Washington and Fred Hutchinson Cancer Research Center
- 10:10 **Floor Discussion**

### 95. ADVANCES IN FUNCTIONAL DATA ANALYSIS

PRESIDIO A, 3RD FLOOR

Sponsor: IMS  
Organizer: Hans Mueller, University of California at Davis  
Chair: Hans Mueller, University of California at Davis

- 8:30 **Functional Principal Components Analysis with Survey Data**  
Hervé, Cardot\*, Institut de Mathématiques, Université de Bourgogne
- 9:00 **Concept of Density for Functional Data**  
Peter Hall\*, University of Melbourne and University of California at Davis and Aurore Delaigle, University of Bristol

- 9:30 **Deciding the Dimension of Effective Dimension Reduction Space for Functional Data**  
Yehua Li\*, University of Georgia and Tailen Hsing, University of Michigan

### 10:00 Floor Discussion

### 96. NEW STATISTICAL CHALLENGES AND ADVANCEMENTS IN GENOME-WIDE ASSOCIATION STUDIES

CROCKETT A/B, 4TH FLOOR

Sponsor: ENAR  
Organizer: Ching-Ti Liu, Yale University  
Chair: Ching-Ti Liu, Yale University

- 8:30 **On the Adjustment for Covariates in Genetic Association Analysis: A Novel, Simple Principle to Infer Causality**  
Stijn Vansteelandt, University of Ghent and Christoph Lange\*, Harvard School of Public Health
- 8:55 **Family-based Association Test for Multiple Traits Speaker**  
Heping Zhang\*, Yale University, Ching-Ti Liu, Boston University, Xueqin Wang, Sun-Yat Sen University, Wensheng Zhu, Yale University
- 9:20 **A Penalized Likelihood Approach to Haplotype Specific Analysis**  
Jung-Ying Tzeng and Howard D. Bondell\*, North Carolina State University
- 9:45 **Statistical Methods for Gene Mapping using High Density SNPs in Family Samples**  
Josée Dupuis\*, Boston University School of Public Health
- 10:10 **Floor Discussion**

### 97. RESPONSE-ADAPTIVE DESIGNS FOR CLINICAL TRIALS

TEXAS BALLROOM C, 4TH FLOOR

Sponsor: ASA Biopharmaceutical Section  
Organizer: Xuelin Huang, The University of Texas M.D. Anderson Cancer Center  
Chair: Yisheng Li, The University of Texas M.D. Anderson Cancer Center

- 8:30 **Response Adaptive Randomization in Non-Inferiority Trials**  
Lanju Zhang\* and Harry Yang, MedImmune LLC
- 8:55 **Sequential Monitoring Response-Adaptive Randomized Clinical Trials**  
Feifang Hu\* and Hongjian Zhu, University of Virginia
- 9:20 **Using Short-Term Response Information to Facilitate Adaptive Randomization for Survival Clinical Trials**  
Jing Ning\*, Yisheng Li and Donald A. Berry, The University of Texas MD Anderson Cancer Center
- 9:45 **Discussant:**  
Xuelin Huang, The University of Texas M.D. Anderson Cancer Center
- 10:10 **Floor Discussion**

# SCIENTIFIC PROGRAM

## 98. DEVELOPMENT OF BAYESIAN SURVIVAL AND RISK ANALYSIS

INDEPENDENCE, 3RD FLOOR

Sponsor: ENAR

Organizer: Sourish Das, University of Connecticut

Chair: Arpita Ghosh, University of North Carolina, Chapel Hill

**8:30 Classical and Bayes Estimation for Additive Hazards Regression Models**

Stuart R. Lipsitz\*, Brigham and Women's Hospital, Debajyoti Sinha, Florida State University, M. Brent McHenry, Bristol-Myers Squibb, Malay Ghosh, University of Florida

**9:00 Bayesian Development of a Generalized Link Function for Binary Response Data**

Xia Wang\* and Dipak K. Dey, University of Connecticut  
**Analysis of Extreme Drinking in Patients with Alcohol Dependence Using Pareto Regression**

Sourish Das\*, Duke University, Ofer Harel and Dipak K. Dey, University of Connecticut, Jonathan Covault and Henry R. Kranzler, University of Connecticut Health Center, Psychiatry

**10:00 Floor Discussion**

## 99. CONTRIBUTED PAPERS: PROTEOMICS / METABOLOMICS

BONHAM C, 3RD FLOOR

Sponsor: ASA Biometrics Section

Chair: Iryna Lobach, New York University School of Medicine

**8:30 A Bayesian Approach to the Alignment of Mass Spectra**

Xiaoxiao Kong\* and Cavan Reilly, School of Public Health, University of Minnesota

**8:45 Two-dimensional Correlation Optimized Warping Algorithm for Aligning GCXGC-MS Data**

Dabao Zhang\*, Xiaodong Huang, Fred E. Regnier and Min Zhang, Purdue University

**9:00 Monoisotopic Peak Detection for Mass Spectrometry Data**

Mourad Atlas\* and Susmita Datta, University of Louisville

**9:15 A Two-Stage Approach for Detecting Clusters of Peaks with Periodicity in NMR Spectra**

Anna K. Jolly\*, Amita Manatunga and Tianwei Yu, Emory University

**9:30 Sparsity Priors for Protein-protein Interaction Predictions**

Inyoung Kim\*, Virginia Tech, Yin Liu, University of Texas Medical School at Houston, Hongyu Zhao, Yale University

**9:45 Statistically Appraising Affinity-Isolation Experiment Process Quality**

Julia Sharp\*, Clemson University, John Borkowski, Montana State University, Denise Schmoyer and Greg Hurst, Oak Ridge National Laboratory

**10:00 Set Enrichment Analysis Methods for Integromic Studies**

Laila M. Poisson\*, School of Public Health, University of Michigan and Debasish Ghosh, Penn State University

## 100. CONTRIBUTED PAPERS: DETECTING GENE DEPENDENCIES AND CO-EXPRESSION

BONHAM D, 3RD FLOOR

Sponsor: ASA Biometrics Section

Chair: Donatello Telesca, The University of Texas, M.D. Anderson Cancer Center

**8:30 Detecting Non-linear Dependencies in Gene Co-expression Networks**

Alina Andrei\*, University Of Wisconsin-Madison

**8:45 A Nonparametric Approach to Detect Nonlinear Correlation in Gene Expression**

Yian Ann Chen\*, Moffitt Cancer Center, University of South Florida, Jonas S. Almeida, The University of Texas, M.D. Anderson Cancer Center, Adam J. Richards, Medical University of South Carolina, Peter Müller, The University of Texas, M.D. Anderson Cancer Center, Raymond J. Carroll, Texas A&M University, Baerbel Rohrer, Medical University of South Carolina

**9:00 Analysis for Temporal Gene Expressions under Multiple Biological Conditions**

Hong-Bin Fang\*, University of Maryland Greenebaum Cancer Center, Dianliang Deng, University of Regina, Canada, Jiuzhou Song, University of Maryland, Ming Tan, University of Maryland Greenebaum Cancer Center

**9:15 Query Large Scale Microarray Compendium Datasets using a Model-based Bayesian Approach with Variable Selection**

Ming Hu\* and Zhaojun Qin, School of Public Health, University of Michigan

**9:30 Modelling Three dimensional Chromosome Structures Using Gene Expression Data**

Guanghua Xiao, University of Texas Southwestern Medical Center, Xinlei Wang\*, Southern Methodist University, Arkady Khodursky, University of Minnesota

**9:45 Order Reversal Detection (ORD) for Analysis of Splice-Junction Microarrays**

Jonathan A. Gelfond\* and Luiz Penalva, University of Texas Health Science Center, San Antonio

**10:00 Analysis of Cancer-Related Epigenetic Changes in DNA Tandem Repeats**

Michelle R. Lacey\*, Tulane University, Koji Tsumagari and Melanie Ehrlich, Hayward Genetics Center, Tulane University School of Medicine

## 101. CONTRIBUTED PAPERS: NONPARAMETRIC METHODS

TRAVIS A, 3RD FLOOR

Sponsor: ENAR

Chair: Kai Fun Yu, National Institutes of Health

**8:30 Rank Inference for Varying Coefficient Models**

Lan Wang, University of Minnesota, Bo Kai\* and Runze Li, Penn State University

**8:45 Comparison of Treatment Effects-An Empirical Likelihood Based Method**

Haiyan Su\* and Hua Liang, University of Rochester

**9:00 A Multivariate Likelihood-tuned Density Estimator**

Yejin Chung\* and Bruce G. Lindsay, Penn State University



# SCIENTIFIC PROGRAM

9:15	<b>Spline-Based Sieve Semiparametric Generalized Estimating Equation Method</b> Lei Hua* and Ying Zhang, University of Iowa	8:45	<b>Multiple Imputation Based on Restricted Mean Models for Censored Survival Data</b> Lyrica Xiaohong Liu*, Susan Murray and Alex Tsodikov, University of Michigan
9:30	<b>Dimension Reduction for Non-elliptically Distributed Predictors: Second-order Methods</b> Yuxiao Dong* and Bing Li, Penn State University	9:00	<b>Non-parametric Estimation of a Lifetime Distribution with Incomplete Censored Data</b> Chung Chang*, New Jersey Institute of Technology and Wei-Yann Tsai, Mailman School of Public Health, Columbia University
9:45	<b>Rank-based Similarity Metric with Tolerance</b> Aixiang Jiang* and Yu Shyr, Vanderbilt University	9:15	<b>Bayesian Model Averaging for Clustered Data: Imputing Missing Daily Air Pollution Concentrations</b> Howard H. Chang*, Roger D. Peng and Francesca Dominici, Johns Hopkins University
10:00	<b>Clustering Techniques for Histogram-valued Data</b> Jaejik Kim* and Lynne Billard, University of Georgia	9:30	<b>Non-Ignorable Models for Intermittently Missing Categorical Longitudinal Responses</b> Roula Tsonaka*, Katholieke Universiteit Leuven, Belgium, Dimitris Rizopoulos, Erasmus Medical Center, The Netherlands, Geert Verbeke, Katholieke Universiteit Leuven, Belgium
<b>102. CONTRIBUTED PAPERS: BAYESIAN SPATIAL/ TEMPORAL MODELING</b> TRAVIS B, 3RD FLOOR Sponsor: ASA Section on Statistics in Epidemiology Chair: Ronald Gangnon, University of Wisconsin, Madison		9:45	<b>Identification Strategies for Pattern Mixture Models with Covariates</b> Chenguang Wang* and Michael J. Daniels, University of Florida
8:30	<b>Bayesian Modelling of Wind Fields using Surface Data Collected Over Land</b> Margaret Short* and Javier Fochesatto, University of Alaska Fairbanks	10:00	<b>A Comparison of Imputation Methods in a Longitudinal Clinical Trial Count Data</b> Mohamed Alesh*, U.S. Food and Drug Administration
8:45	<b>Bayesian Non-Parametric Approaches for Detecting Abrupt Changes in Disease Maps.</b> Pei Li*, Sudipto Banerjee, Timothy E. Hanson and Alexander M. McBean, University of Minnesota	<b>104. CONTRIBUTED PAPERS: META-ANALYSIS</b> TRAVIS D, 3RD FLOOR Sponsor: ASA Biopharmaceutical Section Chair: Hani Samawi, Georgia Southern University	
9:00	<b>A Marginalized Zero Altered Model for Spatially Correlated Counts with Excessive Zeros</b> Loni P. Philip* and Brent Coull, Harvard School of Public Health	8:30	<b>Meta Analysis of Soil Ingestion Intake for Childhood Risk Assessment</b> Edward J. Stanek III* and Edward J. Calabrese, University of Massachusetts, Amherst
9:15	<b>Space-time Dirichlet Process Mixture Models for Small Area Disease Risk Estimation</b> M.D. M. Hossain*, Biostatistics/Epidemiology/Research Design (BERD) Core Center for Clinical and Translational Sciences, The University of Texas Health Science Center at Houston and Andrew B. Lawson, Medical University of South Carolina	8:45	<b>Methyl Bromide Alternatives in Huelva (Spain): A Case of Meta-analysis Application</b> Dihua Xu* and Bimal K. Sinha, University of Maryland-Baltimore County, Guido Knapp, TU Dortmund University
9:30	<b>Asymptotic Comparison of Predictive Densities for Dependent Observations</b> Xuanyao He*, Richard Smith and Zhengyuan Zhu, University of North Carolina at Chapel Hill	9:00	<b>An Alternative Approach to Meta-analysis with Confidence</b> G.Y. Zou*, University of Western Ontario
9:45	<b>Zero-inflated Bayesian Spatial Models with Repeated Measurements</b> Jing Zhang*, Miami University and Chong Z. He, University of Missouri-Columbia	9:15	<b>A Comparison of Meta-analysis Methods for Rare Events in Clinical Trials</b> Zhiying Xu*, Mark Donovan and David H. Henry, Bristol-Myers Squibb Company
10:00	<b>Bayesian Modeling for Nonstationary Multivariate Spatial Processes</b> Anandamayee Majumdar*, Arizona State University, Debashis Paul, University of California at Davis, Dianne Bautista, The Ohio State University	9:30	<b>Random Effects Meta-analysis with Two-component Normal Mixtures</b> Michael P. LaValley*, Boston University
<b>103. CONTRIBUTED PAPERS: MISSING VALUES IN SURVIVAL AND/OR LONGITUDINAL DATA</b> TRAVIS C, 3RD FLOOR Sponsor: ENAR Chair: Chiu-Hsieh Hsu, Arizona Cancer Center		9:45	<b>Meta-analyses on the Rate of Change Over Time When Individual Patient Data are Partly Available</b> Chengjie Xiong*, Washington University
8:30	<b>Empirical Likelihood Confidence Intervals for the Ratio and Difference of Two Hazard Functions</b> Yichuan Zhao* and Meng Zhao, Georgia State University	10:00	<b>A Meta-analytic Framework for Combining Incomparable Cox Proportional Hazard Models Caused by Omitting Important Covariates</b> Xing Yuan* and Stewart Anderson, University of Pittsburgh

# SCIENTIFIC PROGRAM

## WEDNESDAY, MARCH 18

10:15—10:30 a.m.

### REFRESHMENT BREAK AND VISIT THE EXHIBITORS TEXAS BALLROOM PRE-FUNCTION AREA, 4TH FLOOR

10:30 a.m.—12:15 p.m.

### 105. INTEGRATING GENOMIC AND/OR GENETICS DATA

TEXAS BALLROOM C, 4TH FLOOR

Sponsor: ENAR

Organizer: Naisyin Wang, Texas A&M University

Chair: Naisyin Wang, Texas A&M University

**10:30 Information-integration Approaches to Biological Discovery in High-dimensional Data**

John Quackenbush\*, Dana-Farber Cancer Institute and Harvard University

**10:55 Methods for Identifying the Genetic Variants Associated with High-order Expression Modules**

Hongzhe Li\*, University of Pennsylvania

**11:20 A Graphical Solution for the Multiple Top-k List Problem with Applications in Molecular Medicine**

Michael G. Schimek\*, Danube University, Krems, Austria and Eva Budinska, Masaryk University, Brno, Czech Republic

**11:45 Finite Mixture of Sparse Normal Linear Models in High Dimensional Feature Space with Applications to Genomics Data**

Shili Lin\*, The Ohio State University

**12:10 Floor Discussion**

### 106. APPLICATION OF DYNAMIC TREATMENT REGIMES

PRESIDIO B, 3RD FLOOR

Sponsors: ASA Biopharmaceutical Section, ASA Health Policy Statistics Section

Organizer: Abdus Wahed, University of Pittsburgh

Chair: Abdus Wahed, University of Pittsburgh

**10:30 Estimating the Causal Effect of Lower Tidal Volume Ventilation on Survival in Patients with Acute Lung Injury**

Daniel O. Scharfstein\*, Johns Hopkins Bloomberg School of Public Health

**11:00 STAR\*D, Dynamic Treatment Regimes and Missing Data**

Dan Lizotte, Lacey Gunter, Eric Laber and Susan Murphy\*, University of Michigan

**11:30 A Prostate Cancer Trial with Re-Randomization: How We Spent a Decade Studying Twelve Dynamic Treatment Regimes**

Peter F. Thall\*, University of Texas, M.D. Anderson Cancer Center

**12:00 Floor Discussion**

### 107. DATA SHARING: AN EXAMPLE OF CONFLICTING INCENTIVES

CROCKETT A/B, 4TH FLOOR

Sponsor: ASA Health Policy Statistics Section

Organizer: Sharon-Lise Normand, Harvard University

Chair: Sharon-Lise Normand, Harvard University

**10:30 Data Sharing: An Example of Conflicting Incentives**

Thomas A. Louis\*, Johns Hopkins Bloomberg School of Public Health

**11:00 Access to Data from Publicly Funded Studies: Opportunities and Unintended Consequences**

Constantine Gatsonis\*, Brown University

**11:30 NIH Mandate on Sharing Research Data: The Regulatory Landscape**

Jane Pendergast\*, The University of Iowa

**12:00 Discussant:**

Mary K. Pendergast, Pendergast Consulting

### 108. MAPPING SPATIAL DATA INTO THE FUTURE

PRESIDIO A, 3RD FLOOR

Sponsors: ASA Section on Statistics and the Environment, ASA

Section on Statistics in Defense and National Security

Organizer: Bo Li, National Center for Atmospheric Research

Chair: Bo Li, National Center for Atmospheric Research

**10:30 Testing and Modeling the Cross-Covariance Functions of Multivariate Random Fields**

Marc G. Genton\*, Texas A&M University

**10:55 On Consistent Nonparametric Intensity Estimation for Inhomogeneous Spatial Point Processes**

Yongtao Guan\*, Yale University

**11:20 Probabilistic Quantitative Precipitation Forecasting Using a Two-Stage Spatial Model**

Tilmann Gneiting\*, University of Washington, Veronica J. Berrocal, Duke University, Adrian E. Raftery, University of Washington

**11:45 The Other World of Large Spatial data Sets**

Douglas Nychka\*, National Center for Atmospheric Research

**12:10 Floor Discussion**

### 109. STRATEGIES FOR SUCCESSFUL STATISTICAL CONSULTING/COLLABORATION IN DRUG AND VACCINE DISCOVERY AND DEVELOPMENT

BONHAM C, 3RD FLOOR

Sponsor: ENAR

Organizer: Karen Chiswell, GlaxoSmithKline

Chair: Karen Chiswell, GlaxoSmithKline

**10:30 Statistical Consulting: Earning Your Place on the Team**

Mandy L. Bergquist, GlaxoSmithKline

**10:55 From Population to Cell to Animal to Human**

Borko D. Jovanovic\* and Raymond C. Bergan, Northwestern University Medical School

**11:20 The Role of the Statistical Consultant in Strategic Planning and Development in the Pharmaceutical and Biotechnology Industries**

Bruce E. Rodda\*, Strategic Statistical Consulting LLC and The University of Texas School of Public Health

**11:45 Prioritizing Areas of Therapeutic Interest to Target Product Profiles: Lessons and Examples in Infectious Diseases and Vaccine Development**

Nicole C. Close\*, EmpiriStat, Inc.

**12:10 Floor Discussion**



# SCIENTIFIC PROGRAM

## 110. CONTRIBUTED PAPERS: CLINICAL TRIAL DESIGN

TRAVIS A, 3RD FLOOR

Sponsor: ASA Biopharmaceutical Section

Chair: S. Krishna Padmanabhan, Wyeth Research

### 10:30 Safety and Accuracy of Phase 1 Oncology Trials: Traditional Method vs. Rolling 6 vs. CRM

Arzu Onar-Thomas\* and Zang Xiong, St. Jude Children's Research Hospital

### 10:45 Dose-establishment Designs for Phase I/II Cancer Immunotherapy Trials

Karen Messer\*, Loki Natarajan, Edward Ball and Thomas Lane, Department of Medicine and Moores University of California at San Diego Cancer Center

### 11:00 An Iterative Phase I/II Clinical Trial Design Incorporating Genomic Biomarkers

Ashok Krishnamurthy\*, School of Public Health and Information Sciences, University of Louisville

### 11:15 Reinforcement Learning Design for Cancer Clinical Trials

Yufan Zhao\*, Michael R. Kosorok and Donglin Zeng, University of North Carolina at Chapel Hill

### 11:30 Propensity Score Matching in Randomized Clinical Trial

Zhenzhen Xu\* and John D. Kalbfleisch, University of Michigan

### 11:45 Evaluating Probability of Success for Clinical Trials with Time-to-Event Endpoints

Di Li\*, Todd A. Busman and Martin S. King, Abbott Laboratories

### 12:00 Design of Predictive Power Method with Two Endpoints

Kiya R. Hamilton\* and Leslie A. McClure, University of Alabama at Birmingham

## 111. CONTRIBUTED PAPERS: GENETIC STUDIES WITH RELATED INDIVIDUALS

BONHAM D, 3RD FLOOR

Sponsor: ENAR

Chair: Li Zhang, Cleveland Clinic

### 10:30 Analysis of Twin Data Using SAS

Rui Feng\*, University of Alabama at Birmingham, Gongfu Zhou, Meizhuo Zhang and Heping Zhang, Yale University

### 10:45 Haplotyping Inherited Human Diseases with a Family-Based Design

Qin Li\* and Arthur Berg, University of Florida, Rongling Wu, University of Florida; Department of Public Health Sciences, Penn State University

### 11:00 Imputing Missing Data in Case-Parent Triad Studies

Tracy L. Bergemann\*, University of Minnesota

### 11:15 Evaluating the Heterogeneity of Polygenic Variance Component by Sex and Age on Cardiovascular Risk Factors in Brazilian Families

Suely R. Giolo\*, Federal University of Parana, Brazil and Heart Institute, University of Sao Paulo, Brazil, Julia M. Soler, University of Sao Paulo, Brazil, Alexandre C. Pereira, Heart Institute, University of Sao Paulo, Brazil, Mariza de Andrade, Mayo Clinic, Jose E. Krieger, Heart Institute, University of Sao Paulo, Brazil

### 11:30 A Regularized Regression Approach for Dissecting Genetic Conflicts that Increase Disease Risk in Pregnancy

Yuehua Cui and Shaoyu Li\*, Michigan State University

### 11:45 Linkage Map Construction in Integrated Crosses

Emma Huang\* and Andrew George, CSIRO Mathematical and Information Sciences

### 12:00 Testing for Familial Aggregation of Functional Traits

Yixin Fang\*, Georgia State University and Yuanjia Wang, Columbia University

## 112. CONTRIBUTED PAPERS: VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA

INDEPENDENCE, 3RD FLOOR

Sponsor: ENAR

Chair: Jing Cao, Southern Methodist University

10:30

### Pairwise Variable Selection for High-dimensional Model-based Clustering and Its Application to Microarray Data

Jian Guo\*, Elizaveta Levina, George Michailidis and Ji Zhu, University of Michigan

### 10:45 Fast FSR Variable Selection with Interactions

Dennis D. Boos\*, North Carolina State University, Hugh B. Crews, SAS Institute Inc., Leonard A. Stefanski, North Carolina State University

### 11:00 Variable Selection in High-Dimensional Varying Coefficient Models via the Adaptive Group Lasso

Fengrong Wei\* and Jian Huang, University of Iowa

### 11:15 On the Study of LAD-Lasso in High-dimensional Settings

Xiaoli Gao\*, Oakland University and Jian Huang, University of Iowa

### 11:30 Variable Selection in the Kernel Machine Framework

Michael C. Wu\*, Tianxi Cai and Xihong Lin, Harvard University

### 11:45 Variable Selection for Ordinal Response Models with Applications to High Dimensional Data

Kellie J. Archer\* and Andre Williams, Virginia Commonwealth University

### 12:00 Choose An Optimal Ridge Parameter in Penalized Principal-Components Based on Heritability

Man Jin\*, American Cancer Society, Yuanjia Wang, Columbia University, Yixin Fang, Georgia State University

## 113. CONTRIBUTED PAPERS: CLUSTERED DATA METHODS

TRAVIS B, 3RD FLOOR

Sponsor: ENAR

Chair: Brent Logan, Medical College of Wisconsin

### 10:30 Inference for Variance Components in Generalized Linear Mixed Models for Pooled Binary Response

Joshua M. Tebbs\* and Peng Chen, University of South Carolina, Christopher R. Bilder, University of Nebraska-Lincoln

### 10:45 Association Models for Clustered Data with Bivariate Mixed Responses

Lanjia Lin\* and Debajyoti Sinha, Florida State University, Stuart Lipsitz, Harvard Medical School

### 11:00 Estimation Methods for an Autoregressive Familial Correlation Structure

Roy T. Sabo\*, Virginia Commonwealth University and N. R. Chaganty, Old Dominion University

# SCIENTIFIC PROGRAM

- 11:15** **Inference for Marginal Linear Models with Clustered Longitudinal Data with Potentially Informative Cluster Sizes**  
Ming Wang\*, Emory University, Maiying Kong and Somnath Datta, University of Louisville
- 11:30** **Quasi-Least Squares with Mixed Correlation Structure**  
Jichun Xie\* and Justine Shults, School of Medicine, University of Pennsylvania
- 11:45** **Generalized Varying Coefficient Single-Index Mixed Model**  
Jinsong Chen\*, Inyoung Kim and George R. Terrell, Virginia Polytechnic Institute and State University
- 12:00** **On Nonparametric Tests for Partially Correlated Data with Application to Public Health Issues**  
Hani M. Samawi\*, Lili Yu, Robert Vogel and Laura H. Gunn, Georgia Southern University

## 114. CONTRIBUTED PAPERS: ESTIMATION IN SURVIVAL MODELS

Sponsor: ENAR  
Chair: Yabing Mai, Merck & Co., Inc.

- 10:30** **Semiparametric Inference of Linear Transformation Models with Length-biased Censored Data**  
Jane Paik\* and Zhiliang Ying, Columbia University
- 10:45** **Intercept Estimation for the Semiparametric Accelerated Failure Time Model**  
Ying Ding and Bin Nan, University of Michigan
- 11:00** **Efficient Estimation in the Partly Linear Additive Hazards Regression Model with Current Status Data**  
Xuewen Lu\*, University of Calgary, Peter X.-K. Song and John D. Kalbfleisch, University of Michigan
- 11:15** **Efficient Estimation for the Proportional Odds Model with Bivariate Current Status Data**  
Bin Zhang\*, University of Missouri, Xingwei Tong, Beijing Normal University, Jianguo Sun, University of Missouri
- 11:30** **Semiparametric Cure Rate Models for Current Status Data**  
Guoqing Diao\*, George Mason University
- 11:45** **Accelerated Hazards Mixture Cure Model**  
Jiajia Zhang\*, University of South Carolina and Yingwei Peng, Queen's University

## 115. CONTRIBUTED PAPERS: BIOLOGICS, PHARMACEUTICALS, MEDICAL DEVICES

Sponsor: ASA Biopharmaceutical Section  
Chair: Edward J. Stanek III, University of Massachusetts-Amherst

- 10:30** **Proposed Methodology for Shelf Life Estimation**  
Michelle Quinlan\*, University of Nebraska-Lincoln, James Schwenke, Boehringer Ingelheim Pharmaceuticals, Inc., Walt Stroup, University of Nebraska-Lincoln
- 10:45** **Extending Group Sequential Methods to Observational Medical Product Safety Surveillance**  
Jennifer C. Nelson\*, Andrea Cook and Shanshan Zhao, Group Health Center for Health Studies; University of Washington
- 11:00** **A Conditional Maximized Sequential Probability Ratio Test for Pharmacovigilance**  
Lingling Li\* and Martin Kulldorff, Department of Ambulatory Care and Prevention, Harvard Medical School
- 11:15** **Selection and Evaluation of Biomarkers using Information Theoretic Approach**  
Abel Tilahun\*, Dan Lin, Suzy Van Sanden, Ziv Shkedy, Ariel Alonso and Geert Molenberghs, Universiteit Hasselt
- 11:30** **Patient Focused Method for Assessing In Vitro Drug Combinations Using Growth Rates**  
Maksim Pashkevich\*, Philip Iversen and Harold Brooks, Eli Lilly and Company
- 11:45** **Split-plot Designs in Serial Dilution Bioassay using Robots**  
Jeff Buzas\*, University of Vermont, Carrie Wager and David Lansky, Precision Bioassay
- 12:00** **Log-rank Test Weight Selection for Hazard Ratio with a Change-point**  
Jun Wu\*, Bristol Myers Squibb and Howard Stratton, School of Public Health, SUNY Albany



# ABSTRACTS

SUNDAY, MARCH 15

## POSTER PRESENTATIONS

### 1. POSTERS: CLINICAL TRIALS

Sponsor: ASA Biopharmaceutical Section

#### 1a. COMPARISON OF THREE ADAPTIVE DOSE-FINDING MODELS FOR COMBINATION THERAPY IN PHASE I CLINICAL TRIALS

Rui Qin\*, Mayo Clinic  
Yufen Zhang, University of Minnesota  
Sumithra J. Mandrekar, Mayo Clinic  
Wei Zhang, Boehringer Ingelheim  
Daniel J. Sargent, Mayo Clinic

Combination therapies become more popular in current medical research in pursuit of potential synergy in efficacy. Even though previous knowledge about each individual component may suggest proper dosages for combination therapy, a phase I clinical trial is still required for determining of the optimal dose combination to be further tested for efficacy. We have proposed three models incorporating both toxicity and efficacy within a Bayesian adaptive design in the dose-finding for combination therapy. Their performances under clinical relevant scenarios are evaluated by simulations. We conclude that all three models work reasonably well except for the extreme scenarios with significant non-monotone dose-efficacy curves. Further research are needed to improve the models to accommodate extreme scenarios.

email: qin.rui@mayo.edu

#### 1b. INCORPORATING PATIENT HETEROGENEITY IN ADAPTIVE PHASE I TRIAL DESIGNS

Thomas M. Braun\*, University of Michigan School of Public Health

There is often significant variability among patients enrolled in oncology Phase I trials with regard to the actual number and types of treatments they have received prior to enrolling in a Phase I trial. As a result, there often exists a non-negligible level of variability in the probability of dose-limiting toxicity (DLT) among patients receiving the same dose of an experimental agent. However, existing methodology for adaptive Bayesian designs assumes a homogeneous probability of toxicity among subjects receiving the same dose. We examine the impact of patient heterogeneity, parameterized through frailty models, on the performance of existing Phase I adaptive Bayesian designs and show that in most realistic settings, patient heterogeneity will lead to underestimation of the maximum tolerated dose (MTD). Such a result increases the likelihood of moving forward to a Phase II trial that examines a safe, yet ineffective dose of the

agent. We present both theoretic and simulation-based evidence and propose modifications to existing methods that account for patient heterogeneity and improve identification of the MTD.

email: tombraun@umich.edu

#### 1c. A PHASE 2 CLINICAL TRIAL WITH ADAPTATION ON 2 FACTORS

Richard J. McNally\*, Celgene Corporation  
David McKenzie, Celgene Corporation

A clinical trial is proposed with the dual objectives of finding both an effective and tolerable dose of a treatment and the optimal treatment schedule (e.g., 10, 15, or 20 days of treatment in a 28-day cycle) that will be used in a confirmatory study. Due to the large number of design points, with each point being a particular dose/schedule combination, an adaptive design is proposed. To find the best dose/schedule, techniques are discussed that combine elements from Bayesian dose-ranging studies and response surface methodology. We develop schemes for allocating patients to each design point at each study stage. We also develop rules for stopping the study, either for efficacy or futility, and selecting the best dose/schedule combination. Simulations are presented to show the operating characteristics of each design under both the universal null hypothesis (i.e., a flat response surface) and various alternatives. The study can be run independently or implemented as the “learning” phase of a seamless phase 2/3 trial.

email: rmcnally@celgene.com

#### 1d. ESTIMATING PERCENTILES IN DOSE-RESPONSE CURVES FOR DELAYED RESPONSES USING AN ADAPTIVE COMPOUND URN DESIGN

Rameela Chandrasekhar\*, University at Buffalo/Roswell Park Cancer Institute  
Gregory E. Wilding, University at Buffalo/Roswell Park Cancer Institute

Dose-response studies are conducted primarily in Phase II/III efficacy trials to estimate pre-specified dose percentiles or the underlying dose-response function of an investigational drug. A standard dose selecting trial randomizes subjects to several fixed dose groups and analyzes the data after all the responses have been obtained. Response-adaptive designs have been used as an alternative to the traditional randomization scheme to overcome the ethical and cost disadvantages. We focus our attention on the generalized Polya urn (GPU) model and its extensions. Often in clinical trials, patient responses are not instantly obtained, delaying the randomization of the subsequent subject in an adaptive design to the next dose level. To investigate the effect of delayed response in an adaptive urn design, we extend the compound urn model reviewed by Mugno, Zhus and Rosenberger (Statist. Med. 2004; 23:2137-2150) to a group sequential scheme. We estimate the percentiles of interest and evaluate its properties.

email: rc52@buffalo.edu

# ABSTRACTS

## 1e. PROPORTIONAL ODDS MODEL FOR DESIGN OF DOSE-FINDING CLINICAL TRIALS WITH ORDINAL TOXICITY GRADING

Emily M. Van Meter\*, Medical University of South Carolina  
Elizabeth Garrett-Mayer, Medical University of South Carolina  
Dipankar Bandyopadhyay, Medical University of South Carolina

Currently many phase I clinical trial designs including the continual reassessment method (CRM) dichotomize toxicity based on pre-specified dose-limiting criteria. Since phase I trials use small sample sizes, much information is lost by not accounting for different toxicity grades. Our proposed design incorporates ordinal toxicity endpoints as specified by Common Toxicity Criteria (CTCAE v3.0) and includes toxicity grades 1 and 2 not currently considered in standard designs. We extend the CRM to include ordinal toxicity outcomes using the proportional odds model and compare our results with the traditional CRM. Simulation studies and further sensitivity analysis of the new design comparing various weighting schemes, sample sizes, and cohort sizes show that the proposed proportional odds CRM does as well or better than the standard CRM and estimates the maximum tolerated dose (MTD) with more precision. Our findings suggest that it is beneficial to incorporate ordinal toxicities into phase I trial designs and future studies will compare this proposed design to other phase I trial designs.

email: vanmete@musc.edu

## 1f. SIMULATION OF THE OPTIMAL TIMING OF A MULTIPLE REGIME

Yining Du\*, Ning Wang, Texas Tech University  
Clyde Martin, Texas Tech University

A characteristic of treating a terminal disease with multiple drug therapies is the timing of the drug treatments. We develop a very simple simulation model for such treatments. Mathematically this work is based on the theory of switching systems. We assume that some measure of quality,  $x$ , is being measured and that it is scaled so that 0 represents death and 1 represents remission. For each treatment we assume a discrete linear model. Then for two treatments we have the combined model which switches between two linear systems. Dayawansa and Martin gave necessary and sufficient conditions for the stability of switching systems and in their proof constructed the worst case switching regime. This worst case regime is the best case for the problem here. We would like to switch the treatment in an effort to either drive the system to 1 or to delay as long as possible the inevitable approach to 0. We show that the best possible switching times occur when the system is driven by one treatment as far from zero as possible. This is in contrast to the usual practice. This result is a major deviation from the result of Dayawansa and Martin in that the model is stochastic. Dayawansa and Martin worked in continuous time and with deterministic models. We are able to give optimal strategies for switching both in the deterministic case and the stochastic case.

email: yining.du@ttu.edu

## 2. POSTERS: POWER/SAMPLE SIZE

### 2a. DETERMINATION OF SAMPLE SIZE FOR DEMONSTRATING EFFICACY OF RADIATION COUNTERMEASURES

Ralph L. Kodell, University of Arkansas for Medical Sciences  
Shelly Y. Lensing\*, University of Arkansas for Medical Sciences  
Reid D. Landes, University of Arkansas for Medical Sciences  
K. Sree, Kumar Armed Forces Radiobiology Research Institute  
Martin Hauer-Jensen, University of Arkansas for Medical Sciences

In response to the ever increasing threat of radiological and nuclear terrorism, active development of non-toxic drugs and other countermeasures to protect against or mitigate adverse radiological health effects is ongoing. Although the classical LD<sub>50</sub> study used for decades in preclinical toxicity testing has been largely replaced by experiments with fewer animals, the need to evaluate the radioprotective efficacy of new drugs necessitates the conduct of traditional LD<sub>50</sub> comparative studies (FDA, 2002). However, no readily available method exists for determining the number of animals needed to establish efficacy in such studies. We derived a sample-size formula for comparative potency testing of response modifiers in total body irradiation experiments incorporating elements of FDA's requirements for robust efficacy data where human studies are not ethical or feasible. Monte Carlo simulation demonstrated the formula's performance for Student's t, Wald, and Likelihood Ratio tests in logistic and probit models. Results showed clear potential for justifying use of substantially fewer animals than customarily studied. This work may thus initiate a dialogue among researchers using animals for radioprotection survival studies, institutional animal care and use committees, and drug regulatory bodies to arrive at a legitimate number of animals needed for statistically robust results.

email: sylensing@uams.edu

### 2b. WHEN ICCs GO AWRY: A CASE STUDY FROM A SCHOOL-BASED SMOKING PREVENTION STUDY IN SOUTH AFRICA

Ken Resnicow, School of Public Health, University of Michigan  
Nanhuia Zhang\*, School of Public Health, University of Michigan  
Roger D. Vaughan, Columbia University  
Sasiragha P. Reddy, Medical Research Council of South Africa, Cape Town, South Africa

It is common in public health interventions to randomize and then intervene with intact social groups, such as schools, churches, or worksites, rather than individuals. In Group Randomized Trials (GRTs), it is important to account for the correlation among the individuals in the same group, which is captured by the intraclass correlation coefficient (ICC). In designing Group Randomized Trials, ICCs must be accounted for to ensure adequate statistical power. In a school-based smoking prevention trial conducted in South Africa, the observed ICCs are considerably higher than those previously reported; some are as high as 0.10, which reduces our sample size by a factor of 16 for some psychosocial outcomes. In this paper, we investigate the causes for the high ICCs and use appropriate method to reduce the impact of the large ICCs. Reporting these ICCs helps future

investigators with sample size calculation if they want to conduct similar trials.

email: nhzhang@umich.edu

## **2c. SAMPLE SIZE DETERMINATION FOR A 5 YEAR LONGITUDINAL CLINICAL TRIAL IN CHILDREN: USING SIMULATION**

Yahya A. Daoud\*, Baylor Health Care System

Sunni A. Barnes, Baylor Health Care System

Dunlei Cheng, Baylor Health Care System

Ed DeVol, Baylor University Medical Center

C. Richard Boland, Baylor University Medical Center

As consulting statisticians we are often faced with unique and challenging problems. In a recent grant application we were faced with a request to estimate required sample size for a unique 5-year trial involving children. The overarching goal of the grant is to determine whether the JCV virus plays an active role in the development of colorectal polyps, which are the precursors for cancer. The specific hypothesis to be tested is that children with familial adenomatous polyposis (FAP) develop colonic polyps in association with infection by JCV. The challenge in this case is the fact that the probability of developing JCV and the conditional probability of developing polyps given JCV seroconversion both change with age. Since this is a 5-year study, the probability that a patient will get the JCV infection and then develop polyps increases each year they are in the study. There was no well established method for determining the required sample size for this study. We developed a simulation program using S-Plus to estimate the required sample size. We also compared our results with methods in the literature for estimating sample size when the probability of disease is constant by using mean probabilities and mean conditional probabilities.

email: sunni.barnes@baylorhealth.edu

## **2d. SAMPLE SIZE CALCULATION FOR CLUSTERED BINARY OUTCOMES WITH SIGN TESTS**

Fan Hu\*, Southern Methodist University

William Schucany, Southern Methodist University

Chul Ahn, University of Texas Southwestern Medical Center

We propose a sample size calculation for testing proportions using the sign test when binary outcomes are dependent within clusters. A sample size formula is derived using the test statistic of Datta and Satten (2008) for clustered binary outcomes accounting for the variability due to cluster size. A simulation study is conducted to evaluate the performance of the proposed sample size formula in terms of empirical type I errors and powers. This simulation study shows that empirical type I errors and powers are close to the nominal levels when the number of clusters is greater than 10. Our simulation study also shows that the number of clusters required increases as the imbalance in cluster size increases.

email: fhu@smu.edu

## **2e. SAMPLE SIZE CALCULATION FOR TWO-STAGE RANDOMIZATION DESIGNS WITH CENSORED DATA**

Zhiguo Li\*, Institute for Social Research, University of Michigan  
Susan Murphy, Institute for Social Research, University of Michigan

Some clinical trials are implemented via a two-stage approach. In the first stage, an initial treatment is given to induce remission, and in the second stage, a follow-up therapy is given in the intent of prolonging survival. A question for this kind of trials is how to calculate the sample size needed to guarantee a certain power in comparing different treatment policies. Assuming we want to compare two treatment policies, the sample size can be calculated based on several different test statistics. They include a weighted version of the Kaplan-Meier statistic, a weighted version of the Aalen-Nelson statistic, a weighted version of the empirical estimate of probabilities, and finally also a weighted version of the log-rank test statistic. Weights are necessary here because different subjects are consistent with a policy with different probabilities. Sample size formulas are obtained under different assumptions. Under some assumptions, less information is elicited from the investigator, but the sample size can be more conservative than that obtained when more information is elicited from the investigator. Simulation studies are carried out to compare all the sample size formula in different scenarios.

email: zhiguo@umich.edu

## **2f. COMPARE SAMPLE SIZE ADJUSTMENT METHODS FOR CLUSTER RANDOMIZED TRIALS**

Dhuly Chowdhury\*, RTI International

Hrishikesh Chakraborty, RTI International

Cluster randomized trials, where interventions or treatments are randomly assigned to clusters and conclusions are made in individual level. The individuals within a cluster are correlated and researchers need to account for the within cluster correlations during the sample size calculation and analysis. The required sample sizes for cluster randomized trials are adjusted to account for correlated observations within clusters. There are several statistical methods available to adjust sample size for clustering in cluster randomized trials and two widely used methods are (1) design effect adjustment method using intraclass correlation estimate and (2) adjustment based on between cluster coefficients of variation estimate. In this paper, we calculated the required sample sizes assuming the same hypothesis, power, and size to both adjustment methods for different design settings and conducted simulation exercises to compare the power of using the adjusted sample sizes. We only considered the binary outcomes for all simulation exercises. Our simulation results suggested that if the individuals within a cluster are highly correlated then the design effect adjustment method requires more clusters compared with the coefficients of variation adjusted method to conduct the same cluster randomized trial and if the correlation is low then both methods are equally powerful.

email: dchowdhury@rti.org

# ABSTRACTS

## 2g. PERFORMANCE OF THE HOCHBERG MULTIPLE TESTING PROCEDURE IN CLUSTER RANDOMIZED DESIGNS

Jeffrey J. Wing\*, University of Michigan  
Brisa N. Sánchez, University of Michigan  
Cathie Spino, University of Michigan

The Hochberg step-up procedure (1988) is a simple improvement to the Bonferroni multiple testing procedures which increases the power, preserves the familywise error rate, and maintains the ease of application. While its use is quite common in applications during the analysis stage, there is little guidance in the clinical trial design literature on how to incorporate the procedure into sample size calculations. We performed simulation studies to examine differences in Type I error rates and power under various clinical trial designs when the Hochberg procedure is used in the analysis, assessing both simple randomized and cluster randomized designs. We varied the number of outcome variables (i.e., comparisons) and their correlations and compared results with that of the traditional Bonferroni approach as a benchmark.

email: wingjeff@umich.edu

## 2h. COMPARISON OF SIMON'S TWO-STAGE DESIGN, SEQUENTIAL PROBABILITY RATIO TEST, AND TRIANGULAR TEST IN PHASE II CLINICAL TRIALS

Leigh A. Morton\*, University of Alabama-Birmingham  
David T. Redden, Ph.D., University of Alabama-Birmingham

The main objective of phase II clinical trials is to evaluate the efficacy of a new treatment in a manner that avoids enrolling a large number of patients when a drug is truly ineffective while providing adequate power to declare a drug effective when it is. Therefore, it becomes necessary to evaluate different study designs aimed at reducing expected sample size under the null hypothesis while maintaining power under the alternative hypothesis. Simon suggests a two-stage design in which the study is either terminated or continued by observing the number of responses in the first stage. The trial is continued into the second stage if a sufficient number of responses are observed. The sequential probability ratio test (SPRT), originally proposed by Wald, analyzes data after each new observation is obtained. One of three decisions is then reached: 1) accept the null hypothesis, 2) reject the null hypothesis, or 3) add additional observations and continue with trial based on open-ended boundaries. The triangular test is similar to the Wald's SPRT, but it implements analysis after a discrete number of observations are added and relies on more constrained stopping boundaries. Using simulations, we compare and contrast the three study designs with respect to expected sample size, type I error, and power.

email: bamalam@uab.edu

## 3. POSTERS: MICROARRAY ANALYSIS

### 3a. DIFFERENTIAL DNA METHYLATION: METHODOLOGY AND STUDY DESIGN

Richard E. Kennedy\*, Section on Statistical Genetics, University of Alabama-Birmingham  
Xiangqin Cui, Section on Statistical Genetics, University of Alabama-Birmingham

Studies to identify methylation differences between experimental and control (normal) tissues are increasingly common. Methodologies for examining differential methylation between two experimental conditions are less well developed, particularly for two-channel microarrays. In the latter, the two experimental conditions may serve as controls for each other. We analyzed publicly available datasets for the Nimblegen two-color methylation platform. Using the original datasets, methylation-enriched regions were identified using the Ringo package, and summaries for the degree of methylation were computed. These correlated well with the signal for the methylation-enriched channel, indicating that control DNA hybridizations may not be needed. Artificial differential methylation arrays were constructed by combining methylation-enriched channels from each disease state, which were also analyzed using Ringo. The regions identified using this approach overlapped significantly with regions identified using the original arrays. Taken together, these results offer initial support that separate control hybridizations for each array of differential methylation studies are not necessary.

email: rkennedy@ms.soph.uab.edu

### 3b. USE OF THE ITEM RESPONSE THEORY TO EVALUATE GENE EXPRESSION IN CONGENIC RAT STRAINS

Julia P. Soler\*, University of Sao Paulo, Brazil  
Carlos E. Neves, University of Sao Paulo, Brazil  
Suely R. Giolo, Federal University of Parana, Brazil  
Dalton F. Andrade, Federal University of Santa Catarina, Brazil  
Mariza de Andrade, Mayo Clinic  
Ayumi A. Miyakawa, Heart Institute, University of Sao Paulo, Brazil  
Jose E. Krieger, Heart Institute, University of Sao Paulo, Brazil

In the new field of genetical genomics, researches has focused attention in exploring the quantitative genetic of gene expression and also in relating transcriptional variation to variation in clinical traits. In this work, we explore the potential of this field on the analysis of a microarrays data set which considered measurements of gene expression of 35,129 fragments evaluated in samples from kidney tissue extracted from 5 rat strains (Spontaneously Hypertensive Rat (SHR) and four congenic rat strains) evaluated on two conditions of salt exposition (absence and presence of NaCl). The congenic rat strains were derived for 4 previously mapped blood pressure Quantitative Trait Loci (QTLs) in chromosomes 2 (two positions), 4, and 16. Thus, these animals have specific modifications in the genome that probably generate variations on the phenotypic blood pressure as well as on the gene expression pattern. We considered two strategies to analyze the microarrays data: a classical approach in terms of MAANOVA (MicroArrays Analysis of Variance), and a more robust strategy through Item Response Theory (ITR). Under the ITR

formulation we fitted the Rach and Samejima models, considering genes as individuals and the rat groups as items. As a result, we classify the genes in terms of their expression probability in each rat strain.

email: pavan@ime.usp.br

### **3c. DIFFERENTIATING mRNA EXPRESSION LEVELS OF TUMOR VERSUS NON-TUMOR CELLS IN A CANCER TISSUE**

Li-yu D. Liu\*, National Taiwan University

In recent years the microarray technology has been widely adopted in cancer research. The applications of microarray technology in cancer research include recapture of regulatory mechanism of oncogenes, development of prognosis classifiers, and drug discovery. Cancer research *in vivo* means that the experimentation has been conducted on the living tissue, from which the complex biological kinase can be directly observed. However, if the tissue under experiment is a mixture of tumor and non-tumor cells, the observed gene expressions in a microarray might not well represent the mRNA levels from tumor cell. Microscopy combined with laser-assisted microdissection provides a way to reduce the cellular heterogeneity of the tissue but laser microdissection units remain very expensive. In this study, we propose an alternative to estimate the mRNA levels of tumor cells with little cost to estimate from the microarray gene expression of the mixture tissue via normalization. The gene expression levels before and after normalization are used to conduct feature selection and classification, respectively, and the results will be compared to assess the validity of the proposed methods.

email: lyliu@ntu.edu.tw

### **3d. MOLECULAR CLASSIFICATION OF CANCER: CLASS DISCOVERY AND CLASS PREDICTION BY GENE EXPRESSION USING JAVA VISUALIZATION**

Rui Ding\*, University of Minnesota  
Morris Jong-Min Kim, University of Minnesota  
Morris Deukwoo Kwon, Division of Cancer Epidemiology and Genetics, National Cancer Institute

In this research, we do Java visualization approach to molecular classification of cancer based on gene expression and our new approach is applied to human acute leukemias as an example case. First, we perform a Bayesian variable selection for finding some meaningful genes from 7128 genes. And then with the selected genes, we do Java visualization of gene data for classification, since visualization is a good way to classify data. We randomly choose two genes out of the selected gene data to make a 2D figure, and also randomly choose three out of the selected gene data to make a 3D Figure, meanwhile, we do a dimension reduction from 7128 gene data to three-dimensional view by one of statistical multivariate methods, Principal Component Analysis. Finally, we create the Java visualization website of this research for other researchers who are interested in Molecular Classification of Cancer.

email: dingrui0417@gmail.com

### **3e. BIOMARKER DETECTION METHODS WHEN COMBINING MULTIPLE MULTI-CLASS MICROARRAY STUDIES**

Shuya Lu\*, University of Pittsburgh  
Jia Li, University of Pittsburgh  
George C. Tseng, University of Pittsburgh

As the microarray technology becomes mature in biomedical research, increasing number of datasets has been accumulated. Systematic information integration of multiple studies can provide improved biomarker detection. So far, published meta-analysis methods mostly consider two-class comparison. Methods for combining multi-class studies and pattern concordance are rarely explored. We first consider a natural extension of combining p-values from the traditional ANOVA model. Since p-values from ANOVA do not reflect the expression pattern information within classes, we propose a multi-class correlation measure (MCC) for biomarkers of concordant inter-class patterns across a pair of studies. For both approaches, we focus to identify biomarkers differentially expressed in all studies (ANOVA-maxP, min-MCC). Both ANOVA-maxP and min-MCC are evaluated by simulation studies and by applications to a multi-tissue mouse metabolism data set and a multi-platform mouse trauma data set. The results show complementary strength of the two methods for different biological purposes. When detecting only biomarkers with concordant inter-class patterns across studies, min-MCC has better power. If biomarkers with discordant inter-class patterns across studies are expected and are of interests together with concordant inter-class pattern genes, ANOVA-maxP better serve for the purpose.

email: lushuya@gmail.com

### **3f. STRATEGIES FOR APPLYING GENE SIGNATURES TO PROSPECTIVE CLINICAL STUDIES**

William T. Barry\*, Duke University Medical Center  
Michael Datto, Duke University Medical Center

Genome-wide expression profiling has been used extensively in pre-clinical cancer research, having led to many new insights regarding breast cancer biology and disease heterogeneity. However, many approaches for identifying gene signatures in association to tumor subtypes and clinical outcome are only applicable to retrospective analyses. In order to translate these discoveries into prognostic and predictive biomarkers, the algorithms must be modified to be used in a completely prospective manner. Herein, we generate microarrays on 51 replicate samples (18 patients) obtained from the Duke Breast SPORE tissue repository. Single-gene and multi-gene biomarkers are applied in a prospective manner to assess concordance and validation against standard immunohistochemical assays (IHC) of tumor biology. A single-gene predictor of ER shows complete concordance in 17 of 18 patients and >95% agreement with IHC; likewise, novel single-gene predictors of PR and EGFR are generated from IHC results. Multi-gene signatures of response to chemotherapies have been generated from a Bayesian probit regression model by Potti et al (2006) and will be applied in a randomized neoadjuvant breast cancer trial. Intra-class correlation for adriamycin and docetaxel signatures were 89% and 86%, respectively. Bootstrap-based confidence intervals illustrate the robustness of applying gene signatures in a prospective manner.

email: bill.barry@duke.edu

# ABSTRACTS

## 4. POSTERS: STATISTICAL GENETICS/GENOMICS

### 4a. HAPLOTYPE ANALYSIS OF QUANTITATIVE TRAITS IN OUTCROSSING PLANT POPULATIONS

Wei Hou\*, University of Florida

Rongling Wu, University of Florida and Penn State University

The genetic architecture of quantitative traits can be understood at the haplotype level. For natural outcrossing plant populations, the methods for discovering and modeling haplotypes associated with quantitative traits have not been sufficiently developed. We present a statistical model for haplotype analysis by incorporating genetic and biological characteristics of outcrossing plants. The model allows the simultaneous estimation of outcrossing rate, recombination fraction, haplotype frequencies, linkage disequilibria, and haplotype effects. The model is formulated with a mixture-based maximum likelihood and implemented with the EM algorithm. A Monte Carlo simulation was performed to test the statistical properties of the model, suggesting that the model is robust for parameter estimation. A series of hypothesis tests were formulated to study the pattern and amount of population genetic variation and the genetic architecture of complex traits for outcrossing plants.

email: whou@biostat.ufl.edu

### 4b. A MULTI-STEP APPROACH TO GENETIC ASSOCIATION FOR ASTHMA CHARACTERISTICS IN THE ISLE OF WIGHT BIRTH COHORT

Marianne Huebner\*, Mayo Clinic

Hasan Arshad, University of Southampton, UK

Eric Schauberger, Michigan State University

Karen Friderici, Michigan State University

Marsha Wills-Karp, Cincinnati Childrens Hospital

Wilfried Karmaus, University of South Carolina

Susan Ewart, Michigan State University

A subset of children in the Isle of Wight, UK, 1989-90 birth cohort ( $n=272$ ) were classified as having asthma based on physician diagnosis, positive reaction allergens by skin prick test, and bronchial hyperresponsiveness. SNPs were ranked by comparing hybridization intensities in a pooled genome wide association study in these regions. For the identified genomic region SNPs were genotyped in individual samples to capture the variation across the region. A multi stage survival analysis model was developed to determine the significance of SNPs for asthma characteristics and asthma development at ages 4, 10 and 18 years.

email: huebner.marianne@mayo.edu

### 4c. VALUE OF SNPs IN MODELS THAT PREDICT BREAST CANCER RISK

Mitchell H. Gail\*, Division of Cancer Epidemiology and Genetics, National Cancer Institute

Ruth M. Pfeiffer, Division of Cancer Epidemiology and Genetics, National Cancer Institute

The Breast Cancer Risk Assessment Tool (BCRAT) uses age at menarche, age at first live birth, number of previous biopsies and family history to project breast cancer risk. Recently, seven single nucleotide polymorphisms have been validated as associated with breast cancer risk. We compare a model that adds these seven SNPs, BCRATplus7, with BCRAT. We discuss the added value of BCRATplus7 in terms of discriminatory accuracy and in terms of reduced expected losses for: deciding whether to take tamoxifen to prevent breast cancer; deciding whether to have a mammogram; and allocating scarce resources for mammography. In all these applications, adding the seven SNPs improves the performance of BCRAT very little.

email: gailm@mail.nih.gov

### 4d. AN APPROACH TO DETECT GENE-GENE INTERACTIONS IN SNP DATA BASED ON PROBABILISTIC MEASURES

Ramon Casanova\*, Wake Forest University Health

Josh D. Grab, Wake Forest University Health

Miranda C. Marion, Wake Forest University Health

Paula S. Ramos, Wake Forest University Health

Jasmin Divers, Wake Forest University Health

Carl D. Langefeld, Wake Forest University Health

There is increasing evidence that the origin of complex diseases is related to multiple genes, gene-gene interactions and gene-environment interactions. This has motivated the rapid development of analytic tools to efficiently detect these interactions. This is a problem compounded by the huge size of genomic data sets and different sources of noise such as missing data, phenotypes, genetic heterogeneity, etc. To deal with such a degree of complexity, methods from machine learning and information theory are becoming more popular in the field. Here we propose a new approach for detection of gene-gene interactions in SNP data based on information theory and probabilistic measures. The key idea behind our approach is to probe the data for correlation changes between cases and controls that may be related to the disease. We show using case-control simulated data, that these measures in some situations could outperform a nonlinear support vector machine (SVM) classifier. These techniques are easy to implement and computationally very efficient. Applications to the large genome-wide association study and replication study in lupus ([www.SLEGEN.org](http://www.SLEGEN.org)) will be discussed.

email: casanova@wfubmc.edu

### 4e. CHANGE-POINT IDENTIFICATION IN HIDDEN MARKOV MODELS FOR DNA SEQUENCE SEGMENTATION MODELING

Darfiana Nur\*, University of Newcastle, Australia

Kerrie L. Mengersen, Queensland University of Technology, Australia

Many genome sequences display heterogeneity in base composition in the form of segments with similar structure. Early evidence of segmental genomic structure was noticed early on that in the salivary glands of *Drosophila melanogaster* whereas the problem of statistically segmenting DNA sequence has a history about four decades. One

approach describes DNA sequence structure by a hidden Markov model (HMM). Change-point detection is an identification of abrupt changes in the generated parameters of sequential data. It has proven to be useful in application such as DNA segmentation modeling. This talk focuses on the various change-point identifications of a Bayesian hidden Markov model describing homogeneous segments of DNA sequences. A simulation study will be used to evaluate the change-points followed by the real-life examples.

email: Darfiana.Nur@newcastle.edu.au

## 5. POSTERS: CAUSAL INFERENCE

### 5a. CAUSAL INFERENCE FOR INTERVENTION EFFECTS ON NICOTINE WITHDRAWAL SYMPTOMS

Brian L. Egleston\*, Fox Chase Cancer Center  
Karen L. Cropsey, University of Alabama School of Medicine  
Amy B. Lazev, Fox Chase Cancer Center  
Carolyn J. Heckman, Fox Chase Cancer Center

One problem with assessing effects of smoking cessation interventions on withdrawal symptoms is that symptoms are affected by whether participants abstain from smoking during trials. Those who enter a randomized trial but do not change smoking behavior might not experience withdrawal related symptoms. We provide a hypothetical example that demonstrates why estimating effects within observed abstention groups does not address this problem. We demonstrate how estimation of effects within groups defined by potential abstention that an individual would have in either arm of a study can provide meaningful inferences. We describe a methodology to estimate such effects, and use it to investigate effects of a combined behavioral and nicotine replacement therapy intervention on withdrawal symptoms in a prisoner population.

email: brian.egleston@fccc.edu

### 5b. ESTIMATION OF MARGINAL STRUCTURAL SURVIVAL MODELS IN THE PRESENCE OF COMPETING RISKS

Maarten Bekaert\*, Ghent University, Ghent, Belgium  
Stijn Vansteelandt, Ghent University, Ghent, Belgium  
Karl Mertens, Scientific Institute of Public Health, Brussels, Belgium

The effect on mortality of acquiring a nosocomial infection in the intensive care unit (ICU) is poorly understood because of informative censoring of the survival time by discharge from the ICU and because of time-dependent confounders, which lie on the causal path from infection to mortality. Standard statistical analyses may be severely misleading in such settings and have shown contradictory results. To accommodate informative censoring and because physicians' interest is often in 30-day ICU mortality, we will consider discharge from the ICU as a competing risk. To additionally accommodate time-dependent confounding, we propose marginal structural models for the counterfactual subdistribution hazard which express the subdistribution hazard that would be observed if the entire study population were to acquire infection at a given time in the ICU.

We develop inference for marginal structural subdistribution hazard models and use it to quantify the causal effect of nosocomial infection on mortality in the ICU using data from the National Surveillance Study of Nosocomial Infections in ICU's (Belgium).

email: maarten.bekaert@ugent.be

## 5c. A MARKOV COMPLIANCE CLASS AND OUTCOME MODEL FOR CAUSAL ANALYSIS IN THE LONGITUDINAL SETTING

Xin Gao\*, School of Public Health, University of Michigan  
Michael R. Elliott, School of Public Health, University of Michigan

We propose a Markov compliance class and outcome model for analyzing longitudinal randomized studies when non-compliance is present. In any longitudinal studies, subjects are randomized to the treatment or control group only at baseline, but subjects' compliance behaviors may vary over time. The proposed model considers the problem in the potential outcome framework, and provides causal estimates on the effect of the treatment within principal strata, which are a function of the subject's adherence to various possible randomization assignments. Previous research in this area (Lin, Ten Have, and Elliott 2008) considered the effect of subjects' joint compliance behavior on the joint distribution of the longitudinal outcomes, but not the effect of outcomes at time  $t-1$  on the compliance behaviors at time  $t$ , which is often of great interest to investigators. The proposed Markov compliance class and outcome model provides estimates both on the effect of the adherence on the following outcome, and on the effect of the outcome on the following adherence. The model requires assumptions to be made about the unobservable correlation among a subject's potential outcomes. We conduct a sensitivity analysis by varying the correlation. We analyze the longitudinal Suicide CBT Study using the proposed method and estimate the parameters and causal effects using both expectation-maximization (EM) and Markov chain Monte Carlo (MCMC) methodology.

email: xingao@umich.edu

## 6. POSTERS: IMAGING

### 6a. A BAYESIAN GENERALIZED NON-LINEAR PREDICTIVE MODEL OF TREATMENT EFFICACY USING qMRI

Jincao Wu\*, University of Michigan  
Timothy D. Johnson, University of Michigan

The prognosis for patients with high-grade gliomas is poor with a median survival of one year after diagnosis. The assessment of treatment efficacy is typically unavailable until about 8 to 10 weeks post treatment. Investigators hypothesize that recently developed Quantitative MRI (qMRI) techniques can predict the treatment efficacy only 3 weeks from the initiation of the therapy thereby allowing second line therapies to begin earlier. The purpose of this work is to build a predictive model for the treatment efficacy based on qMRI data and baseline prognostic factors. We use 1 year survival status as the outcome and propose a Bayesian joint model. In the first

# ABSTRACTS

stage, we smooth the qMRI data using a pairwise-difference prior and derive summary statistics. In the second stage, these statistics are used in a generalized non-linear model with a Multivariate Adaptive Regression Spline (MARS) basis in the systematic component and a probit link. Gibbs sampling and reversible jump Markov chain Monte Carlo are applied iteratively between the two stages to estimate the posterior. Bayesian model averaging is employed to derive the final predictive model.

email: jincaowu@umich.edu

## 6b. SPATIAL POINT PROCESS MODELING OF GROUP fMRI DATA

Timothy D. Johnson\*, University of Michigan  
Thomas E. Nichols, GlaxoSmithKline; University of Oxford, FMRIB;  
University of Michigan  
Lei Xu, Vanderbilt University  
Tor D. Wager, Columbia University

We propose a Bayesian hierarchical spatial model for multi-subject fMRI data. While there has been much work on univariate modeling of each voxel for single- and multi-subject data, and some work on spatial modeling for single-subject data, there has been virtually no work on spatial models that explicitly account for inter-subject variability in activation location. Most previous models use Gaussian mixtures for the activation shape. At the first level, we use Gaussian mixtures for the probability that a voxel belongs to an activated region. Spatial correlation is accounted for in the mixing weights. At the second level, mixture component means are clustered about individual activation centers and a priori are assumed to arise from a Cox cluster process. At the third level, individual activation centers are clustered about population centers, again arising from a Cox cluster process. At the fourth level, population centers are a priori, modeled as a homogeneous Poisson process. Our approach incorporates the unknown number of mixture components and individual centers into the model as parameters whose posterior intensities are estimated by reversible jump Markov Chain Monte Carlo. We demonstrate our method on a recently published fMRI study.

email: tdjtdj@umich.edu

## 6c. META-ANALYSIS OF fMRI DATA VIA A BAYESIAN COX CLUSTER PROCESS

Jian Kang\*, University of Michigan  
Timothy D. Johnson, University of Michigan  
Thomas E. Nichols, GlaxoSmithKline; University of Oxford, FMRIB;  
University of Michigan  
Tor D. Wager, Columbia University

Most functional magnetic imaging studies (fMRI) are small in size due to cost and difficulty in recruiting special patient populations. Since the same psychological paradigms are used in many studies, there is growing interest in meta-analyses of these data. Typical data available for fMRI meta-analyses consists of all activation foci from several experiments. To date the most widely used method is the Activation Likelihood Estimation (ALE) method, either on its own or as part of a larger multi-stage method. The ALE method has known

shortcomings, in particular only producing null-hypothesis inferences and providing no interpretable fitted model. In contrast, our model, a Bayesian spatial Cox cluster process model, provides an explicit fitted model and interpretable parameters. In particular our model provides information, via posterior intensity functions, about the most likely locations of activation centers at a population level and the inter-experiment spread of activated foci about these population centers. In our model, the observed activated foci are the offspring of a latent realization (population centers) of a parent process. A priori, the offspring arise from a Cox cluster process while the parent process is a homogeneous Poisson process. We demonstrate our method on an emotion activation meta-analysis of 169 studies.

email: jiankang@umich.edu

## 6d. EXTRACTION OF THE HEMODYNAMIC RESPONSE FUNCTION AND PARAMETER ESTIMATION FOR THE TWO GAMMA DIFFERENCE MODEL

Joel C. O'Hair\*, Southern Methodist University  
Richard F. Gunst, Southern Methodist University  
William R. Schucany, Southern Methodist University  
Wayne A. Woodward, Southern Methodist University

Stimulation of certain brain regions induces changes in the concentration of oxygen in those areas of the brain. The nature of this hemodynamic response is an essential component of understanding brain function. The purpose of this work is to find the best method of ascertaining the characteristics of the hemodynamic response from an fMRI time series. The structure of the hemodynamic response, referred to as the hemodynamic response function (HRF), can be estimated without assuming any mathematical form or by estimating the parameters of an appropriate model. Four nonparametric methods of extracting the HRF are compared using simulation methods. These four methods "extraction by deconvolution, Wiener filter extraction, LS-F extraction, and LS-T extraction" produce a nonparametric estimate of the HRF. There are five methods of parametric estimation of the HRF considered in this article. These include the Convolved HRF fit to the original fMRI time series, and a parametric model fit to the time series resulting from each of the four nonparametric extraction methods. These methods are also compared using simulation. It is shown that the LS-T extraction often provides the most desirable nonparametric estimation of the HRF, while the Convolved HRF fit usually provides the best parametric estimation.

email: imrunnin2win@yahoo.com

## 6e. WAVELET PACKET RESAMPLING FOR fMRI EXPERIMENTS

Ohn Jo Koh\*, Southern Methodist University  
William R. Schucany, Southern Methodist University  
Richard F. Gunst, Southern Methodist University  
Wayne A. Woodward, Southern Methodist University

Identification of activated brain regions in response to external stimuli is important for understanding human brain activity. The 4D spatiotemporal wavelet packet resampling method by Patel

et al. generates null data that preserve average background spatial correlation better than the wavelet resampling method. Activated regions are determined by testing observed data against these null distributions. Instead of resampling axial slices first and then across axial slices, a newly developed method resamples three dimensions of the brain using a 3D wavelet packet decomposition. A new statistic that measures spatial decorrelation is investigated for determining an orthogonal basis for wavelet packet decomposition. In addition, a resampling method that generates the null distribution for comparison of control versus treatment is examined.

email: okoh@smu.edu

## 6f. DIRICHLET PROCESS MODELS FOR CHANGES IN fMRI VISUAL FIELD

Raymond G. Hoffmann\*, Medical College of Wisconsin  
Pippa Simpson, Medical College of Wisconsin  
Shun-Hwa Li, Medical College of Wisconsin  
Ke Yan, Medical College of Wisconsin  
Edgar A. DeYoe, Medical College of Wisconsin  
Daniel B. Rowe, Medical College of Wisconsin

The Visual Field Map (VFM) is a circular region that maps the visual cortex to a virtual retina. The relationship between the dynamic image presented to the eye and the virtual retina can be used to identify changes in the visual system. These changes could be a result of surgery near the components of the visual system or result from progression of a chronic disease. The visual field map is a nonisotropic, nonhomogeneous set of points that represent the activation of voxels in the visual cortex assessed in an fMRI scanner. A wedge shaped mask (18 to 90 degrees of arc) of the image is used to simulate the effect of surgical damage. A Bayesian non-parametric mixture model, a Dependent Dirichlet Process, uses a Dirichlet prior on a space of 2D density functions  $G$  to model the intensity of the stochastic process that generates the points in the VFM, an inverse gamma on the precision of the distribution and a gamma prior on the mixing parameter for the number of the distributions needed to model the DDP. The posterior probability of the DDP model on the disk quantifies the probable location of the wedge-shaped mask compared to a reference scan.

email: rhoffmann@mcw.edu

## 7. POSTERS: SURVIVAL ANALYSIS

### 7a. ANALYZING PATIENT SURVIVAL AFTER DECEASED-DONOR KIDNEY TRANSPLANTS: THE NOVEL USE OF TIME-VARYING COVARIATES

Arwin M. Thomasson\*, University of Pennsylvania  
Peter P. Reese, University of Pennsylvania  
Justine Shults, University of Pennsylvania

Analysis of patient survival after deceased-donor kidney transplantation has typically focused on kidneys received from standard criteria donors. However, due to the limited availability of high-quality organs, clinicians have become increasingly interested in the outcomes associated with potentially less-desirable organs. In

particular, there is a focus on kidneys from donors with an increased risk of HIV infection, from donors who experienced cardiac death, and from donors with other undesirable health characteristics. In this presentation we describe the statistical methods used in analysis of a retrospective cohort study using data from the Organ Procurement and Transplantation Network. In particular, we focus on the use of logistic regression and non-proportional Cox regression for a comparison of rates of transplant rejection and patient survival for the different donor types. In addition, we demonstrate the application of quasi-least squares regression for multiple correlated binary outcomes to analyze absolute two-year allograft survival and delayed graft function.

email: arwin@mail.med.upenn.edu

### 7b. SURVOMATIC: A USER-FRIENDLY PACKAGE FOR ANALYSIS OF SURVIVAL AND MORTALITY DATA

Alex F. Bokov\*, University of Texas Health Science Center at San Antonio  
Scott D. Pletcher, Baylor College of Medicine, Department of Molecular and Human Genetics and Huffington Center on Aging  
Jonathan A.L. Gelfond, University of Texas Health Science Center at San Antonio

The log-rank test is widely used in biomedical research for determining whether two distributions of survival times are significantly different from one another. However, the log-rank test can fail to detect even a large difference between survival curves if they cross or if one group dies at younger or older ages than the other only during a certain segment of the overall lifespan for that population. This is particularly problematic in the field of aging and longevity research, where the focus is on survival at extreme ages rather than just mean or median survival. Quantile regression and fitting of mortality models (such as the Gompertz model) to the data are both alternative approaches which are more sensitive and more robust to complex differences between survivorship functions. Here we present an open source, cross-platform, R-based software package that incorporates two quantile regression tests as well as maximum likelihood estimation of best-fit mortality parameters. This package comes with an optional graphical interface that makes it possible for researchers to use without having to know the R statistical language. Download site: <http://rmodest.r-forge.r-project.org/>.

email: dxykbay02@sneakemail.com

### 7c. ASSOCIATION BETWEEN PROGRESSION-FREE AND OVERALL SURVIVAL IN RANDOMIZED CLINICAL TRIALS

Kristine Broglio\*, U.T. M.D. Anderson Cancer Center  
Donald Berry, U.T. M.D. Anderson Cancer Center

Overall survival (OS) is the gold standard endpoint for new drug development in oncology. Progression-free survival (PFS) is a more subjective endpoint, but is assessed prior to OS, allowing for perhaps smaller and faster studies. There is debate over whether PFS can be considered a surrogate marker for OS and whether PFS is a measure

# ABSTRACTS

of direct clinical benefit. We evaluated the relationship between PFS and OS in three settings including 1) trial design 2) trial analysis and 3) meta-analysis. We simulated survival data for hypothetical clinical trials where patients were randomized to either a standard of care or an experimental treatment arm. We assumed that median PFS was 6 months for standard of care and 12 months for the experimental regimen. OS was considered as the sum of PFS and survival subsequent to disease progression (SS). We calculated 1) the probability of finding a statistically significant difference in OS for varying lengths of SS 2) the probability of finding a statistically significant difference in OS based on the observed p-value for PFS 3) the association between hazard ratios for PFS and OS in a meta-analysis setting. When SS is 12 months, the probability of detecting a significant OS benefit is 33%, 37% and 32% for studies designed to have 80%, 85%, and 90% power to detect the 6 month difference in PFS respectively. If the p-value for PFS observed in a particular study is highly statistically significant ( $p = 0.0001$ ), the probability of also estimating a statistically significance OS benefit ranges from 90% if SS is 4 months to 30% when SS is 24 months. As SS increases, the strength of the association between the estimated treatment effects for PFS and OS will decrease. Even when there is a true treatment benefit as measured by PFS, the probability of observing a statistically significant benefit as measured by OS depends on the size of the observed PFS benefit and the play of chance, but most importantly, the length of SS.

email: kbroglion@mdanderson.org

## 7d. ON AN EMPIRICAL METHOD FOR A GENERALISED VERSION OF THE YANG AND PRENTICE MODEL

Carl M. DiCasoli\*, North Carolina State University  
Sujit K. Ghosh, North Carolina State University  
Subhashis Ghosal, North Carolina State University

In survival data analysis, the proportional hazards (PH), accelerated failure time (AFT), and proportional odds models (POM) are commonly used semiparametric models for the comparison of survivability in subjects. These models assume that the survival curves do not cross. However, in some survival applications, the survival curves pertaining to the two groups of subjects under the study may cross each other. Hence, these three models stated above may no longer be suitable for making inference. Yang and Prentice (Biometrika, 2005 92(1):1-17) proposed a model which separately models the short-term and long-term hazard ratios generalising both PH and POM. This feature allows for the survival functions to cross. We study the estimation procedure in the Yang-Prentice model using the empirical likelihood approach. This method is extended to a regression version involving predictors, where the posterior sample is computed. Good properties of this method are also examined.

email: cdpiano27@hotmail.com

## 7e. BAYESIAN HAZARD RATE ESTIMATION AND SUFFICIENT DIMENSION REDUCTION

Shraddha S. Mehta\*, Purdue University

Surya T. Tokdar, Carnegie Mellon University

Jayanta K. Ghosh, Purdue University and Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, Kolkata, India  
Bruce A. Craig, Purdue University

Logistic Gaussian process priors have been used in nonparametric Bayesian density estimation and sufficient dimension reduction in a density regression setting. In this talk, we will describe the extension of this approach to survival modeling with right-censored data. The method simultaneously estimates the covariate subspace as well as the hazard rate given this subspace. As a result, this method overcomes the dimensionality problem and the inability to estimate non-linear covariate effects when using the traditional proportional hazards model. In addition, sufficient dimension reduction approach provides the most parsimonious representation of the covariate space in relation to the hazard. Currently, the hazard rate is estimated using a partial likelihood approach. Simulation studies are used to compare this method, random survival forests, and Cox's proportional hazards model using variable importance measures and estimates of treatment effect. Future work will utilize the full-likelihood instead of partial likelihood in estimating the hazard rate.

email: ssmehta@purdue.edu

## 7f. ESTIMATING MEDIATION EFFECTS IN SURVIVAL ANALYSIS WITH CENSORED DATA

Yanhui Sun\*, University of Alabama at Birmingham  
Chichi Aban, University of Alabama at Birmingham  
Gary R. Cutter, University of Alabama at Birmingham  
David L. Roth, University of Alabama at Birmingham

This study examines the estimates of mediated effect calculated by the product of coefficients method using two survival analyses: log-survival and log-hazard time models. Study duration and loss-to-follow-up time are introduced to simulate and control the amount of censored data in the analysis. The mediation effects assessed by statistical tests varying sample size, study duration and parameter combinations were examined using Sobel first order formula, PRODCLIN method, Goodman unbiased formula and empirical method. Our results show all four methods yielded similar results that all the tests with study duration=5 and sample size >500 and over 80% of the tests with study duration=3 show significant mediation effects. When standard errors fall between 0 and 1, both Goodman and Sobel methods produce smaller estimates of standard errors than empirical method in both procedures. Standard errors were comparable between log-survival and log-hazard models. When standard errors fall between 0 and 1, standard errors from log-hazard model are about 4 times larger than log-survival model. No significant difference was found for testing mediation effects among four methods within same study duration and same sample size. The amount of censored data affects results most. Log-survival model is recommended since it produces more stable standard errors.

email: yanhui@uab.edu



## **7g. JOINT MODELING OF SURVIVAL AND BINOMIAL DATA BASED ON GENERALIZED SELF-CONSISTENCY WITH APPLICATION TO PROSTATE CANCER STAGE-SPECIFIC INCIDENCE**

Chen Hu\*, School of Public Health, University of Michigan  
Alex Tsodikov, School of Public Health, University of Michigan

Stage-specific cancer incidence represents a random vector of joint bivariate response represented by the age at diagnosis, and cancer stage. Factors affecting the unobserved tumor progression and the history of metastasis before diagnosis are of particular interest. Semiparametric models with time-dependent covariates are considered for the joint response. We extend the framework of the generalized self-consistency approach (Tsodikov 2003 JRSSB) and use EM algorithm for maximum likelihood estimation and model building. This method is illustrated by real data from the Surveillance, Epidemiology and End Results (SEER) program and by simulation studies.

email: chenhu@umich.edu

## **7h. SURVIVAL ANALYSIS WITH ERROR PRONE TIME-VARYING COVARIATES: A RISK SET CALIBRATION APPROACH**

Xiaomei Liao\*, Harvard School of Public Health  
David Zucker, Hebrew University, Jerusalem, Israel  
Yi Li, Harvard School of Public Health  
Donna Spiegelman, Harvard School of Public Health

Occupational and environmental epidemiologists are often interested in estimating the prospective effect of time-varying exposure variables such as the cumulative exposure or average cumulative exposure, in relation to chronic disease endpoints such as cancer incidence and mortality. From exposure validation studies, it is apparent that many of these variables are measured with moderate to substantial error. Although the ordinary regression calibration approach is valid and efficient for measurement error correction of relative risk estimates from the Cox model with time-independent point exposures when the disease is rare, it is not adaptable for use with time-varying exposures. By recalibrating within each risk set, the risk set regression method is proposed for this setting. An algorithm for a bias-corrected point estimate of the relative risk using an RRC approach is presented, followed by the derivation of an estimate of its variance, resulting in a sandwich estimator. Emphasis is on methods which apply to the main study/external validation study design. Simulation studies with different error models are carried out to show the validity and efficiency of the method, compared to the 'naïve' cox model, and the method is applied to a study of diet and cancer from the Nurses' Health Study.

email: xliao@hsph.harvard.edu

## **7i. LONGITUDINAL CHANGES IN CAROTID IMT AND RISK OF MI, STROKE AND CHD: THE CARDIOVASCULAR HEALTH STUDY**

David Yanez, University of Washington  
Michal Juraska\*, University of Washington  
Bruce M. Psaty, University of Washington  
Mary Cushman, University of Vermont  
Cam Solomon, CHSCC, University of Washington  
Joseph F. Polak, Tufts University  
Daniel O'Leary, Tufts University

Ordinary regression calibration (ORC) is a popular method employed as a bias-correcting technique in regression models with covariate measurement error. In the analysis of failure time data, ORC does not account for possible dependence of survival or censoring probabilities on mismeasured covariates, suggesting a potential bias reduction by recalibrating within each risk set (RSRC). Xie et al. (2001) proposed such a method for measurement error correction for time independent predictors. The purpose of this work is to evaluate differences in ORC and RSRC in a Cox regression model with mismeasured time-dependent covariates. To this end, an association study of longitudinal change in carotid intima-media thickness (IMT) and the risk of MI, stroke and CHD is presented. As one might expect, ignoring measurement error induced bias in the analysis leads to erroneous conclusions. Both the RSRC and ORC method produce greater cross-sectional and smaller longitudinal effects of IMT on the risk of each CVD event as compared to the naïve procedure. A reduction in bias is accompanied by an increase in the standard error of the longitudinal IMT effect whereas the precision of the cross-sectional IMT effect remains stable across the considered methods.

email: mjuraska@u.washington.edu

## **8. POSTERS: MISSING DATA**

### **8a. A HOT-DECK MULTIPLE IMPUTATION PROCEDURE FOR GAPS IN LONGITUDINAL EVENT HISTORIES**

Chia-Ning Wang\*, University of Michigan  
Roderick Little, University of Michigan  
Bin Nan, University of Michigan  
Sioban Harlow, University of Michigan

In many longitudinal cohort studies or clinical trials, subjects are assessed for the transition to an intermediate state and to a final event, such as an occurrence of a disease-related non-fatal event and death, respectively. In our specific application, the intermediate state is a measure of menopausal transition based on information on menstrual cycles, and the final event is the final menstrual period (FMP). The distribution of the occurrence time for the intermediate state and the distribution of the duration from the intermediate state to the final event are two primary research interests of the data analysis. However, a difficulty arises when some subjects have gaps in their event history. These gaps can create problems in determining the time of transition to the intermediate state, and simple approaches such as ignoring gap times or dropping cases with missing gaps have obvious limitations. A better approach is to impute the missing information for the gaps.

# ABSTRACTS

In this study, predictive mean matching is used to multiply impute by matching gaps to completely recorded histories, conditional on longitudinal characteristics, and the time of FMP (which may be censored). This procedure is applied to an important data set for assessing various measures of menopausal transition and FMP.

email: cnwang@umich.edu

## 8b. A MULTIPLE IMPUTATION APPROACH TO THE ANALYSIS OF INTERVAL-CENSORED FAILURE TIME DATA WITH THE ADDITIVE HAZARDS MODEL

Ling Chen\*, University of Missouri-Columbia  
Jianguo Sun, University of Missouri-Columbia

This paper discusses regression analysis of interval-censored failure time data, which occur in many fields including demographical, epidemiological, financial, medical, and sociological studies. For the problem, we focus on the situation where the survival time of interest can be described by the additive hazards model and a multiple imputation approach is presented for inference. A major advantage of the approach is its simplicity and it can be easily implemented by using the existing software packages for right-censored failure time data. Extensive simulation studies are conducted and indicate that the approach performs well for practical situations and is comparable to the existing methods. The methodology is applied to a set of interval-censored failure time data arising from an AIDS clinical trial.

email: lcbm9@mizzou.edu

## 8c. ASSESSING THE CONVERGENCE OF MULTIPLE IMPUTATION ALGORITHMS USING A SEQUENCE OF REGRESSION MODELS

Jian Zhu\*, University of Michigan  
Trivellore E. Raghunathan, University of Michigan

Multiple imputation algorithms using a sequence of regression models, or multiple imputation using chained equations, are commonly used to handle non-responses in complex survey studies. Although such algorithms have several advantages over joint modeling of all survey variables, they have a theoretical limitation that the specified conditional distributions could be incompatible and therefore the underlying joint distribution to which the algorithms attempt to converge may not exist. Previous simulation studies (van Buuren et al 2006) show that multiple imputation algorithms using incompatible conditional distributions seem to work well for some cases. We focus on general multivariate data to assess the convergence properties of the imputation algorithms using various types of conditional distributions. Additionally, we evaluate the impact of incompatible models on imputation results through simulation studies. We also use a longitudinal study with a general missing pattern to illustrate the performance of the imputation algorithms.

email: jianzhu@umich.edu

## 8d. A COMPARISON OF MISSING DATA METHODS FOR QUALITY OF LIFE MEASURES IN A CLINICAL TRIAL WITH LONG-TERM FOLLOW-UP

Paul Kolm\*, Christiana Care Health System  
Wei Zhang, Christiana Care Health System  
John A. Spertus, Mid America Heart Institute  
David J. Maron, Vanderbilt University  
William E. Boden, Buffalo General Hospital  
William S. Weintraub, Christiana Care Health System

The primary outcomes of clinical trials are typically well-defined events such as survival, non-fatal cardiovascular events (e.g., stroke, myocardial infarction) and recurrence of cancer. Many trials also include secondary measures assessing patient quality of life and health status. The latter types of outcomes are assessed at different times over the length of the trial or an additional follow-up period. Inevitably, missing patient values occur either intermittently or through dropout. Generally, more missing values occur with increasing length of follow-up and present a challenge for longitudinal analyses. Analytic methods for handling missing data have advanced in recent years with respect to statistical sophistication. In this study, we compare results of applying several methods for missing data, including last value carried forward, missing at random, pattern mixture and propensity scores, to a quality of life measure obtained from patients in a large cardiovascular clinical trial spanning up to 7 years of follow-up.

email: pkolm@christianacare.org

## 8e. ANALYSIS OF NON-IGNORABLE MISSING AND LEFT-CENSORED LONGITUDINAL BIOMARKER DATA USING WEIGHTED PSEUDO LIKELIHOOD METHOD

Abdus Sattar\*, University of Pittsburgh  
Lisa Weissfeld, University of Pittsburgh

In a longitudinal study of biomarker data collected during a hospital stay, observations may be missing due to administrative reasons, the death of the subject or the subject's discharge from the hospital, resulting in non-ignorable missing data. In addition to non-ignorable missingness, there are left-censoring in biomarker measurements due to the inherent limit of detection and the quantification limit of the bioassays. Standard likelihood-based methods for the analysis of longitudinal data, e.g. mixed model, do not include a mechanism that accounts for the different reasons for missingness and left-censoring. We have proposed to extend the theory of random effects Tobit regression for the left-censored data to weighted random effects Tobit regression using weighted pseudo likelihood theory for the non-ignorable missing and left-censored longitudinal biomarker data. The proposed method is applied to non-ignorable missing and left-censored interleukin-6 (IL-6) biomarker data obtained from the Genetic and Inflammatory Markers of Sepsis (GenIMS) study, a large, multicenter, cohort study. An extensive simulation study was performed to compare the performance of the proposed model with a number of widely used models.

email: mas196@pitt.edu

## **8f. 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 on the outcomes between variables measured at different levels are of interest. Missing data can complicate analysis using hierarchical models and can occur at all levels, in both outcomes and covariates. Ignoring the subjects with missing data usually leads to biased estimates, yet less attention has been paid to the analysis based on hierarchical models with incomplete data. We use a combination of the EM algorithm and multiple imputation to develop approximate maximum likelihood estimates of the parameters in hierarchical models, assuming missing at random (MAR) and ignorable missing mechanism (Rubin, 1976; Little and Rubin, 2002). In this paper we consider a binary response with missing values as well as continuous and binary covariates with missing values at each level. Simulation study is used to demonstrate that our proposed method has desirable repeated sampling properties. The method is also applied to a survey data.

email: yonzhang@umich.edu

## **8g. AN APPROACH TO SENSITIVITY ANALYSIS OF NESTED CASE-CONTROL STUDIES WITH OUTCOME-DEPENDENT FOLLOW-UP**

Kenneth J. Wilkins\*, Infectious Disease Clinical Research Program, University of the Health Sciences

This paper presents a likelihood-based approach to case-control analysis, appropriate when conducted within a well-defined cohort that (1) affords estimation of disease incidence, and (2) involves distinct stages of outcome-dependent follow-up. A Department of Defense / Veterans Affairs (VA) trauma casualty cohort motivates this method, with nested sampling of cases and controls from medical databases to target estimation of disease-exposure association. Subjects are followed from trauma through the evacuation chain to military hospitals for bone infection (osteomyelitis). Late-developing cases are ascertained within the VA system. However, the probability of additional VA follow-up may involve some dependence on disease risk; estimates of overall risk are thus sensitive to unverifiable assumptions about this dependence. Using cohort-based estimates of disease incidence by stage, the proposed approach adapts the case-control weighted targeted maximum likelihood estimation of Rose & van der Laan (2008) to examine how estimated associations change under distinct outcome/follow-up dependency assumptions posited within a sensitivity analysis. The paper concludes with simulations and an illustration using the U.S. Military HIV Natural History Study cohort.

email: kwilkins@post.harvard.edu

## **8h. MISSING ANIMALS IN TOXIC TREATMENT STUDIES**

Pippa M. Simpson\*, Medical College of Wisconsin  
Shun H. Li, Medical College of Wisconsin  
Ke Yan, Medical College of Wisconsin  
Bevan E. Huang, CSIRO  
Calvin Williams, Medical College of Wisconsin  
Dipeca Haribhai, Medical College of Wisconsin  
Raymond G. Hoffmann, Medical College of Wisconsin

Animals often have missing outcomes in studies of toxicity. Not infrequently it is because they become moribund and must be sacrificed. When death is not the outcome of interest, the primary outcome may be missing because of its value or other factors related to the outcome. Under these conditions the missing pattern is not MAR, but MNAR, and must be modeled. For example, a study of a carcinogen may have the number of tumors palpated in mice over time as an outcome, but the mice are sacrificed when the tumor burden is too great. Our analysis is motivated by a study of genetically engineered mice treated with different T regulatory cells with the weight pattern over time as an outcome of interest. One of the main criteria for sacrificing mice was excessive weight loss. In this case there is information about the possible mechanism for missing data. We use a nonlinear random coefficient model and look at various ways to model the dropout pattern. This includes a logistic regression model for dropout at each time conditional on the mouse still being in the study, on previous values of weight and possibly other variables.

email: pippam.simpson@gmail.com

## **8i. PSEUDOLIKELIHOOD RATIO TESTS WITH BIASED OBSERVATIONS**

Bin Zhang\*, Boston University  
Joan X. Hu, Simon Fraser University

This paper considers pseudolikelihood ratio tests with biased observations using auxiliary information. The pseudolikelihood functions are constructed without specifying the association between the primary variable and the auxiliary variables. We derive the asymptotic distributions of the test statistics and examine finite-sample properties of the testing procedures via simulation. The methodology is illustrated by an example involving kindergarten readiness skills in children with sickle cell disease.

email: binzhang@bu.edu

# ABSTRACTS

## 9. POSTERS: SPATIAL/TEMPORAL MODELING AND ENVIRONMENTAL/ECOLOGICAL APPLICATIONS

### 9a. ESTIMATING THE MAXIMUM GROWTH RATE OF HARMFUL ALGAL BLOOMS USING A COMBINED MODEL METHOD

Margaret A. Cohen\*, University of North Carolina at Wilmington

This paper proposes a creative new method of estimating the maximum growth of harmful algal blooms. Traditionally marine scientists have calculated the maximum growth rate using a linear method which can be influenced by the choice of endpoints. A more objective statistical method of estimating growth in a sigmoidal curve using the rate at the point of inflection was presented, but the estimates were viewed as too large. In response, we proposed a hybrid approach that was the combination of the Logistic, Weibull, and Gompertz models. This paper illustrates this combined method along with examples of harmful algal blooms.

email: mac1166@uncw.edu

### 9b. A SPATIAL SCAN STATISTIC FOR MULTINOMIAL DATA

Inkyung Jung\*, University of Texas Health Science Center at San Antonio

Martin Kulldorff, Harvard Medical School

Otuuke J. Richard, Makerere University, Kampala, Uganda

As a geographical cluster detection analysis tool, the spatial scan statistic has been developed for different types of data such as Bernoulli, Poisson, ordinal, exponential and normal. Another interesting data type is multinomial. For example, one may want to find clusters where the disease type distribution is statistically significantly different from the rest of the study region when there are different types of disease which have no ordinal structure. In this paper, we propose a spatial scan statistic for such data, which is useful for geographical cluster detection analysis for categorical data without any intrinsic order information. The proposed method is illustrated using meningitis data in two counties of UK and the performance of the method is evaluated through a simulation study.

email: inkyung.jung@gmail.com

### 9c. LONGITUDINAL SPATIAL POINT PROCESSES FOR RESIDENTIAL HISTORIES

Patrick E. Brown\*, Cancer Care Ontario  
Peter Henrys, Lancaster University

An individual's location at time of diagnosis can be useful information for an epidemiological study, as clustering of case locations together relative to the population's distribution could indicate an environmental spatially-structured risk factor. However, individuals could have changed residences between exposure and diagnosis, or exposure could be cumulative over years spent in different residences. In these cases, it is clustering of residential histories rather than of single locations which should be examined. This talk describes a

longitudinal spatial point process for modeling residential histories. Statistical properties of this process are derived and measure of clustering based on the K function is presented. A simulation study shows the ability of this methodology to detect different types of spatial dependence, and the method is a real dataset of residential histories related to lung cancer cases in the city of Winnipeg.

email: patrick.brown@utoronto.ca

### 9d. HIERARCHICAL DYNAMIC MODELING OF SPATIAL-TEMPORAL BINARY DATA

Yanbing Zheng\*, University of Kentucky

Jun Zhu, University of Wisconsin-Madison

Brian Aukema, Natural Resources Canada, Canadian Forest Service and Ecosystem Science and Management Program, University of Northern British Columbia

We develop spatial-temporal generalized linear mixed models for spatial-temporal binary data observed on a spatial lattice and repeatedly over discrete time points. To account for spatial and temporal dependence, we introduce a spatial-temporal random effect in the link function and model by a diffusion-convection dynamic model. We propose a Bayesian hierarchical model for statistical inference and devise Markov chain Monte Carlo algorithms for computation. We illustrate the methodology by an example of outbreaks of mountain pine beetle on the Chilcotin Plateau of British Columbia, Canada. We examine the effect of environmental factors while accounting for the potential spatial and temporal dependence.

email: yanbing.zheng@uky.edu

### 9e. TWO-STAGE GENERALIZED METHOD OF MOMENTS ESTIMATION WITH APPLICATIONS IN SPATIO-TEMPORAL MODELS

Yun Bai\*, University of Michigan

Peter X.K. Song, University of Michigan

Trivellore Raghunathan, University of Michigan

Spatio-temporal process modeling has received increasing attention in recent statistical research. However, due to the high dimensionality of the data, likelihood-based approaches for estimation of the spatio-temporal covariance structure are computationally prohibitive. In this paper, we propose a two-stage generalized method of moments (GMM) method to estimate spatio-temporal covariance structures. Estimating equations are formulated separately for spatial and temporal processes using pair-wise composite likelihood, which significantly reduces the dimensionality. This often results in a larger set of estimating equations than the number of parameters, so we apply GMM to construct a quadratic inference function to for estimation. The optimal weight matrix is the covariance between the spatial and temporal score functions, which accounts for spatio-temporal inter-correlation and hence improves the efficiency of parameter estimation. To deal with the issue of numerical instability, a well-known difficulty in the implementation, we propose a two-stage estimation procedure, in a similar spirit to the method of inference functions for margins (IFM). Theoretically, the method will yield consistent estimation and the estimation efficiency can be improved

through the GMM over the analysis ignoring the spatio-temporal inter-correlation. We conducted simulations of the proposed methods with various covariance structures to illustrate ideas.

email: yunbai@umich.edu

## **9f. SPATIO-TEMPORAL MODELLING FOR LUPUS INCIDENCE IN TORONTO, CANADA SINCE 1965**

Ye Li\*, University of Toronto  
Patrick E. Brown, Cancer Care Ontario

The data provided by the Toronto Western Hospital Lupus clinic provide a fairly complete capture of residence locations of lupus cases in Toronto since 1965. These data span several census periods, with each census using different geographical boundaries for census tracts. A fine grid (175m\*175m) was created over entire Greater Toronto Area, expected count were calculated for each grid cell by obtaining the expected count for each Age, Sex and Year group of the census data. Using that as an offset, a Gaussian Markov Random Field is fit on the grids assuming risk surface does not change over time. The posterior sample of risk surfaces is used to construct an estimated risk surface and detect areas likely to be 'clusters' of excess risk. Reporting bias is examined by including distance from the clinic as a covariate, allowing for the clinic changing locations in the 1980's. Future work planned involves allowing the risk surface to change over time, and using Markov Random Fields to approximate continuous spatial surfaces. The results of this research will be used to identify environmental risk factors for the disease.

email: ye.li@utoronto.ca

## **9g. SPATIAL MODELING OF AIR POLLUTION EXPOSURE, MEASUREMENT ERROR AND ADVERSE BIRTH OUTCOMES**

Simone Gray\*, Duke University  
Alan Gelfand, Duke University  
Marie Lynn Miranda, Nicholas School of the Environment  
Sharon Edwards, Nicholas School of the Environment

When estimating personal exposure it is customary to use measurements from the closest monitoring station as a simple proxy for personal exposure. Evidently, measurement error is introduced as the estimated exposure is not equal to the monitored exposure. This talk presents work that attempts to better understand the relationship between maternal exposure to air pollution and pregnancy outcomes, while accounting for the associated measurement error. Induced by a process model specification for exposure reflecting sparsity of monitoring sites, we construct a spatial model that allows uncertainty in exposure to increase as the distance between maternal residence and the location of the closest monitor increases. We illustrate with air quality data from the EPA. We extend this method to account for missing data values in the monitors that are inactive by design and measured every three or every six days. We assume that error increases as the time from the nearest recorded measurement increases and incorporate that error term into a temporal component of the model. The statistical analyses are implemented using hierarchical modeling within a Bayesian perspective.

email: simone@stat.duke.edu

## **9h. STATISTICAL ANALYSIS OF THE EFFECTS OF AIR POLLUTION ON CHILDREN'S HEALTH**

Elizabeth A. Stanwyck\*, University of Maryland Baltimore County  
Bimal Sinha, University of Maryland Baltimore County

Effects of air pollution on human health, especially on the health of children, have been a concern for many years and serious attempts have been made to effectively measure such effects. While one end of the equation, namely human health effects, can be either observed or measured without much error, the other end of the equation, specific pollutants and their amounts at the (micro) individual level, is very hard to determine. Fortunately, some interesting and useful statistical models can be used to describe such uncertainties, and more importantly, to meaningfully link the two sides. In this presentation a general overview of the problem and some solutions from both frequentist and Bayesian points of view will be presented. Then an adaptation of these procedures to deal with real data available in the United States will be discussed. Also, separate results of a study involving a few Chinese cities based on an entirely different approach will be given and a comparison of the two approaches using a simulation study will be mentioned.

email: estanwy1@math.umbc.edu

## **9i. THE EFFECT OF RAINFALL ON VISITS TO PEDIATRIC EMERGENCY ROOMS FOR DIARRHEA**

Shun H. Li\*, Medical College of Wisconsin  
Pippa M. Simpson, Medical College of Wisconsin  
Stephen Stanhope, Medical College of Wisconsin  
Ke Yan, Medical College of Wisconsin  
Marc Gorelick, Medical College of Wisconsin  
Bevan E. Huang, CSIRO  
Raymond G. Hoffmann, Medical College of Wisconsin

Modeling the effect of one time series on another where there are innovations and both cyclical and acyclical trends requires an extensive modeling building process to examine the temporal relationships. In a study of the association between gastroenteritis and rainfall in children residing in the Lake Michigan watershed, surface water and well water are regularly contaminated with chemicals and other daily wastes. Rainwater carries additional wastes from the surface into the drinking water. Ancillary measurements include checking for cryptosporidium, giardia and coliform bacteria, but not viral particles. Cracks in the pipelines increase the chance of contamination due to infiltration during rainfall. An example of an innovation was the occurrence of an intense storm where the runoff exceeded the capacity of the normal water filtration plants. Autoregressive integrated moving average (ARIMA) models with innovations were tested and evaluated for their validity, accuracy and reliability. Since the incubation time could be variable, models were considered with and without seasonal considerations and with and without lag differences. The model building procedure involved iterative cycles consisting of three stages: (1) building the ARIMA model, (2) model estimation, and (3) diagnostic checking.

email: psimpson@mcw.edu

# ABSTRACTS

## 9j. STOCHASTIC MODELS OF FLOW THROUGH A RANDOM GRAPH

Nicholas M. Murray\*, Texas Tech University  
Clyde Martin, Texas Tech University  
Dorothy Wallace, Dartmouth College

Playa Lakes can be found throughout the Panhandle of Texas and are formed by seasonal rains. Many species populate the Playa Lakes. In each species, it is possible to track genetic adaptations which can be extremely localized, for example, present in a single Playa Lake. It is then possible to track the passing of a specific genetic trait throughout the Playa Lake system. The goal of this research is to acquire an estimate of the time required for a genetic trait to travel between the two most distant Playa Lakes. This is accomplished by constructing a simple stochastic model simulating the movement of populations carrying the genetic trait. The model must take into account that traits can only be passed between nearest neighbors and playa lakes can dry out or be filled on a seemingly random basis. The problem reduces to a stochastic flow on a graph that is changing randomly. For small sets of nodes (playas) the problem has the potential for an analytic solution. However, for systems of the same order of magnitude as the playa system (20,000 nodes) experimental results are the best possible. From this model, a time estimate can be constructed for any system size.

email nicholas.m.murray@ttu.edu

## 9k. BAYESIAN ALIGNMENT OF CONTINUOUS MOLECULAR SHAPES

Irina Czogiel\*, University of Nottingham, UK  
Ian L. Dryden, University of Nottingham, UK  
Christopher J. Brignell, University of Nottingham, UK

A frequent objective in drug design is to find molecules with a high binding affinity towards a certain target protein. If no structural information about the target protein is available, putative ligands are often superimposed with the structure of a reference ligand which is known to bind to the target under consideration. If a ligand can be aligned closely, it is likely to exhibit a similar biochemical activity and hence drug potency. Here, we propose a statistical model for evaluating and comparing molecular shapes using methods from the field of statistical shape analysis. In order to account for the continuous nature of molecules, we combine these methods with techniques used in spatial statistics and apply kriging to predict the values of the considered molecular properties (e.g. partial atomic charge) in three-dimensional space. Superimposing entire fields rather than discrete points solves the problem that there is usually no clear one-to-one correspondence between the atoms of the considered molecules. Using similar concepts, we also propose an algorithm for the simultaneous alignment of molecular fields. Our methods work well on a data set comprising 31 steroid molecules which has been used as a test bed for various alignment techniques.

email: pmxic@nottingham.ac.uk

## 9l. MODELLING SPATIO-TEMPORAL TRENDS OF FOREST HEALTH MONITORING DATA

Nicole H. Augustin\*, University of Bath Mathematical Sciences, Bath, UK  
Monica Musio, University of Cagliari, Italy  
Klaus von Wilpert, Forest Research Centre Baden-Wurttemberg, Freiburg, Germany  
Edgar Kublin, Forest Research Centre Baden-Wurttemberg, Freiburg, Germany  
Simon N. Wood, University of Bath Mathematical Sciences, Bath, UK  
Martin Schumacher, University Hospital Freiburg University, Freiburg, Germany

Forest health monitoring surveys are in operation in Europe since the early 1980s due to forest eco-system damage by air pollution. Here we model yearly data on spruce tree defoliation from a monitoring survey carried out in Baden-Wurttemberg, Germany since 1983. On an irregular grid defoliation and other site specific variables are recorded. The temporal trend of defoliation differs between areas because of site characteristics and pollution levels, making it necessary to allow for space-time interaction. We use a generalized additive mixed model combined with scale invariant tensor product smooths of the space-time dimension. In addition to a temporal trend due to site characteristics and other conditions modeled with the space-time smooth, we account for random temporal correlation at site level using an auto-regressive moving average process. The results show that since 2003 there is significant evidence for an increased trend in defoliation in spruce. The defoliation can mainly be associated with recent drought years due to climate change and cumulative effects of pollution.

email: n.h.augustin@bath.ac.uk

## 10. POSTERS: CATEGORICAL DATA ANALYSIS AND SURVEY RESEARCH

### 10a. SYNTHESIZING CATEGORICAL DATASETS TO ENHANCE INFERENCE

Veronica J. Berrocal\*, Duke University  
Alan E. Gelfand, Duke University  
Sourab Bhattacharya, Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute  
Marie L. Miranda, Duke University, Nicholas School of the Environment  
Geeta Swamy, Duke University

A common data analysis setting consists of a collection of datasets of varying sizes that are all relevant to a particular scientific question, but which include different subsets of the relevant variables, presumably with some overlap. Here, we present a method to synthesize incomplete categorical datasets drawn from an incompletely cross-tabulated population, for which the marginal probabilities are known and where at least one of the dataset is completely observed. We show that our method can still be applied even when the complete dataset refers to a subset of the population not obtained via simple random sampling. We also demonstrate that in the incomplete categorical datasets scenario synthesizing datasets is expected to reduce uncertainty about the cell probabilities in the associated multi-way

contingency table as well as for derived quantities such as relative risks and odds ratios, but the improvement is not guaranteed. This differs from the complete datasets scenario, where we are assured to tighten inference. To illustrate our method, we present simulated examples motivated by an adverse birth outcomes investigation.

email: vjb2@stat.duke.edu

## **10b. USE OF SECONDARY DATA ANALYSIS AND INSTANTANEOUS STATES IN A DISCRETE-STATE MODEL OF DIABETIC HEART DISEASE**

Jacob Barhak, University of Michigan  
Deanna JM Isaman, University of Michigan  
Wen Ye\*, University of Michigan

With the increasing burden of chronic diseases on the health care system, Markov-type models are becoming popular to guide disease management. Typically, these models use secondary data analysis to estimate transitions in the model; however there are frequent discrepancies between the theoretical model and the design of the studies being used. This paper demonstrates a likelihood approach to correctly model the design of clinical studies when 1) the theoretical model may include an instantaneous state of distinct interest, and 2) the secondary data can not be used to estimate a single parameter in the theoretical model (e.g., a study may ignore intermediary stages of disease). Our approach not only accommodates the two conditions above, but allows using more than one study to estimate model parameters and allows model refinement. We name our approach the Lemonade Method in the spirit of “when study data give you lemons, make lemonade.” This method is applied to a model of heart disease in diabetes.

email: wye@umich.edu

## **10c. SAMPLING TABLES GIVEN A SET OF CONDITIONALS**

Juyoun Lee\*, Penn State University  
Aleksandra Slavkovic, Penn State University

Federal agencies and other organizations publish a data summarized in arrays of non-negative integers, called a contingency table. When releasing the data, it is necessary to prevent the sensitive information of the individuals from being disclosed. In statistical disclosure limitation, we must maintain a balance between disclosure risk and data utility for the purposes of statistical inference. One method of achieving this balance is to release partial information about the original data; in practice, many federal agencies and medical institutions release data summarized in the form of marginal sums, conditional probabilities, or odds-ratios. Sampling methods for multi-way contingency tables given a set of marginal sums have been studied in diverse ways while there is almost no literature about sampling of tables given a set of conditional probabilities. Here, we focus on a set of conditional probabilities instead of a set of marginal sums. We describe the Markov chain Monte Carlo (MCMC) algorithm based on the algebraic tools for sampling contingency tables with given conditional probabilities. This algorithm can be used for Bayesian computation of posterior distribution and assessment of data utility

and disclosure in statistical disclosure limitation. We demonstrate the MCMC algorithm with examples and discuss its advantages and disadvantages. We then discuss their feasibility for sampling tables given a set of conditional probabilities.

email: jxl982@psu.edu

## **10d. EXAMINING THE ROBUSTNESS OF FULLY SYNTHETIC DATA TECHNIQUES FOR DATA WITH BINARY VARIABLES**

Gregory Matthews\*, University of Connecticut

There is a growing demand for public use data while at the same time increasing concerns about the privacy of personal information. One proposed method for accomplishing both goals is to release data sets which do not contain real values but yield the same inferences as the actual data. The idea, first proposed by Rubin (1993), is to view confidential data as missing and use multiple imputation techniques to create synthetic data sets. In this paper, we compare techniques for creating synthetic data sets in simple scenarios with a binary variable.

email: gjm112@gmail.com

## **10e. BAYESIAN MODEL-BASED ESTIMATES OF DIABETES INCIDENCE BY STATE**

Theodore J. Thompson, Centers for Disease Control and Prevention  
Betsy L. Cadwell, Centers for Disease Control and Prevention  
James P. Boyle, Centers for Disease Control and Prevention  
Lawrence Barker, Centers for Disease Control and Prevention

The Behavioral Risk Factor Surveillance System (BRFSS) is a state based random-digit-dialed survey of the U.S. civilian, noninstitutionalized population aged > 18 years. The fraction of the population diagnosed with diabetes within the past year (incidence) can be determined from the diabetes module of the BRFSS, not administered by all states in all years. Direct design-based estimates of diabetes incidence for a single state in a single year are not sufficiently precise to be useful. We develop a multilevel model that provides the first practical yearly estimates of diabetes incidence for all 50 states and the District of Columbia. Using area level models that treat design-based estimates and their estimated variances as data, we provide estimates for the years 1999 to 2006. State level covariates (Census division, diabetes prevalence, percent completing high school) are included. State-level temporal random effects are modeled as first order autoregressive processes. Unlike some competing methods, Bayesian models constrain all estimates to the interval (0, 1). Posterior distributions of diabetes incidence by state and year and posterior distributions of state ranks by year are provided.

email: tat5@cdc.gov

# ABSTRACTS

## 10f. HEALTH DISPARITY INDICES - SIMULATIONS OF UNDERLYING DEPENDENCIES

Stuart A. Gansky\*, University of California, San Francisco, Center to Address Disparities in Children's Oral Health

Nancy F. Cheng, University of California, San Francisco, Center to Address Disparities in Children's Oral Health

The US National Center for Health Statistics (NCHS) recent monograph for reporting health disparities (HDs) (Keppel et al. 2005) includes commonly used health disparities indices (HDIs). Some authors claim all HDIs depend on the underlying prevalence (e.g. Scanlan 2006). With a simple simulation we recently illustrated that some HDIs depend on underlying prevalence but others do not (Cheng et al. 2008). This investigation studied how HDIs are affected by underlying factors (e.g. prevalence, sample size, standard error, and strength of association). HDIs included absolute and relative measures, Slope Index of Inequality (SII), Relative Index of Inequality (RII) for mean and ratio, health concentration index (C), and entropy (Theil's index). HDIs were estimated for the California Oral Health Needs Assessment of Children 2004-5, a complex stratified cluster sample survey ( $N=21,399$ ), to assess associations of race/ethnicity and socioeconomic position (SEP) (%free/reduced-price lunch (FRL) program in schools) to oral health outcomes such as rampant caries. Simulations varied underlying prevalence while keeping other factors constant (e.g. sample size, strength of association with SEP covariate). Then other factors were varied in simulations. We studied the behavior of HDIs based on underlying properties. Support: US DHHS NIH/NIDCR R03DE018116.

email: stuart.gansky@ucsf.edu

## 10g. NEW DEVELOPMENT OF OPTIMAL COEFFICIENTS FOR A BEST LINEAR UNBIASED ESTIMATOR OF THE TOTAL FOR SIMPLE RANDOM SAMPLING WITH REPLACEMENT USING GODAMBE'S GENERAL LINEAR ESTIMATOR

Shuli Yu\*, School of Public Health, University of Massachusetts-Amherst

Edward J. Stanek III, School of Public Health, University of Massachusetts-Amherst

Questions about real populations/subjects are appropriately framed in models that closely match the setting, rather than in artificial superpopulation or infinite population frameworks. Finite population mixed models (FPMM) are appropriately framed, but are limited to simple random sampling without replacement settings (one or two stage). Godambe (1955) defined and proved that there is no best linear unbiased estimator (BLUE) for a sample from a finite population in a general setting that allowed for with replacement sampling. We discuss Godambe's results in the context of the FPMM, obtaining optimal coefficients for the BLUE of the population total based on a model for with replacement sample sets of size  $n=2$  from a finite population of size  $N=3$ . Optimal coefficients are obtained when parameters are distinct and not equal to zero after specifying unbiased constraints and one additional constraint for over-parameterization. The solutions are consistent with the results of our separate study regarding a simple random sampling without replacement of size  $n=2$

from  $N=3$ . Substituting the optimal coefficients into the estimator gives a solution with zero mean squared error (MSE).

email: shuli@schoolph.umass.edu

## 11. POSTERS: VARIABLE/MODEL SELECTION

### 11a. NONPARAMETRIC BAYES CONDITIONAL DISTRIBUTION MODELING WITH VARIABLE SELECTION

Yeonseung Chung\*, Harvard School of Public Health  
David Dunson, Duke University

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 first introduce the probit stick-breaking process (PSBP) as a prior for an uncountable collection of predictor-dependent random probability measures, and propose a PSBP mixture (PSBPM) of normal regressions for modeling the conditional distributions. A global variable selection structure is incorporated to discard unimportant predictors, while allowing estimation of posterior inclusion probabilities. Local variable selection is conducted relying on the conditional distribution estimates at different predictor points. An efficient stochastic search algorithm is proposed for posterior computation. The methods are illustrated through simulation and applied to an epidemiologic study.

email: chungy@email.unc.edu

### 11b. A NEW APPROACH TO HIGH DIMENSIONAL VARIABLE SELECTION

Xingye Qiao\*, University of North Carolina-Chapel Hill  
Yufeng Liu, University of North Carolina-Chapel Hill  
J.S. Marron, University of North Carolina-Chapel Hill

Popular methods of variable selection, such as SAM, focus on each variable individually. However, some heuristic data examples show that improvement is available by testing on multiple variables simultaneously. We propose an innovative approach to variable selection based on testing significance of variables in a moving window with varying centers and sizes, i.e. a multi-scale approach. Re-ordering of the variables according to the summarized score of each variable is taken to improve the ranking of the variables. A new visualization gives a convenient summary of the test results.

email: xyqiao@email.unc.edu



## **11c. TESTING FOR CONDITIONAL INDEPENDENCE VIA MULTIPLE MODELS: AN ORTHOPEDIC APPLICATION FOR THE SIX SEGMENT FOOT MODEL**

Sergey Tarima\*, Medical College of Wisconsin  
Xue-Cheng Liu, Medical College of Wisconsin  
Roger Lyon, Medical College of Wisconsin  
John Thometz, Medical College of Wisconsin  
Channing Tassone, Medical College of Wisconsin

A hypothesis about conditional independence between two random variables X and Y given Z is often tested via an assumed regression model  $E(g(Y|X,Z))$ , where  $g(y)$  is monotone. After fitting a chosen model the Wald test is regularly used for testing significance of the regression coefficient at X (we assume no interaction with Z). However, this model is not the only model applicable for testing conditional independence. The other regression models, such as  $E(h(Y|X,Z))$  or  $E(f(X)|Y,Z)$  can be used as well,  $h(y)$  and  $f(x)$  are monotone functions. We consider a simple method for testing conditional independence based on multiple models. Our method secures asymptotically the type I error. We illustrate application of this approach on orthopedic patients for testing conditional independence between kinematic parameters measured by the six-segment foot model. The six-segment foot model is designed for 3D measurements of the foot and ankle joint in the patient with spastic cerebral palsy (CP), which allows assessing the degree of foot deformities objectively. A total of 5 typically developing children and 8 children with CP, aged from 6 to 18 year-old, were enrolled in the study.

email: starima@mcw.edu

## **11d. MODEL CHECKING FOR BAYESIAN ESTIMATION OF STATE DIABETES INCIDENCE RATES**

James P. Boyle\*, Centers for Disease Control  
Betsy L. Cadwell, Centers for Disease Control  
Theodore J. Thompson, Centers for Disease Control  
Lawrence Barker, Centers for Disease Control

Direct design-based survey estimates of annual state incidence rates of diagnosed diabetes and their variances were obtained from the Behavioral Risk Factor Surveillance System (BRFSS). These estimates were for the eight years 1999 through 2006, for three age groups (20-44, 45-64, and 65+ years) and 51 states (including Washington D.C.). Of the  $8(51)(3) = 1224$  incidence rates, only 1011 were available because of missing survey data. Furthermore, many of the 1011 direct estimates were imprecise with only 87 estimates having a coefficient of variation < 20%. To improve precision, a model-based approach applying Bayesian models to direct design-based estimates and their estimated variances as data was used. Several multilevel models were fit, ranging from very simple models with no covariates and independent errors to more complicated models with first order autoregressive time series errors and state level covariates from the U.S. Census Bureau. Models were ranked with the deviance information criterion (DIC). We checked the fit of the model with minimum DIC through simulated values from the posterior predictive distribution of replicated data (posterior predictive checking), and found the posterior predictive p-values acceptable.

email: jboyle@cdc.gov

## **11e. CHECKING TRANSFORMATION MODELS WITH CENSORED DATA**

Li Chen\*, University of North Carolina at Chapel Hill

Transformation models provide a very general framework for studying the effects of (possibly time-dependent) covariates on survival time and recurrent event times. Assessing the adequacy of these models is an important task because model misspecification affects the validity of inference and the accuracy of prediction. In this article, we introduce appropriate residuals for these models and consider the cumulative sums of the residuals. The cumulative-sum processes under the assumed model are shown to converge weakly to zero-Gaussian processes whose distributions can be approximated through Monte Carlo simulation. These results enable one to assess, both visually and numerically, how unusual the observed residual patterns are in reference to their null distributions. The residual patterns can also be used to determine the nature of model misspecification. Extensive simulation studies demonstrate that the proposed methods perform well in practical situations. A colon cancer study is provided for illustration.

email: lchen@bios.unc.edu

## **11f. CLASSIFICATION OF FUNCTIONAL DATA: A SEGMENTATION APPROACH**

Bin Li\*, Louisiana State University  
Qingzhao Yu, Louisiana State University

We suggest a method for combining the classical method of linear discriminant analysis and support vector machine to functional data where the predictor variables are curves or functions. This procedure, which we call Segment Discriminant Analysis (SDA), is particularly useful for irregular functional data, characterized by spatial heterogeneity and local patterns like spikes. In addition, SDA allows us to (1) reduce the computation and storage burden by using a fraction of curves; (2) select important markers and extract features automatically; (3) incorporate prior knowledge from the investigators. We apply SDA to two public domain data sets and discuss the understanding developed from the study.

email: bli@lsu.edu

## **11g. NETWORK EXPLORATION VIA THE ADAPTIVE LASSO AND SCAD PENALTIES**

Jianqing Fan, Princeton University  
Yang Feng\*, Princeton University  
Yichao Wu, North Carolina State University

Graphical models are frequently used to explore networks, such as genetic networks, among a set of variables. This is usually carried out via exploring the sparsity of the precision matrix of the variables under consideration. Penalized likelihood methods are often used in such explorations. Yet, positive-definiteness constraints of precision matrices make the optimization problem challenging. We introduce non-concave penalties and the adaptive LASSO penalty to attenuate the bias problem in the network estimation. Through the local linear

# ABSTRACTS

approximation to the non-concave penalty functions, the problem of precision matrix estimation is recast as a sequence of penalized likelihood problems with a weighted L1 penalty and solved using the efficient algorithm of Friedman et al. (2008). Our estimation schemes are applied to two real datasets. Simulation experiments and asymptotic theory are used to justify our proposed methods.

email: yangfeng@princeton.edu

## 12. POSTERS: DIAGNOSTIC TESTS

### 12a. DETERMINING PRESENCE OF GB VIRUS TYPE C IN HIV POSITIVE SUBJECTS

Carmen J. Smith\*, University of Iowa  
Kathryn Chaloner, University of Iowa

GB virus type C (GBV-C) is a virus that appears to interact with HIV. According to several studies, HIV positive individuals infected with GBV-C live longer than HIV positive individuals without GBV-C. However, there is no gold standard for detecting the presence of GBV-C in blood samples. One commercial test (Roche) and three locally developed ELISA tests (M5, M6, and GNA) were run on 100 stored blood samples in the UI Stapleton laboratory. The objective is to investigate relationships between the tests and develop an algorithm for classifying samples as either positive or negative for GBV-C. The results are explored graphically, and maximum likelihood is used to fit mixtures of normal distributions to the results of each test. Model selection criteria and graphical inspection are used to examine the results. None of the tests clearly show two distinct populations. Further research is needed before these tests can be used routinely.

email: carmen-j-smith@uiowa.edu

### 12b. MODELING SENSITIVITY AND SPECIFICITY WITH A TIME-VARYING REFERENCE STANDARD WITHIN A LONGITUDINAL SETTING

Qin Yu\*, University of Rochester  
Wan Tang, University of Rochester  
Sue Marcus, Department of Psychiatry, Mount Sinai School of Medicine  
Yan Ma, University of Rochester  
Xin M. Tu, University of Rochester

Diagnostic tests are used in a wide range of behavioral, medical, psychosocial, and health-care related research. Test sensitivity and specificity are the most popular measures of accuracy for diagnostic tests. Available methods for longitudinal study designs assume fixed gold or reference standards and as such do not apply to studies with dynamically changing reference standards, which are especially popular in psychosocial research. In this paper, we develop a novel approach to address missing data and other related issues for modeling sensitivity and specificity within such a time-varying reference standard setting. The approach is illustrated with real data in sexual health research.

email: qin\_yu@urmc.rochester.edu

### 12c. COMPARISON OF CORRELATED CORRELATION COEFFICIENTS USING BOOTSTRAPPING

Juhee Song\*, Scott & White Hospital  
Jeffrey D. Hart, Texas A&M University

Many clinical researches have been studied comparison of a gold standard measure and other surrogate measures using diagnostic test to find the best surrogate measure. In many cases a cutoff point for diagnostic test is not well established and depends on study population. Instead of applying diagnostic test, correlations between a gold standard measure and other surrogate measures were compared to find the best surrogate measure. A method of comparing correlated Pearson's correlation coefficients using the Fisher's transformation proposed by Meng et al was adapted and modified. A test that the highest correlation coefficient is significantly better than one of other correlation coefficients was studied. A critical value from Gaussian distribution may not be appropriate, since the distribution of a correlation coefficient, which is not the highest, subtracts from the highest correlation coefficient is not normally distributed. A linear model that considered a standardized gold standard as response variable, and standardized other surrogate measures as explanatory variables was motivated and all parameters were estimated with least square estimation. Bootstrap samples from surrogate measures and residuals were taken to construct the sampling distribution of a test statistic. Simulation study was done.

email: jhsong@gmail.com

### 12d. METHODS FOR CALIBRATING BIVARIATE LABORATORY DATA

Ke Yan\*, Medical College of Wisconsin  
Raymond G. Hoffmann, Medical College of Wisconsin  
Shi-Hwan Li, Medical College of Wisconsin  
Robert Montgomery, Medical College of Wisconsin  
Pippa Simpson, Medical College of Wisconsin

A laboratory calibrates a response curve for an assay to a standard by using known quantities ( $X$ ). The observed assay values ( $Y$ ) are used with each standard  $X$  value to obtain a least squares fit. Provided the curve is monotonic, the least squares fit can be inverted to form a calibration curve that gives an adjusted value ( $X$ ) for any new observation ( $Y$ ). Thus the calibration problem is often described as an inverse regression problem. Von Willebrand's Disease (VWD) is a clotting disorder where there are two related measurements which characterize the disease. Moreover, the ratio of the two is also quite important diagnostically. Both of these measurements need to be calibrated, to standards and preferably not separately. Working with this data has motivated us to develop and compare methods for the bivariate calibration problem. Although the univariate calibration problem has been extensively studied, results for the bivariate calibration problem appear to be almost non-existent in the statistical literature. The results depend on the level of correlation between the two measurements, as well as how close the ratio is to one.

email: hoffmann@mcw.edu

## **13. POSTERS: NONPARAMETRIC METHODS**

### **13a. ESTIMATING THE VARIANCE OF BJE UNDER DISCRETE ASSUMPTION**

Yishi Wang\*, University of North Carolina-Wilmington  
Cuixian Chen, University of North Carolina-Wilmington

The Buckley-James estimator is a widely recognized approach in dealing right censored linear regression models. There have been a lot of discussions in the literatures on the estimation of the Buckley-James estimator as well as its asymptotic distribution. But so far, no simulation has been done to estimate the standard error of the estimator. Kong and Yu (2007) studied the asymptotic distribution under discrete assumption. Based on their methodology, we recalculate the asymptotic variance and formulate the estimation of variance by using plug-in estimators. The simulation suggests that the estimation is a great approximation compared with the empirical SD. The large sample property of the estimator is discussed.

email: wangy@uncw.edu

### **13b. THE GMLE BASED BUCKLEY-JAMES ESTIMATOR WITH MODIFIED CASE-COHORT DATA**

Cuixian Chen\*, University of North Carolina-Wilmington

We consider the estimation problem under the linear regression model with the modified case-cohort design. The extensions of the Buckley-James estimator (BJE) under the case-cohort designs have been studied under an additional assumption that the censoring variable and the covariate are independent. If this assumption is violated, as is the case in a typical real data set in the literature, our simulation results suggest that those extensions are not consistent and we propose a new extension. Our estimator is based on the generalized maximum likelihood estimator (GMLE) of the underlying distributions. We propose a self-consistent algorithm, which is quite different from the one for ultivariate interval-censored data.

email: chenc@uncw.edu

### **13c. BOOTSTRAP CONFIDENCE INTERVALS FOR THE PREDICTORS OF TREATMENT MEANS IN A ONE FACTOR EXPERIMENTAL DESIGN WITH A FINITE POPULATION**

Bo Xu\*, School of Public Health, University of Massachusetts-Amherst  
Edward Stanek, School of Public Health, University of Massachusetts-Amherst

A one factor experimental design is developed based on the potential observable outcome framework, sampling from finite population of units and random allocation of treatments. The predictors for treatment means are obtained by Royall's (1976) prediction theory where the treatment deviations are represented as realized random effects. When the variance components are unknown, the empirical

predictors are considered. The confidence intervals for the empirical predictors are calculated using bootstrapping methods. Several bootstrap methods, such as bootstrapping with replacement (BWR) or bootstrapping without replacement (BWO) are introduced. Each bootstrap method is developed to account for the sampling from the finite population of units and random allocation of treatments. Comparisons of the different bootstrapping methods are made methodologically, and via simulation. We discuss these comparisons, with a goal of recommending an appropriate bootstrapping method for statistical inference in the one factor experimental design.

email: bxu@schoolph.umass.edu

### **13d. GENERALIZED ANOVA FOR CURRENTLY MODELING MEAN AND VARIANCE WITHIN A LONGITUDINAL DATA SETTING**

Hui Zhang\*, University of Rochester Medical Center  
Xin Tu, University of Rochester Medical Center

Although widely used for comparing multiple samples in biomedical and psychosocial research, the analysis of variance (ANOVA) model suffers from a series of flaws that not only raise questions about conclusions drawn from its use, but also undercut its many potential applications to modern clinical and observational research. In this paper, we propose a new class of generalized ANOVA models to concurrently address all these fundamental flaws underlying this popular multi-group comparison approach so that it can be applied to many immediate as well as potential applications ranging from addressing an age-old technical issue in applying ANOVA to cutting-edge methodological challenges arising from the emerging effectiveness research paradigm. By integrating the classic theory of U-statistics with the state-of-the-art concepts such as the inverse probability weighted estimates, we develop distribution-free inference for this new class of models to address missing data for longitudinal clinical trials and cohort studies. We illustrate the proposed class of models with both real and simulated study data, with the latter investigating behaviors of model estimates under small and moderate sample sizes.

email: hui\_zhang@urmc.rochester.edu

### **13e. PRACTICAL ESTIMATION AND DISCUSSION OF NEURONAL PHASE-RESPONSE (PHASE-RESETTING) CURVES**

Daniel G. Polhamus\*, University of Texas, San Antonio  
Charles J. Wilson, University of Texas, San Antonio  
Carlos A. Paladini, University of Texas, San Antonio

Phase-resetting curves provide valuable insight to neural interconnectivity and, as both an oscillatory and neuronal stimulation model, are well documented. Estimation of the PRC from real-life experimental neurons is complicated by the inherent variability of biological electrical systems and human interaction with these signals. In the process of discussing current empirical methodology for the estimation of phase-resetting curves in the context of simulated and real data, we comment on aberrations from the expected behavior. Much of this variability can be traced back to distributional characteristics of the baseline inter-spike intervals (ISI). We propose

# ABSTRACTS

corrections to current empirical methodology and demonstrate the robust nature of our proposed PRC estimation, relative to the aforementioned aberrations and data.

email: daniel.polhamus@utsa.edu

## 14. POSTERS: STATISTICAL MODELS AND METHODS

### 14a. FURTHER DEVELOPMENT OF SEMI-PARAMETRIC METHODS IN BAYESIAN BETA REGRESSION

Christopher J. Swearingen\*, Medical University of South Carolina  
Dipankar Bandyopadhyay, Medical University of South Carolina

Beta Regression is a generalized linear model that estimates both location and precision parameters of a dependent variable that assumes a Beta distribution. This model is extremely flexible, allowing independent covariate prediction of change in location, such as response between treatment groups, as well as changes in precision, such as heteroskedasticity of response between groups. To date, only one Bayesian approach to Beta Regression has been published [1] and illustrated the usage of penalized splines to model a non-linear location covariate. However, this work assumed constant precision in its modeling and did not consider how a non-linear covariate may impact precision. Our examination of penalized spline models examining both constant and non-constant precision assumptions is presented, based in part upon a previous analysis [2] of 285 ischemic stroke lesion volumes which detected a non-linear association between a covariate and median infarct volume.

- [1] Branscum, A., Johnson, W. & Thurmond, M. Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics* 49, 287-301 (2007).  
[2] Nicholas, J. Swearingen, C.J. et al. The effect of statin pre-treatment on infarct volume in ischemic stroke. *Neuroepidemiology*. 31, 48-56 (2008).

email: christopher.swearingen@alumni.musc.edu

### 14b. PENALIZED MAXIMUM LIKELIHOOD ESTIMATION IN LOGISTIC DOSE RESPONSE MODELS

Amy E. Wagler\*, University of Texas at El Paso

Logistic regression is a widely utilized method of modeling quantal dose response data. These models typically employ maximum likelihood (ML) methods for estimating the model parameters. Maximum likelihood estimators (MLEs) are well-known to be biased at smaller sample sizes. This bias can lead to very misleading results when the number of doses or the number of replications per dose is small. For example, past simulations indicate that the empirical coverage level for quantities such as the effective dose (ED) can be as low as 18.8% for models utilizing ML estimation with low numbers of dose levels. Additionally, when there are small numbers of doses or few replications per dose, the ML estimators often fail to converge. An

alternative estimator, the Penalized Maximum Likelihood Estimator (pMLE), is considered in order to address the bias present in the MLE and the problem of non-convergence. Simulations confirm that less bias is present when utilizing the pMLE in place of the MLE in quantal dose-response models with and without a covariate. Additionally, simulations compare the empirical coverage levels of intervals for response probabilities and delta and Fieller intervals for the effective dose (ED).

email: awagler2@utep.edu

### 14c. ACHIEVING COVARIANCE ROBUSTNESS IN THE LINEAR MIXED MODEL

Matthew J. Gurka, University of Virginia,  
Keith E. Muller\*, University of Florida

The flexibility of the general linear mixed model in describing how repeated measurements on individuals covary has helped create a myth of robustness to mis-specification of the covariance model, especially in large samples. Previous work has revealed biased inference about the fixed effects, particularly for small samples, when the covariance structure is misspecified. In contrast to previous work, we prove analytically that incorrectly assuming homogeneity within subjects (i.e., compound symmetry) leads to biased inference about the fixed effects, in both small and large samples. An unstructured covariance assumption has appeal when interest lies in the fixed effects. However, convergence can be a problem in small samples. We rely on a combination of theoretical and numerical results to examine the impact of various covariance model strategies on accuracy of inference about the fixed effects. Strategies include varying the approximation for the test statistic, such as the Kenward-Roger approach, and using the sandwich estimator of the fixed effects covariance matrix. However, we avoid model selection techniques. The results lead to practical guidance in achieving accurate inference about fixed effects inference with unknown covariance structure.

email: Keith.Muller@biostat.ufl.edu

### 14d. BAYESIAN MIXTURES FOR MODELING THE CORRELATION OF LONGITUDINAL DATA

Lei Qian\*, University of California at Los Angeles  
Robert Weiss, University of California at Los Angeles

In longitudinal data analysis, parametric covariance models rely on strong assumptions, while the unstructured covariance model has too many parameters and can not be fit to high dimensional unbalanced data. We propose two rich families of Bayesian semi-parametric stationary correlation mixture models. One approach is a convex combination of simple structure correlation matrices, the second approach models correlations as a convex monotone B-spline function. We compare our models to each other and to standard models using DIC, cross-validation and log marginal likelihood. Simulations show that our model works well and DIC, cross-validation and log marginal likelihood choose similar models. We illustrate our methods with an unbalanced dataset of CD4 cell counts.

email: leiqian@ucla.edu

## **14e. ESTIMATION OF PROBABILITY DISTRIBUTIONS USING CONTROL THEORETIC SPLINES**

Janelle K. Charles\*, Texas Tech University  
Clyde F. Martin, Texas Tech University

We examine the relationship between optimal control and statistics. We explore the use of control theoretic smoothing splines in the estimation of continuous probability distribution functions defined on a finite interval  $[0, T]$ , where the data is summarized by empirical probability distributions. In particular, we consider the estimation of distributions of the form  $\exp(f(t))$ , where there is no restriction on the sign of  $f(t)$ . The construction of the optimal smoothed curve,  $y(t)$ , is based on the minimization of an integral cost function done through the application of the Hilbert Projection Theorem, which guarantees that a unique minimum exists.

email: janelle.k.charles@ttu.edu

## **14f. APPLICATION OF THE KALMAN FILTER ALGORITHM TO ESTIMATE A FUNCTIONAL MIXED MODEL**

Meihua Wu\*, University of Michigan  
Brisa N. Sánchez, University of Michigan

Applications of functional data analysis (FDA) methods in epidemiology remain limited in large population-level studies. One such FDA method is functional mixed models (FMMs). In functional mixed models, the dependent variable for each individual in the study is a function, which is sampled at a finite set of time points. The average across individuals is a function, which may depend on covariates. Covariate effects are also represented by functions. In small studies, FMMs can be easily estimated using standard mixed model software. However, when using this software, the computation time grows proportional to the cube of the number of observations. Thus, in large studies (in terms of the number of individuals) mixed model software fails. The Kalman filter algorithm has been proposed as a computationally efficient approach to estimate FMMs, and has been used effectively in moderate size datasets where the set of sample times is the same for all individuals. We employ the Kalman filter to estimate a FMM for a large dataset where the sample times differ across individuals. We discuss some challenges of this application, and propose some solutions to make the estimation feasible.

email: meihuawu@umich.edu

## **15. STATISTICAL ANALYSIS OF METABOLOMICS DATA**

### **STATISTICAL ISSUES IN METABOLOMICS**

David L. Banks\*, Duke University

Statistical analysis of metabolomics data poses significant challenges in multivariate metrology. This talk lays out the measurement issues, and describes how one can create uncertainty budgets and make cross-platform inferences using the Mandel bundle-of-lines model. It

also addresses the problem of data mining for signal detection in two applications, recommending Random Forests as a tool that works well and that seems heuristically appropriate.

email: banks@stat.duke.edu

## **STATISTICAL WAYS TO CHOOSE A DISTANCE MEASURE FOR METABOLOMIC DATA**

Philip M. Dixon\*, Iowa State University

Most methods for analyzing metabolomic data require the choice of a distance measure, but this choice is commonly made for non-statistical reasons. This talk will discuss two related issues: how to use properties of the data to choose a distance measure and statistical criteria to evaluate the choice of measure. In various data sets with biological replicates, I find a power-law relationship between the variance among replicates and the mean signal. This suggests a new class of variance-weighted distance measures. These measures are compared to traditional distance measures using two novel evaluation criteria that focus on the performance of the distance measure. These are the repeatability given resampling of metabolites and the consistency with a structure expected to be present in the dataset. These methods are illustrated using data on metabolites in the herbal plant genus Echinacea.

email: pdixon@iastate.edu

## **INCORPORATING INTERACTIVE GRAPHICS INTO METABOLOMICS DATA PRE-PROCESSING**

Dianne Cook\*, Iowa State University  
Michael Lawrence,  
Suh-yeon Choi,  
Heike Hofmann,  
Eve Wurtele,

In metabolomics experiments the purpose is to determine quantities of metabolites present in a sample and compare between samples. Methods that were developed for detecting a small number of metabolites are now being used to detect hundreds of metabolites. Many of the metabolites are also unidentifiable because they do not exist in any library of compounds. To determine the quantities of metabolites the GC-MS raw data needs to be processed to extract the peaks, for each mass-to charge ratio ( $m/z$ ), and group the peaks that occur in close temporal locations to produce a metabolite profile. The raw data suffers from background noise, and temporal shifts from one sample to another. Thus, pre-processing of the data is messy. A plus is that there are always replicates for each treatment and these can be used to stabilize the estimation of the metabolite profiles. In our work we have incorporated interactive graphics to improve the quality of the pre-processing. Here are examples of ways we can check the pre-processing: (1) Compare replicates and probe places where there is disagreement in the samples from the same treatment; (2) Investigate diagnostic statistics measuring the peak fitting, to check the particularly problematic fits; (3) Examine and adjust the grouping of peaks into metabolite profiles.

email: dcook@iastate.edu

# ABSTRACTS

## 16. ADVANCED STATISTICAL METHODS FOR HEALTH SERVICES RESEARCH

### CHARACTERIZING PATTERNS OF TREATMENT UTILIZATION FOR YOUTH WITH ADHD

Gary Klein\*, Carnegie Mellon University  
Joel Greenhouse, Carnegie Mellon University  
Abigail Schlesinger, Western Psychiatric Institute and Clinic, University of Pittsburgh  
Bradley Stein, Western Psychiatric Institute and Clinic, University of Pittsburgh, Community Care Behavioral Health Organization, RAND Corporation

The primary treatments for Attention Deficit/Hyperactivity Disorder (ADHD) in school-aged children are medication and behavioral interventions. Using two sources of Medicaid claims data, behavioral health claims and retail pharmacy claims, we characterize the long-term utilization of these treatments in a more representative population than is typically found in clinical trials data. Methods used include survival analysis, multi-state Markov models, and trajectory analysis. At the outset, we discuss some of the difficulties of working with these types of administrative data, especially for an inexperienced researcher and clinician trying to establish a collaborative relationship. These include the understanding, merging, and cleaning of two very different databases as well as defining suitable outcomes and covariates which may not be easily extrapolated from the given data.

email: gklein@stat.cmu.edu

### STATISTICAL STRATEGIES FOR POSTMARKET SURVEILLANCE OF MEDICAL DEVICES

Sharon-Lise T. Normand\*, Department of Health Care Policy, Harvard Medical School

A key problem faced by policy makers is how to assess evidence about the effectiveness and safety of new therapies after they have been approved. Because the evidence for approval is often based on the results of small controlled clinical trials, (1) the patient population and the provider population in the real world can differ dramatically from the trial populations and (2) adverse events are difficult to detect. Approaches to inference on the basis of observational data, focusing on the role of the treatment assignment mechanism, are discussed. Methods are illustrated to examine the safety and effectiveness of drug-eluting coronary stenting compared to bare metal stenting. Issues related to comparison groups and sensitivity to unmeasured confounders are discussed and demonstrated.

email: sharon@hcp.med.harvard.edu

## THE ROLE OF HEALTH AND HEALTH BEHAVIORS IN THE FORMATION AND DISSOLUTION OF FRIENDSHIP TIES

A James O'Malley\*, Harvard Medical School  
Nicholas A. Christakis, Harvard Medical School

An important question in social network analysis is whether observed association in individual characteristics between connected individuals in a network is a consequence of social influences (i.e., forces that act once the tie is formed) or whether individuals who are similar (dissimilar) are more likely to form (break) ties. The latter mechanism is known as homophily and is commonly described as "birds of a feather flock together." In this talk we describe an approach for estimating the magnitude of the homophily effect, and examine the extent to which ties in a social network form or dissolve as a function of the similarity (or lack thereof) of individuals' health-related traits. We also investigate whether the health traits have greater effects than the non-health "unchangeable" traits. If so, this would suggest that those traits which are affected by sociological phenomena have a greater influence on the dynamic behavior of the network. The health behaviors and traits considered are: Body mass index, smoking, depression, and hypertension. For comparison, we also consider two non-health traits, height and handedness (left or right-handed) that are immutable.

email: omalley@hcp.med.harvard.edu

## 17. MODEL SPECIFICATION AND UNCERTAINTY IN ECOLOGICAL ANALYSES

### DATA-MODEL INTEGRATION FOR UNDERSTANDING BELOWGROUND ECOSYSTEMS

Kiona Ogle\*, University of Wyoming

Developing a mechanistic understanding of belowground ecosystem dynamics (e.g., soil carbon storage and fluxes) is critical to understanding whole-ecosystem behavior (e.g., net ecosystem carbon exchange). In particular, soils are important players in the global carbon cycle, and developing a quantitative understanding of the belowground system is critical to forecasting climate change impacts. Significant advances have been made in belowground ecosystem ecology, but several challenges remain. One important problem is the ability to rigorously partition the effects of different belowground processes and to identify how they vary across space and time. Modern statistical and computational tools, combined with field experiments and ecological process modeling, provide a rigorous approach for reconstructing belowground processes. The approach employs a hierarchical Bayesian (HB) framework that simultaneously analyzes diverse data sources within the context of process-based models. Process parameters and latent variables are partially constrained by information from published studies (priors) and by the underlying structure of the ecological process model. An example is presented that couples diverse laboratory and field data with soil carbon flux models to partition sources (e.g., plant- vs. microbial-derived, old vs. new, shallow vs. deep carbon) of soil carbon efflux in a desert ecosystem.

email: kogle@uwyo.edu

## A BAYESIAN BIOCLIMATE MODEL FOR THE LOWER TROPHIC ECOSYSTEM IN THE NORTH PACIFIC OCEAN

Christopher K. Wikle\*, University of Missouri

A wide variety of physical-biological models of varying complexity have been developed for components of the U.S. West Coast upwelling ecosystem. For example, the physical-biological interface is effective for demonstrating high-resolution properties of phytoplankton bloom and zooplankton population dynamical response. We develop a simple stochastic three-component model (nitrogen, phytoplankton, zooplankton) in a hierarchical Bayesian framework. The underlying system is multivariate and highly nonlinear. The model is demonstrated for the upwelling region of the North Pacific. The example demonstrates the interplay between model formulation, data, and interpretation of posterior distributions in relatively simple stochastic models of a complicated physical-biological system. Ultimately, such models can be used for pan-regional syntheses and climate change impact studies for coastal ocean ecosystems.

email: wiklec@missouri.edu

## MODELING AND INFERENCE OF ANIMAL MOVEMENT IN RESPONSE TO LANDSCAPES

Jun Zhu\*, University of Wisconsin-Madison

Jeff Tracey, Colorado State University

Kevin Crooks, Colorado State University

Movement of animals in landscape is an important subject in ecology and conservation biology, yet many of the models used by ecologists do not account for landscape features and thus may not be conducive to the analysis of animal movement data. Here we present new statistical models that feature animal movement in relation to objects in a landscape. Statistical inference including parameter estimation and model assessment is developed. For illustration, we show results of simulated data and a real movement data set collected on rattlesnakes in San Diego, California.

email: jun.zhu.e@gmail.com

## 18. ANALYSIS CHALLENGES OF MODERN LONGITUDINAL BIOMEDICAL DATA

### INCORPORATING CORRELATION FOR MULTIVARIATE FAILURE TIME DATA WHEN CLUSTER SIZE IS LARGE

Li Wang, Oregon State University

Lan Xue, Oregon State University

Annie Qu\*, University of Illinois at Urbana-Champaign

We propose a new estimation method for multivariate failure time data using the quadratic inference function (QIF) approach. The proposed method efficiently incorporates within-cluster correlations. Therefore it is more efficient than those which ignore within-cluster correlation. Furthermore, the proposed method is easy to implement. Unlike the weighted estimating equations in Cai & Prentice (1995), it is not necessary to explicitly estimate the correlation parameters. This simplification is particularly useful in analyzing data with large cluster size where it is difficult to estimate intraclass correlation. Under certain regularity conditions, we show the consistency and asymptotic normality of the proposed QIF estimators. A Chi-squared test is also developed for hypothesis testing. We conduct extensive Monte Carlo simulation studies to assess the finite sample performance of the proposed methods.

email: anniequ@illinois.edu

### VARIABLE SELECTION IN ADDITIVE MIXED MODELS FOR LONGITUDINAL DATA

Daowen Zhang\*, North Carolina State University

In longitudinal studies with a potentially large number of covariates, investigators are often interested in identifying important variables that are predictive of the response. Suppose we can a priori divide the covariates into two groups: one where parametric effects are adequate and the other where nonparametric modeling is required. In this research, we propose a new method to simultaneously select important parametric covariates and nonparametric covariates in additive mixed models for longitudinal data. Simulation will be used to evaluate the performance of the new method and a real data analysis is used to illustrate its application.

email: zhang@stat.ncsu.edu

# ABSTRACTS

## INDIVIDUALIZED PREDICTION IN PROSTATE CANCER STUDIES USING A JOINT LONGITUDINAL-SURVIVAL-CURE MODEL

Menggang Yu\*, Indiana University, School of Medicine

For monitoring patients treated for prostate cancer, Prostate Specific Antigen (PSA) is measured periodically after they receive treatment. Increases in PSA are suggestive of recurrence of the cancer and are used in making decisions about possible new treatments. The data from studies of such patients typically consist of longitudinal PSA measurements, censored event times and baseline covariates. Methods for the combined analysis of both longitudinal and survival data have been developed in recent years, with the main emphasis being on modeling and estimation. We analyze a training data set from prostate cancer study in which the patients are treated with radiation therapy using a joint model that has been extended by adding a mixture structure to the survival model component of the model. Here we focus on utilizing the model to make individualized prediction of disease progression for censored and alive patients. Results are illustrated on a validation data set.

email: menggang.yu@gmail.com

## LONGITUDINAL ANALYSIS OF SURGICAL TRIALS WITH NON-COMPLIANCE

Patrick J. Heagerty\*, University of Washington  
Colleen Sitrani, University of Washington

Randomized surgical trials with the goal of evaluating the long-term benefit of surgical intervention as compared to a non-surgical treatment are often faced with serious patient non-compliance. Frequently subjects assigned to surgery delay or subsequently refuse surgery, while non-surgical subjects may ultimately seek and receive surgery. There are several statistical challenges associated with longitudinal "as-treated" analyses that seek to estimate average causal effects attributable to surgery. In this presentation we adopt an underlying longitudinal causal mixed model that is a natural example of a structural nested mean model, and then compare the performance of alternative analysis methods when endogenous processes lead to patient crossover (e.g. from non-surgical to surgical). Standard linear mixed models may not be valid yet can perform surprisingly well when selection bias is modest and non-differential. In contrast, Causal estimation methods such as G-estimation and instrumental variable approached can be valid and their implementation in this setting will be reviewed.

email: heagerty@u.washington.edu

## 19. RECENT ADVANCES ON FEATURE SELECTION AND ITS APPLICATIONS

### FEATURE SELECTION IN GLM WITH LARGE MODEL SPACES

Jiahua Chen\*, University of British Columbia  
Zehua Chen, The National University of Singapore

In genome-wide association studies, hundreds of thousands single nucleotide polymorphisms are screened to identify the ones that are most responsible to the genetic variation under investigation. At the same time, the sample size or the number of biological replications are at most in thousands. In such applications, the number of independent variables far exceeds the sample size. In such "large-p-small-n" situations, the classical variable selection criteria such as BIC are found far too liberal. The extended Bayes information criterion proposed by Chen and Chen (2008), in contrast, provides effective control on false discovery rate while retaining comparable positive discovery rate. Furthermore, the criterion has been shown to be consistent at identifying the true set of variables in the normal linear regression model. In this talk, we investigate the property of the extended Bayes information criterion under the generalized linear model. In particular, we show that under some mild conditions, the extended Bayes information criterion remains consistent.

email: jhchen@stat.ubc.ca

### HIGHER CRITICISM THRESHOLDING: OPTIMAL FEATURE SELECTION WHEN FEATURES ARE RARE AND WEAK

Jiashun Jin\*, Carnegie Mellon University  
David L. Donoho, Stanford University

For high-dimensional classification, an important approach is to use thresholding to select features. However, it remains an open problem how to set the threshold. In this paper, we propose a new approach which sets the threshold by Higher Criticism, a recent statistic proposed in Donoho and Jin (2004). We show that the Higher Criticism thresholding is optimal for many classifiers and for many circumstances. Comparison to recent classification methods (including the Least Centroid Shrunk and False Discovery Rate Thresholding (FDRT)) are investigated both with simulated data and with microarray data.

email: jiashun@stat.cmu.edu

### WEIGHTED WILCOXON-TYPE SMOOTHLY CLIPPED ABSOLUTE DEVIATION METHOD

Lan Wang\*, University of Minnesota  
Runze Li, The Pennsylvania State University

Shrinkage-type variable selection procedures have recently seen increasing applications in biomedical research. However, their performance can be adversely influenced by outliers in either the response or the covariate space. This paper proposes a weighted Wilcoxon-type smoothly clipped absolute deviation (WW-SCAD)

method, which deals with robust variable selection and robust estimation simultaneously. The new procedure can be conveniently implemented with the statistical software R. We establish that the WW-SCAD correctly identifies the set of zero coefficients with probability approaching one and estimates the nonzero coefficients with the rate  $n^{-1/2}$ . Moreover, with appropriately chosen weights the WW-SCAD is robust with respect to outliers in both the x and y directions. The important special case with constant weights yields an oracle-type estimator with high efficiency at the presence of heavier-tailed random errors. The robustness of the WW-SCAD is partly justified by its asymptotic performance under local shrinking contamination. We propose a BIC-type tuning parameter selector for the WW-SCAD. The performance of the WW-SCAD is demonstrated via simulations and by an application to a study that investigates the effects of personal characteristics and dietary factors on plasma beta-carotene level.

email: lan@stat.umn.edu

## ULTRAHIGH DIMENSIONAL VARIABLE SELECTION: BEYOND THE LINEAR MODEL

Jianqing Fan\*, Princeton University  
 Richard Samworth, University of Cambridge  
 Yichao Wu, North Carolina State University

Variable selection in ultrahigh-dimensional space characterizes many contemporary problems in scientific discovery and decision making. A frequently-used technique is the independent screening such as the correlation ranking (Fan and Lv, 2008) or feature selection using a two-sample  $t$ -test in high-dimensional classification. Within the context of the linear model, Fan and Lv (2008) showed that this simple correlation ranking possesses a sure independence screening property under certain conditions and that its revision, called iteratively sure independent screening (ISIS), is needed when the features are marginally unrelated but jointly related to the response variable. In this paper, we extend ISIS, without explicit definition of residuals, to a general pseudo-likelihood framework, which includes generalized linear models as a special case. Even in the least-squares setting, the new method improves ISIS by allowing variable deletion in the iterative process. Our technique allows us to select important features in ultrahigh-dimensional classification where the popularly used two-sample  $t$ -method fails. A new technique is introduced to reduce the false discovery rate in the feature screening stage. Several simulated and a real data example are presented to illustrate the methodology.

email: jqfan@princeton.edu

## 20. ANALYSIS OF GENOME-WIDE SNP ARRAYS

### PHARMACOGENOMICS META-ANALYSIS - A SIMULATION STUDY

Mei Yang\*, Boston University  
 Meng Chen, Pfizer Global Research & Development

For most currently performed pharmacogenomics studies, trials tend to be small with low power. Meta-analysis is often used to

borrow information across related trials to dependably detect genetic associations. We studied the operating characteristics of different meta-analysis approaches in detecting genetic signals for binary safety outcomes under the dominant model. Data were simulated under three scenarios to represent circumstances with different levels of heterogeneity. Frequentist approaches (i.e. fixed/random effect model and cumulative meta-analysis) and Bayesian approaches (i.e. full Bayesian and empirical Bayes) were performed on each scenario. Under the moderate heterogeneity scenario, which is most common, we found Bayesian model gave narrower width of interval estimate for the genetic effect compared with the random effect model. And effect estimate for each trial was shrunken towards the overall mean. Furthermore, Bayesian models provide us the options to model early stage trials. If they are exchangeable to the later ones, Bayesian model with non-informative priors on all trials is more appropriate; if they provide useful information that needs to be incorporated, empirical Bayes model with priors formed from early trials is more appropriate; if they are totally irrelevant, Bayesian model with non-informative priors on later trials is more appropriate.

email: meiyang@bu.edu

## SIMULTANEOUS BAYESIAN MULTIPLE SHRINKAGE INFERENCE FOR GENETIC ASSOCIATION STUDIES ALLOWING FOR MODE OF INHERITANCE UNCERTAINTY

Nicholas M. Pajewski\*, University of Alabama at Birmingham  
 Purushottam W. Laud, Medical College of Wisconsin

A desirable model for use in high-dimensional genetic association studies would simultaneously consider the effect of all genetic markers and covariates. These studies commonly require considering a large number of markers, most of which are unrelated to the phenotype. Moreover, at each marker the model should allow a variety of modes of inheritance, namely, additive, dominant, recessive, or over-dominant effects. Recently, MacLehose and Dunson (2007) described a flexible multiple shrinkage approach to high-dimensional model building via Bayesian nonparametric priors. The use of these priors facilitates data-driven shrinkage to a random number of random prior locations. Adapting such techniques, we develop Bayesian bivariate semi-parametric shrinkage priors that can be used to allow flexible shrinkage towards the various inheritance modes and, within each mode, shrinkage towards a random number of random effect sizes. The proposed method offers improved power over parametric alternatives, while naturally incorporating the uncertainty in the choice of inheritance mode. We illustrate the proposed method on simulated data based on the International HapMap Project for both quantitative traits and case-control designs.

email: npajewski@ms.soph.uab.edu

# ABSTRACTS

## USING CASES FROM GENOME-WIDE ASSOCIATION STUDIES TO STRENGTHEN INFERENCE ON THE ASSOCIATION BETWEEN SINGLE NUCLEOTIDE POLYMORPHISMS AND A SECONDARY PHENOTYPE

Huilin Li\*, National Cancer Institute  
Mitchell H. Gail, National Cancer Institute

Whole genome scan case-control association studies offer the opportunity to study the association of genotypes with a secondary phenotype in addition to case-control status. Because controls may represent a random sample from the population, they can be used for this purpose. However, the cases can provide additional useful information. We examine under what conditions cases can be used without introducing bias in estimating the association between a genotype and a secondary phenotype. We consider a two-stage model to incorporate the bias from using cases, and propose an Empirical Bayes method to combine the case and control data to estimate the association with reduced mean square error, based on a balance between bias and variance. Both simulated and real data examples suggest the advantage of the newly proposed Empirical Bayes estimator. We also show how to extend the method to accommodate multiple case types and one control group.

email: lih5@mail.nih.gov

## AN ALGORITHM FOR CONSTRUCTING AN IMPRINTED MAP OF THE CANCER GENOME

Louie R. Wu\*, Buchholz High School  
Yao Li, University of Florida  
Rongling Wu, University of Florida and Penn State University  
Arthur Berg, University of Florida

Genetic imprinting is the differential expression of an allele due to its parental origin. Increasing evidence shows that the number of imprinted genes is much higher than previously thought and that imprinting may influence cancer risk. Here we will present a statistical model for detecting genes that trigger imprinting effects on cancer risk by using genotyped single nucleotide polymorphisms (SNPs). The model is derived with a set of unrelated families (composed of two parents and offspring) sampled from a natural population. The model is incorporated with one of the existing hypotheses about cancer pathogenesis - chromosomal instability due to aneuploidy, a syndrome caused by extra or missing chromosomes and constituting some of the most widely recognized genetic disorders in humans. We develop an algorithm for estimating and testing the imprinting effects of genes on cancer risk by distinguishing the parental origin of chromosome doubling. If the SNPs used are assayed from an entire genome, the proposed model can be expanded to construct an imprinted map of the cancer genome. The imprinting model, whose statistical properties are investigated through simulation studies, will provide a useful tool for studying the genetic architecture of cancer risk.

email: rwu@hes.hmc.psu.edu

## A MIXTURE MODEL FOR THE ANALYSIS OF ALLELIC EXPRESSION IMBALANCE

Rui Xiao\*, University of Michigan  
Michael Boehnke, University of Michigan  
Laura Scott, University of Michigan

Genetic polymorphisms that regulate gene expression account for major phenotypic diversity in human, and may also predispose to complex diseases and determine variability in quantitative traits. Polymorphisms affecting gene expression in cis will cause differentiated expression levels depending on which of the two alleles of the SNP of interest are present. Such an allelic expression imbalance (AEI) can be detected in individuals heterozygous for a transcribed SNP. The use of AEI is complementary to testing for SNP-gene expression association and has the advantage of testing both alleles within the same environment in each individual. To detect the association between the SNP of interest and the AEI, we propose a mixture model and corresponding expectation-maximum (EM) algorithm that take into account the linkage disequilibrium (LD) structure between the SNP of interest and the transcribed SNP. In the typical case of small sample size, we describe an approach to estimate the mixing proportion by using LD information from the HapMap to increase power.

email: xiaor@umich.edu

## MAPPING IMPRINTED QUANTITATIVE TRAIT LOCI UNDERLYING ENDOSPERM TRAIT IN FLOWERING PLANT: A VARIANCE COMPONENT APPROACH

Gengxin Li\*, Michigan State University  
Yuehua Cui, Michigan State University

Genomic imprinting has been thought to play an important role in the development in flowing plant. Empirical studies have shown that some economically important endosperm traits are genetically controlled by imprinted genes. However, the exact number and location of the imprinted genes are largely unknown due to the lack of efficient statistical mapping methods. Methods developed for diploid population cannot be directly applied due to the unique triploid inheritance structure of the endosperm genome. Here we propose a statistical variance component framework by utilizing the nature of sex-specific alleles shared identical-by-decent among sibpairs in experimental crosses to map imprinted quantitative trait loci (iQTL) underlying endosperm traits. We propose a new variance component partition method based on the nature of the triploid inheritance pattern and develop an efficient restricted maximum likelihood estimation method in a genome-wide interval scan for estimating and testing the effects of iQTL. Cytoplasmic maternal effect which is believed to have primary influences on yield and grain quality is also considered when testing for genomic imprinting. Extension to multiple QTL analysis is proposed. Both simulation study and real data analysis indicate good performance and powerfulness of the developed approach.

email: ligengxi@stt.msu.edu

## A BAYESIAN CHANGE-POINT ALGORITHM FOR DETECTING COPY NUMBER ALTERATION

Fridtjof Thomas\*, University of Tennessee 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 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 and compare this approach to other approaches currently in use for this purpose. The Bayesian change-point 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

## 21. BIOMARKERS AND DIAGNOSTIC TESTS

### INFORMATION THEORETIC APPROACH TO SURROGATE MARKERS EVALUATION FOR TIME-TO-EVENT CLINICAL ENDPOINTS

Pryseley N. Assam\*, Hasselt University-Belgium  
Abel E. Tilahun, Hasselt University-Belgium  
Ariel Alonso, Hasselt University-Belgium  
Geert Molenberghs, Hasselt University-Belgium

Recent work in the area of surrogate markers validation in the multitrial framework led to definitions in terms of the quality of trial-and individual-level association between a potential surrogate and a true (clinical) endpoint (Buyse et al. 2000, Biostatistics 1, 49- 69). A drawback is that different settings have led to different measures at the individual level. A unified framework for the different settings was developed based on information theory, for outcomes with distribution in the exponential family, leading to a definition of surrogacy with an intuitive interpretation (Alonso and Molenberghs, 2007, Biometrics 63, 180 - 186). The later approach has been applied to a wide range of settings but not to time-to-event (censored) true endpoint, which is the primary objective of this work with focus on the individual-level association. The performance of four measures for surrogate markers validation based on information theoretic approach (ITA) for time-to-event true endpoints was investigated through a simulation study and applied to a case study.

email: pryseley.assamnkouibert@uhasselt.be

## MEDIAN REGRESSION FOR LONGITUDINAL BIOMARKER MEASUREMENTS SUBJECT TO DETECTION LIMIT

Kong Lan\*, School of Public Health, University of Pittsburgh  
Minjae Lee, School of Public Health, University of Pittsburgh

It has become increasingly popular for medical researchers to investigate whether a certain biomarker is useful for the diagnosis and prognosis of a disease. The biomarker measurements are often left censored due to detection limits. Most exiting methods handled the censored observations with maximum likelihood estimation approach. The robust left-censored regression model based on the least absolute deviations (LAD) method has been presented mainly in the field of econometrics. We describe how the LAD approach can be applied to the longitudinal left censored data. We derive the asymptotic properties of the LAD estimators in the median regression model. We conduct a simulation study to evaluate our proposed method and use a dataset from a sepsis study of inflammatory biomarkers for demonstration.

email: leem2@upmc.edu

## A MULTIPLE IMPUTATION APPROACH FOR LEFT-CENSORED BIOMARKERS WITH LIMITS OF DETECTION

Minjae Lee\*, School of Public Health, University of Pittsburgh  
Lan Kong, School of Public Health, University of Pittsburgh

We often encounter left-censored biomarker measurements subject to the limits of detection (LOD). Ignoring or replacing the censored observations with naive imputation method lead to biased estimates of the parameters in the regression analysis. Maximum likelihood methods have been developed when the distribution of the biomarkers are assumed normal. However, the computation can be very intensive or even prohibitive as the number of censored biomarkers increases. Motivated by a sepsis study, where a panel of biomarkers were measured to investigate the association between the sepsis and the biomarkers such as cytokines and coagulation markers, we propose a multiple imputation(MI) approach based on Tobit regression and Gibbs sampling. We conduct simulation study to evaluate the performance of our MI approach and use a sepsis dataset for demonstration.

email: mil21@pitt.edu

## ESTIMATION AND COMPARISON OF THE PREDICTIVENESS CURVE FOR REPEATED MEASURES DESIGN

Kwonho Jeong\*, University of Pittsburgh  
Abdus M. Sattar, University of Pittsburgh  
Lisa Weissfeld, University of Pittsburgh

In the Genetic and Inflammatory Marker of Sepsis (GenIMS) study (a large multicenter cohort study), a number of pro-inflammatory and anti-inflammatory continuous biomarkers associated with severe sepsis and death have been measured longitudinally. In this work,

# ABSTRACTS

we are proposing to extend the theory of the predictiveness curve (PC) that has been developed by Huang, Pepe and Feng (Bcs2007) for longitudinally measured continuous biomarkers data. We fitted the PC using longitudinally measured GenIMS biomarker data for comparison of their effectiveness in predicting the risk of death. The PC has provided a common scale (zero to one) across various markers for comparing the usefulness of a given marker relative to other potential markers. Using this graphical tool, we have compared population distribution of risk of death for a number of competitive biomarkers associated with the disease. An extensive simulation study has been undertaken to establish the properties of the proposed methods under differing scenarios.

email: kwj2@pitt.edu

## AN EVALUATION OF LOGIC FOREST FOR IDENTIFICATION OF DISEASE BIOMARKERS

Bethany J. Wolf\*, Medical University of South Carolina  
Elizabeth H. Slate, Medical University of South Carolina  
Elizabeth G. Hill, Medical University of South Carolina

Adequate screening tools allowing physicians to diagnose diseases in asymptomatic individuals or identify individuals at elevated risk of developing disease have potential to reduce overall disease related mortality. Multiple studies cite the need for noninvasive tests that are both sensitive and specific. Diagnostic tests based on multiple biomarkers may lead to enhanced sensitivity and specificity. Statistical methodologies that can model complex biologic interactions and that are easily interpretable allow for translation of biomarker research into diagnostic tools. Logic regression, a relatively new multivariable regression method that predicts binary outcomes using logical combinations of binary predictors, has the capability to model the complex interactions in biologic systems in easily interpretable models. However the performance of logic regression degrades in noisy data. We implement an extension of logic regression methodology to an ensemble of logic trees, which we call Logic Forest. We conduct a simulation study to compare the ability of logic regression and logic forest to identify interactions among variables that are predictive of disease status. Our findings indicate Logic Forest is superior to logic regression for identifying important predictors, particularly in noisy data. Logic Forest provides a new statistical tool capable of identifying predictors and predictor interactions associated with disease.

email: wolfb@musc.edu

## ESTIMATES OF OBSERVED SENSITIVITY AND SPECIFICITY MUST BE CORRECTED WHEN REPORTING THE RESULTS OF THE SECOND TEST IN A SCREENING TRIAL CONDUCTED IN SERIES

Brandy M. Ringham\*, University of Colorado  
Deborah H. Glueck, University of Colorado

Recommendations for cancer screening are based, in part, on trials that use a series design to compare screening modalities. In a series design, because the decision to conduct the second test depends on the results of the first test, the estimates of sensitivity and specificity for the second test are conditional. Conditional sensitivity and specificity

estimates may differ from unconditional estimates, i.e., those that would have been observed if the second test were used alone. All estimates of diagnostic accuracy may be biased if the true state of disease is not observed. In a common design for screening trials, diagnosis is made solely by use of a reference test. Reference tests are given only to participants who have abnormal screening test results or who experience signs and symptoms of disease during the follow-up period. The remaining participants are assumed to be disease free. The study investigator calculates estimates of diagnostic accuracy using the observed cases of disease. These estimates may not mirror the true results. Using parametric assumptions to derive formulae, we show that the chance of occult disease, and the correlation between the screening tests affect the difference between the unconditional true and the conditional observed sensitivity and specificity.

email: brandelwine@hotmail.com

## BAYESIAN HIERARCHICAL MODELING OF PROBABILITIES FROM REPEATED BINARY DIAGNOSTIC TESTS

Daniel P. Beavers\*, Baylor University  
James D. Stamey, Baylor University  
John W. Seaman III, Baylor University

Fallible diagnostic tests introduce bias into statistical inference via misclassified binary data. One approach to assess the quality of a diagnostic test is to test individuals repeatedly, which can provide sufficient data to estimate the sensitivity and specificity of the test as well as the latent cohort disease prevalence, assuming certain conditions hold. In our work we consider a Bayesian approach to modeling the population prevalence as well as the test sensitivity and specificity. We add hierarchical random variability components to the misclassification models to estimate the inter-individual variability. Model performance is assessed using the deviance information criteria, and we compare our results to existing frequentist methods.

email: daniel\_beavers@baylor.edu

## 22. CAUSAL INFERENCE

### CAUSAL INFERENCE WITH LONGITUDINAL SUBPOPULATION INDICATORS

Booil Jo\*, Stanford University

This paper focuses on settings where treatment compliance is measured over time, but the outcome is not measured in parallel, which is very common in randomized intervention trials. We propose to formulate compliance and/or outcome trajectory strata considering potential outcomes under only one treatment condition and then estimate differential treatment effects conditioning on these strata. The latent strata formulated based on this approach can be thought of as coarse principal strata. The advantage of this approach is that it is possible to construct subpopulation trajectory strata without pairing compliance and outcome information. Our proposed approach nicely complements that of Lin, Ten Have, & Elliot (in press) and deals with situations that are difficult to handle in their framework.

email: booil@stanford.edu



San Antonio, Texas

## A CAUSAL MODEL OF BASELINE AND POST-TREATMENT CONFOUNDING FOR OBSERVATIONAL STUDIES

Chen-pin Wang\*, University of Texas Health Science Center-San Antonio

In clinical practice, the choice of oral glucose-lowering agents in type 2 diabetics depends on patients' recent glucose levels as well as other predictors for cardiovascular diseases (CVD). Prior studies suggested that how well patients' responses to these drugs in terms of the change in glucose levels may have differential impacts on patients' cardiovascular outcomes. This presentation considers a causal model to study the direct effect of two types of glucose-lowering agents on CVD while accounting for confounding due to (i) medication choice at baseline and (ii) heterogeneity in responding to the medication. The proposed model integrates both the inverse probability weighting (IPW) and principal stratification (PS) modeling techniques. The PS component models the variation in drug response to hyperglycemia, while an IPW estimate, nested within each principal stratum, is used to quantify the PS effect. More specifically, the principal strata are identified by latent variable modeling of glycemia trajectory classes, and IPW refers to the inverse of latent class specific propensity score. We demonstrate the proposed model using a clinical cohort of type 2 diabetics from the VA health care system. The impact of the IPW estimate and misspecification in PS modeling will be discussed.

email: wangc3@uthscsa.edu

## ACCOUNTING FOR UNMEASURED CONFOUNDERS WITH LATENT VARIABLE

Haiqun Lin\*, Yale University School of Public Health

Inference about the effects attributable to a treatment condition in non-experimental, i.e. observational studies is critical in many areas of public health research. The pivotal aspect of observational studies is that the treatment assignment is not under control of the investigator. Most existing methods adjust for measured confounding variables under the assumption of no unmeasured confounding. In this paper, we present a method for estimating the effect attributable to a treatment condition when there exist unmeasured confounders. We regard unobserved confounders as a latent variable. We first formulate a joint model of the possibly time-varying treatments and the associated responses that can account for unmeasured (as well as measured confounding variables). The marginal estimate of the effect attributable to a possibly time-varying treatment can be obtained by using a weighted generalized estimation equation model using weights constructed from the joint models. Our method is illustrated with the analysis of a data set from ACCESS (Access to Community Care and Effective Services and Support) and validated through simulation studies.

email: haiqun.lin@yale.edu

## A CAUSAL SELECTION MODEL TO COMPARE TREATMENT GROUPS IN A SUBSET SELECTED POST-RANDOMIZATION WITH APPLICATION TO AN HIV ANTIRETROVIRAL IMMUNOTHERAPY TRIAL

Robin Mogg\*, University of Pennsylvania School of Medicine  
Marshall M. Joffe, University of Pennsylvania School of Medicine  
Devan V. Mehrotra, Merck Research Laboratories  
Thomas R. Ten Have, University of Pennsylvania School of Medicine

A successful therapeutic HIV vaccine could offer significant advantages in fighting the HIV epidemic, including reducing the costs and potential toxicities associated with antiretroviral therapy (ART) by allowing patients to have structured treatment interruptions and potentially contributing to a delay in the onset of AIDS-defining illnesses or death. In this paper, we describe a proof-of-concept (POC) antiretroviral immunotherapy (ARI) trial that was initiated to assess whether immunization with an experimental HIV vaccine has a measurable impact on the control of viremia following subsequent interruption of ART. For a number of reasons, some patients will not interrupt ART after vaccination; as such, a comparison of outcomes among vaccine and placebo patients who interrupt ART must adjust for a potential selective effect of the vaccine on the post-randomization event of ART interruption. We propose a selection model that utilizes the principal stratification framework developed by Frangakis and Rubin (2002) without imposing assumptions on the direction of the selective effect of the vaccine. Methods to test the causal effect of the vaccine on viral load among the principal stratum of patients who would interrupt treatment regardless of randomization assignment are developed. Finite sample properties of the testing methodology are assessed via computer simulation.

email: robin\_mogg@merck.com

## ESTIMATING DRUG EFFECTS IN THE PRESENCE OF PLACEBO RESPONSE

Bengt O. Muthén\*, UCLA  
Hendricks C. Brown, University of South Florida

Placebo-controlled randomized trials on antidepressants and other drugs often show a response for a sizeable percentage of the subjects in the placebo group. Potential placebo responders can be assumed to exist also in the drug group, making it difficult to assess the drug effect. A key drug research focus should be to estimate the percentage of individuals among those who responded to the drug who would not have responded to the placebo ("Drug only responders"). This talk investigates a finite mixture model approach to uncover percentages of up to four potential mixture components: Never Responders, Drug only responders, Placebo only responders, and Always responders. Two examples are used to illustrate the modeling, a twelve-week antidepressant trial with a continuous outcome (Hamilton D Score) and a 7-week schizophrenia trial with a binary outcome (illness level). Growth mixture modeling (Muthén & Asparouhov, 2008) is used to uncover the different mixture components.

email: bmuthen@ucla.edu

# ABSTRACTS

## INFERENCE ON TREATMENT EFFECTS FROM A RANDOMIZED CLINICAL TRIAL IN THE PRESENCE OF PREMATURE TREATMENT DISCONTINUATION: THE SYNERGY TRIAL

Min Zhang\*, University of Michigan

Anastasios A. Tsiatis, North Carolina State University

Marie Davidian, North Carolina State University

Karen S. Pieper, Duke Clinical Research Institute

Kenneth Mahaffey, Duke Clinical Research Institute

The SYNERGY trial was a randomized, open-label clinical trial designed to compare two anti-coagulant drugs on the basis of various time-to-event endpoints. As usual, the protocol dictated circumstances, such as occurrence of a serious adverse event, under which it was mandatory for a subject to discontinue his/her assigned treatment. In addition, as in the execution of many trials, some subjects did not complete their assigned treatment regimens but rather discontinued study drug prematurely for other, "optional" reasons not dictated by the protocol; e.g., switching to the other study treatment or stopping treatment altogether at their or their provider's discretion. In this situation, as an adjunct to the usual intent-to-treat analysis, interest may focus on inference on the "true" treatment effect; i.e., the difference in survival distributions were all subjects in the population to follow the assigned regimens and, if to discontinue treatment, do so only for mandatory reasons. Approaches to inference on this effect used commonly in practice are ad hoc and hence are not generally valid. We use SYNERGY as a motivating case study to propose generally-applicable methods for estimation and testing of this "true" treatment effect by placing the problem in the context of causal inference on dynamic treatment regimes.

email: mzhangst@umich.edu

## DETECTION OF SURROGATES USING A POTENTIAL OUTCOMES FRAMEWORK

Andreas G. Klein\*, University of Western Ontario

This paper proposes a definition of surrogacy that is based on potential outcome notation. Surrogacy is defined as an association between individual causal effects of a treatment on an intermediate variable and an outcome. The idea that a surrogate marker is an indicator of an unobserved intermediate mechanism that lies in the pathway of a causal process is given an exact formalization. Methodological difficulties with regard to Prentice's concept of surrogacy are discussed. A simulation-based procedure is presented that - under certain assumptions about the data and the nature of the causal process - can quantify the association between the individual causal effects on the intermediate variable and the outcome. The application of the new procedure is illustrated using an empirical data set.

email: aklein25@uwo.ca

## 23. AN EM APPROACH FOR PARTIAL CORRELATION AND MISSING DATA

Gina D'Angelo\*, Washington University School of Medicine  
Chengjie Xiong, Washington University School of Medicine

In the cognitive neuroscience area it is often of interest to study brain co-activation and relationships among these regional brain measures and psychometric measures. Many of these studies by design have frequency matching on age and gender, and often it is necessary to adjust for these covariates. Partial correlation is a statistical measure that can be used to correlate two variables while adjusting for other variables. In the presence of data that are missing at random, complete case analysis will lead to biased and inefficient results. We will extend the partial correlation coefficient in the presence of missing data using the EM algorithm and compare it with a multiple imputation method and complete case analysis. Another objective of this work is to compare partial correlations between groups, and we will extend the correlation analysis of variance (CORANOVA) approach (Bilker et al., 2002) to handle missing at random data. We will compare the EM algorithm to the multiple imputation method and complete case analysis method using simulation studies and these methods will be illustrated with a depression diffusion tensor imaging (DTI) study.

email: gina@wubios.wustl.edu

## LATENT VARIABLE REGRESSION FOR MULTIPLE OUTCOMES WITH SOME PREDICTORS MISSING NON-RANDOMLY

Jieqiong Bao\*, Emory University  
Amita Manatunga, Emory University  
Andrew Taylor, Emory University

We are interested in developing a latent model for predicting multiple outcomes in terms of observed covariates. A complication occurs because some covariates are not always observed due to the nature of sampling scheme. For example, in the renal image studies, it is of interest to predict the need for furosemide and kidney obstruction based on characteristics of renal images. The predictors that are crucial for predicting kidney obstruction are only present when only a patient receives furosemide based on a clinical decision. Two observed outcomes are the need for furosemide and the presence of kidney obstruction for each kidney which are evaluated by three experts. We propose a latent variable regression model to predict the underlying outcome while accounting for the missingness of the predictors. Our modeling framework provides estimates for the between-rater variability and within-rater variability simultaneously under maximum likelihood framework. We apply to the nuclear medicine data from MAG3 renography studies.

email: jbao@sph.emory.edu



## OUTFLUENCE -- THE IMPACT OF MISSING VALUES

Ofer Harel\*, University of Connecticut

There are numerous measures that assess the effect of an observation, group of observations, a variable, or variables and observations on the regression estimation. Incomplete data is a common difficulty in data analysis. I introduce a new measure which assesses the effect of a missing observation, a group of missing observations, an incomplete variable or any combination of these, on the overall estimation. I call this measurement "outfluence". The outfluence measure can be used in a regression analysis context or any other parametric settings. I illustrate the major benefits of outfluence using biomedical examples.

email: ofer.harel@uconn.edu

## A DOUBLE ROBUST LOCAL MULTIPLE IMPUTATION

Chiu-Hsieh Hsu\*, University of Arizona  
Qi Long, Emory University

A robust local multiple imputation approach is proposed to recover information for missing observations. To conduct the imputation, we use two working models to create two predictive scores. One score is derived from a linear regression model to predict the missing values. The other is derived from a logistic regression model to predict the missing probabilities. The two scores are then used to decide the resampling and imputing probabilities via kernel regression for each missing observation. Under an assumption of missing at random mechanism, the imputation approach is shown to be robust to misspecification of either of the two working models and to be robust to misspecification of the distribution of the data. In addition, simulation comparisons with other methods suggest that the method works well in a wide range of populations. The approach is demonstrated on a dataset from a colorectal polyp prevention trial.

email: phsu@azcc.arizona.edu

## AVOID ECOLOGICAL FALLACY: USING BART TO IMPUTE MISSING ORDINAL DATA

Song Zhang\*, University of Texas Southwestern Medical Center  
Tina Shih, University of Texas M.D. Anderson Cancer Center  
Peter Muller, University of Texas M.D. Anderson Cancer Center

Ecological fallacy is a situation that can occur when making inference about an individual based on aggregate data from a group. In health disparity research, a missing individual-level social-economical-status (SES) variable is usually replaced by a census-based SES statistic, which is considered a proxy for the individual SES. We use a real data example to demonstrate the potential biases associated with the census-based approach, as a result of ecological fallacy. We further propose a Bayesian additive regression tree (BART) method to impute missing ordinal data (household income category), utilizing statistical learning from a different dataset with the income variable observed. The imputation based on BART is shown to be a better proxy for the missing variable than census-based SES, and it avoids ecological fallacy when making inference about the important factors in the disparity of colorectal cancer screening utilization.

email: song.zhang@utsouthwestern.edu

## META-ANALYSIS OF STUDIES WITH MISSING DATA

Ying Yuan\*, University of Texas M.D. Anderson Cancer  
Roderick Little, University of Michigan

Consider a meta-analysis of studies with varying proportions of patient-level missing data, and assume that each primary study has made certain missing data adjustments so that the reported estimates of treatment effect size and variance are valid. These estimates of treatment effects can be combined across studies by standard meta-analytic methods, employing a random-effects model to account for heterogeneity across studies. However, we note that a meta-analysis based on the standard random-effects model will lead to biased estimates when the attrition rates of primary studies depend on the size of the underlying study-level treatment effect. Perhaps ignorable within each study, this type of missing data is in fact not ignorable in a meta-analysis. We propose three methods to correct the bias resulting from such missing data in a meta-analysis: re-weighting the DerSimonian-Laird estimate by the completion rate; incorporating the completion rate into a Bayesian random-effects model; and inference based on a Bayesian shared-parameter model that includes the completion rate. We illustrate these methods through a meta-analysis of 16 published randomized trials that examined combined pharmacotherapy and psychological treatment for depression.

email: yyuan@mdanderson.org

## IMPROVING EFFICIENCY AND ROBUSTNESS OF THE DOUBLY ROBUST ESTIMATOR FOR A POPULATION MEAN WITH INCOMPLETE DATA

Weihua Cao\*, North Carolina State University  
Anastasios A. Tsiatis, North Carolina State University  
Marie Davidian, North Carolina State University

Considerable recent interest has focused on doubly robust estimators for a population mean response in the presence of incomplete data, which involve models for both the propensity score and the regression of outcome on covariates. The 'usual' doubly robust estimator may yield severely biased inferences if neither of these models is correctly specified and can exhibit nonnegligible bias if the estimated propensity score is close to zero for some observations. We propose alternative doubly robust estimators that achieve comparable or improved performance relative to existing methods, even with some estimated propensity scores close to zero.

email: wcao5@ncsu.edu

## 24. POWER/SAMPLE SIZE

### POWER ANALYSIS FOR MEDIATION EFFECT IN LONGITUDINAL STUDIES

Cuiling Wang\*, Albert Einstein College of Medicine

Mediation effect is often of great interest in longitudinal epidemiological research. Current literature on longitudinal mediation analysis mostly focuses on time dependent mediators, while in some studies both the independent variable and the mediator are time

# ABSTRACTS

independent. Power analysis methods for such study designs have not been adequately addressed. In this work we derive formulas based on asymptotic theory for calculation of power for the longitudinal mediation effect of a time-independent mediator on a time-independent predictor on repeated measures outcome, with and without drop out. Performance of the formulae for limited sample sizes were examined through simulation studies. The method was applied to a project in the design of Mobility Biology & Interventions Study at Einstein a longitudinal (MOBILISE) study where the mediation effect of markers of cerebral microvascular status and function on the associations of cardiovascular risk factors with decline in gait velocity.

email: cuwang@aecon.yu.edu

## POWER AND TYPE I ERROR RATES IN REPEATED MEASURES EXPERIMENTS AS THE NUMBER OF TIME POINTS IN A FIXED LENGTH TIME INTERVAL INCREASE AND UNDER SEVERAL COVARIANCE STRUCTURES FOR THE REPEATED MEASURES

John D. Keighley\*, University of Kansas Medical Center  
Dallas E. Johnson, Kansas State University

Linear mixed models are by far the most popular modeling choice for longitudinal analyses because they offer flexible mean and covariance structure choices. However, researchers can be challenged at designing future experiments for which linear mixed models are used. In particular, how many time points should they collect? This basic question was addressed in this paper by calculating power as a function of equally spaced time points with different expected response functions and covariance structures. The main focus was on testing treatment by time interaction and using various single degree of freedom contrasts. Countering expectations, power decreased or remained constant with an increase in the number of time points. The simulation methods and results in this paper can guide researchers in determining the number of time points for their future study.

email: jkeighle@kumc.edu

## INTRACLUSTER CORRELATION ADJUSTMENTS TO MAINTAIN POWER IN CLUSTER TRIALS FOR BINARY VARIABLES

Hrishikesh Chakraborty\*, RTI International  
Janet Moore, RTI International  
Tyler D. Hartwell, RTI International

Adequately powered sample size calculations for cluster randomized trials primarily depend on the event rate variability, effect size, average cluster size, and intracluster correlation (ICC). Furthermore, an ICC estimate depends on event rate variability among clusters, cluster size, and number of clusters. We evaluated the impact of event rates, event rate variations, cluster size, cluster size variations for different numbers of clusters. We also evaluated how the event rate changes at the end of the trial effect ICC estimates. We created one simulation exercise to investigate how different event rates, event rate variations, cluster size, and cluster size variations impact ICC estimates and 95%

confidence intervals. A separate simulation exercise in four different trial scenarios examined the impact of an intervention or drug effect in the intervention group on ICC estimates and 95% confidence intervals and on sample size. The first simulation results suggest that the ICC value depends upon the event rate and event rate variations in addition to the cluster size, cluster size variations, and number of clusters. The second simulation exercise suggested that adjusting the sample size will help to preserve the appropriate power at the end of the trial.

email: hchakraborty@rti.org

## A TWO-STAGE ADAPTIVE DESIGN CAN INCREASE POWER FOR MULTIPLE COMPARISONS

Deborah H. Glueck\*, University of Colorado  
Anis Karimpour-Fard, University of Colorado  
Keith E. Muller, University of Florida

Testing the association between multiple polymorphisms and a disease state involves multiple hypothesis testing. All of the many methods to avoid inflating the Type I error rate do so at the price of a decline in power. For multiple comparison procedures like the Benjamini-Hochberg procedure, the power decreases as a function of the number of hypotheses tested. We propose a two-stage adaptive design to increase power for multiple comparisons. For the two-stage process, 1) estimate the number of null hypotheses in the experiment and choose a truncation point based on this estimate. 2) Conduct a multiple comparison procedure for all hypotheses whose p-values are less than the truncation point. The two-stage adaptive design typically increases power while still providing control of the type I error rate. The increase in power occurs because fewer hypotheses are considered.

email: Deborah.Glueck@uchsc.edu

## LONG TERM SURVIVOR MODELS AND TWO COMPONENT MIXTURE MODELS

Wonkuk Kim\*, University of South Florida

The test of whether survival data follow a long term survivor model or a two component mixture model can be made out using the likelihood ratio test. The sample size and the asymptotic power of the likelihood ratio test are calculated under the generalized type I censoring. An example when the component density function is an exponential distribution is given.

email: wkim@cas.usf.edu

## DETERMINATION OF SAMPLE SIZE FOR VALIDATION STUDY IN PHARMACOGENOMICS

Youlan Rao\*Yoonkyung Lee, The Ohio State University  
Jason C. Hsu, The Ohio State University

Pharmacogenomics aims at co-development of a drug that targets a subgroup of patients with efficacy and a device that identifies the responder group through patients' genetic variations. Development

of such a prognostic device includes a training stage and a validation stage. The transition from the training stage to the validation stage typically involves change of platforms as a subset of the genes predictive of drug response are identified in the first stage and only those are used in the second stage. With the change in consideration, this paper concerns how to determine sample sizes for the validation stage to meet pre-specified sensitivity and specificity requirements in order to avoid futility of pharmacogenomic development. In particular, taking microarrays which measure gene expression levels as a primary device, we show how to decide the numbers of subjects per group, replicated samples per subject, replicated probes per sample for the validation experiment. The change of platforms is taken into account in the sample sizes calculation by linear mixed effect modeling. Our formulation of sensitivity and specificity requirements calls for confidence lower bounds of both measures, which lead to a slightly conservative procedure. The procedure is illustrated in a proof-of-concept mice experiment.

email: rao@stat.osu.edu

## R PROGRAMS FOR CALCULATING SAMPLE SIZE AND POWER IN BIOEQUIVALENCE TRIALS

Qinfang Xiang\*, Endo Pharmaceuticals, Inc.

Crossover designs are the primary statistical designs for bioavailability and bioequivalence studies. Sample size and power calculation could be challenging for high order crossover designs since the calculations involve non-central t-distribution and therefore numerical integration or approximation has to be used. Some existing standard statistical software for power and sample size calculations do not provide such functionalities. R language is open source for statistical computing and has been widely used in academics and industry around the world. This presentation introduces a set of R programs to determine sample size and power for average bioequivalence testing using crossover designs, including standard  $2\times 2$  crossover designs, commonly used four higher order crossover designs, replicated  $2\times 2m$  (me2) crossover designs, and Williams designs. Some power curves and sample size tables are also easily to be generated using the provided R programs.

email: xiang.qinfang@endo.com

## 25. MULTIVARIATE SURVIVAL

### NONPARAMETRIC QUANTILE ESTIMATION FOR SUCCESSIVE EVENTS SUBJECT TO CENSORING

Adin-Cristian Andrei\*, University of Wisconsin-Madison

Quantiles represent important and useful summary measures used in survival analysis. For example, in a cancer clinical trial investigators may want to know the quantiles of the time-to-death, measured from disease relapse, given that one has been disease-free for one year. Conditional quantile estimates based on the conditional Kaplan-Meier curve are not consistent, due to induced dependent censoring. Methodology for properly estimating such quantities in successive or recurrent time-to-event settings is lacking. By consistently estimating the joint distribution of the successive times involved, we develop consistent nonparametric conditional quantile estimators and provide confidence intervals by inverting a test statistic. Simulations performed

in a variety of scenarios confirm the good functional characteristics of the method. An example from the International Breast Cancer Study Group Trial V is used to illustrate the practical usefulness of this methodology.

email: andrei@biostat.wisc.edu

### NONPARAMETRIC AND SEMIPARAMETRIC ESTIMATIONS FOR BIVARIATE FAILURE TIME DISTRIBUTION WITH INTERVAL SAMPLING

Hong Zhu\*, Johns Hopkins University  
Mei-Cheng Wang, Johns Hopkins University

In medical follow-up studies, ordered bivariate survival data are frequently encountered when bivariate failure events are used as the outcomes to identify the progression of a disease. In cancer studies interest could be focused on bivariate failure times, for example, time from birth to cancer-onset and time from cancer onset to death. This paper considers a sampling scheme where the first failure event is (cancer-onset) identified within a calendar time interval, the time-origin (birth) can be retrospectively confirmed, and the occurrence of the second event (death) is observed subject to right censoring. To analyze the bivariate failure time data, it is important to recognize the presence of bias arising due to the complex interval sampling. In this paper, nonparametric and semiparametric methods are developed for estimating the bivariate failure time distribution under stationary and semi-stationary conditions. Statistical methods based on a nonparametric model and a semiparametric copula model are developed to address the problems. The research could also be extended to the situation when the bivariate failure events are observed subject to both interval and prevalent sampling.

email: hongzhu@jhsph.edu

### PARTIALLY MONOTONE SPLINE ESTIMATION WITH BIVARIATE CURRENT STATUS DATA

Yuan Wu\*, University of Iowa  
Ying Zhang, University of Iowa

Assuming that the data structure of two failure times is referred as bivariate current status data, a partially monotone spline estimation of the joint distribution is proposed. The consistency properties of the estimation are studied. A bootstrap test for the dependency between the two failure times is implemented. Simulation results are presented.

email: yuan-wu@uiowa.edu

### COMPARISON OF STATE OCCUPATION, ENTRY, EXIT AND WAITING TIMES IN K INDEPENDENT MULTISTATE MODELS UNDER CURRENT STATUS DATA

Ling Lan\*, Medical College of Georgia  
Somnath Datta, University of Louisville

We propose distance based nonparametric bootstrap tests comparing the occupation probabilities and entry, exit and waiting times in a

# ABSTRACTS

given state in two or more multistate systems each of which has the same topology. The actual transition times are subjected to current status censoring resulting from a single random inspection time for each individual. A detailed simulation study shows that the proposed test has close to nominal size and reasonable power. An illustrative application to a pubertal development data set obtained from the NHANES III is also presented.

email: llan@mcg.edu

## ADDITIONAL HAZARDS MODEL FOR CASE-COHORT STUDIES WITH MULTIPLE DISEASE OUTCOMES

Sangwook Kang\*, University of Georgia  
Jianwe Cai, University of North Carolina at Chapel Hill

A case-cohort study design is widely used to reduce the cost of large cohort studies while achieving the same goals, especially when the disease rate is low. A key advantage of the case-cohort study design is its ability to use the same subcohort for several diseases or for subtypes of disease. In order to compare the effect of a risk factor on different types of diseases, times to different events need to be modeled simultaneously. Valid statistical methods which take the correlations among the outcomes from the same subject into account need to be developed. To this end, we consider marginal additive hazards model for case-cohort studies with multiple disease outcomes. We also consider the generalized case-cohort designs which do not require sampling of all the cases, which is more realistic for multiple disease outcomes. We propose an estimating equation approach for parameter estimation with two different types of weights. Asymptotic properties of the proposed estimators are investigated and their finite sample properties are assessed via simulations studies. The proposed methods are applied to the Busselton Health Study.

email: skang@uga.edu

## MODELLING CUMULATIVE INCIDENCES OF DEMENTIA AND DEMENTIA-FREE DEATH USING A NOVEL THREE-PARAMETER LOGISTIC FUNCTION

Yu Cheng\*, University of Pittsburgh

Parametric modelling of univariate cumulative incidence functions and logistic models have been studied extensively. However, to the best of our knowledge, there is no study using logistic models to characterize cumulative incidence functions. We hence propose a novel parametric model which is an extension of a three-parameter logistic function. The modified model can accommodate various shapes of cumulative incidence functions and be easily implemented using standard statistical software. The simulation studies demonstrate the good performance of the proposed model when it is correctly specified and the robustness of the model when the underlying cumulative incidence function does not follow the three-parameter logistic function. The practical utility of the modified three-parameter logistic model is illustrated using the data from the Cache County Study of dementia.

email: yucheng95@gmail.com

## GENERALIZED t-TEST FOR CENSORED DATA

Mi-Ok Kim\*, Cincinnati Children's Hospital Medical Center

We consider the problem of testing that two populations are identical with respect to the distribution of a continuous variable subject to random right censoring against the alternative that values tend to be larger in one population. In practice the alternative is usually formulated by the proportional hazards model and log-rank test is standard. In application, however, data often suggests departure from the proportional hazards model, in which case log-rank test loses its optimality and weighted versions of log-rank test have been proposed (see Harrington and Fleming 1982; Peto and Peto 1972; Prentice and Marek 1979). However, the selection of the weights is problematic as the power depends on where and how the two hazard curves differ, while such information is usually unknown. O'Brien (1988) formulated this problem as heterogeneity in the difference of log-rank scores and proposed an extension of log-rank test where group membership is regressed against the log-rank scores using a quadratic model. We consider two extensions: firstly, we note that the power of the test is limited by the use of second degree polynomials and propose an extension with splines. Secondly we consider an extension that allows covariates and different censoring distributions.

email: miok.kim@cchmc.org

## 26. PANEL DISCUSSION: BAYESIAN METHODS IN CLINICAL TRIALS: LEVERAGING INDUSTRY-ACADEMIC PARTNERSHIPS

### PANEL DISCUSSION: BAYESIAN METHODS IN CLINICAL TRIALS: LEVERAGING INDUSTRY-ACADEMIC PARTNERSHIPS

Panelists:

Amy Xia, Amgen Corporation  
Stacy Lindborg, Eli Lilly and Company  
Gary Rosner, MD Anderson Cancer Center  
Gene Pennello, Center for Devices and Radiological Health, U.S. Food and Drug Administration

This panel discussion will stress the ways academic research in this area can be modified to actually meet the needs of industry as they try to run Bayesian trials, either internally (safety/efficacy/dosing) or for FDA review (confirmatory).

email: carli002@umn.edu



**MONDAY, MARCH 16, 2009**

**10:30 am - 12:15**

**27. RECENT ADVANCEMENTS IN LONGITUDINAL ANALYSIS**

**JOINT MODELING OF LONGITUDINAL CATEGORICAL DATA AND SURVIVAL DATA**

Jianwen Cai\*, University of North Carolina at Chapel Hill  
Jaeun Choi, University of North Carolina at Chapel Hill  
Donglin Zeng, University of North Carolina at Chapel Hill

In many biomedical studies, it is of interest to study the covariate effect on both longitudinal categorical outcomes and survival outcomes. For example, in cancer research, it is of interest to study the treatment effect on both quality of life which is a categorical outcome measured longitudinally and survival time. In this talk, we will discuss such joint models. Random effects are introduced into the simultaneous models to account for dependence between longitudinal categorical outcome and survival time due to unobserved factors. EM algorithms are used to derive the point estimates for the parameters in the proposed model and profile likelihood function is used to estimate their variances. The asymptotic properties are established for our proposed estimators. Finally, simulation studies are conducted to examine the finite-sample properties of the proposed estimators and a liver transplantation data set is analyzed to illustrate our approaches.

email: cai@bios.unc.edu

**DROPOUT IN LONGITUDINAL CLINICAL TRIALS WITH BINARY OUTCOME**

Mike G. Kenward\*, London School of Hygiene and Tropical Medicine  
Rhian M. Daniel, London School of Hygiene and Tropical Medicine

There is a large literature on handling dropout in clinical trials, much concerned with continuous outcomes. Additional complications arise with a binary outcome because of the different forms that the treatment effect can take, and the importance of non-likelihood methods of analysis. In this talk we consider methods of analysis for making marginal treatment comparisons under dropout, and in the process compare and link a number of approaches to such analyses, including forms of weighted and unweighted estimating equations, multiple imputation, and likelihood. Relationships between underlying structure and bias under missing at random dropout mechanisms are explored under the different analyses. Some overall conclusions are drawn, particularly in terms of bias and precision of the treatment estimates and robustness of the resulting inferences.

email: mike.kenward@lshtm.ac.uk

**VARIABLE SELECTION IN LONGITUDINAL DATA USING REGULARIZED LIKELIHOODS**

Xihong Lin\*, Harvard School of Public Health

Variable selection becomes an increasingly important problem for longitudinal studies with high-dimensional genomic and proteomic

data, e.g., genome-wide association studies with longitudinal phenotypes. We propose variable selection methods using regularized likelihoods including the seamless L<sub>0</sub> penalized likelihood, for mixed models. The theoretical properties of these methods are studied and their finite sample performance is evaluated using simulations. The methods are illustrated using several data examples.

email: xlin@hsph.harvard.edu

**FUNCTIONAL LATENT FEATURE MODELS FOR DATA WITH LONGITUDINAL COVARIATE PROCESSES**

Erning Li\*, Texas A&M University  
Yehua Li, University of Georgia  
Nae-Yuh Wang, Johns Hopkins University School of Medicine  
Naisyin Wang, Texas A&M University

We consider a joint model approach to study the association between nonparametric latent features of longitudinal processes and a primary endpoint. Our modeling strategy is closely related to generalized functional linear models (GFLM), but has several marked differences. We argue that the key assumption in the common GFLM approach that the estimation variation in eigenfunctions is negligible is not necessarily true and is purely determined by the nature of data. We propose estimation procedures and supportive theory that allow the investigation regardless of the validity of this assumption. Our approach takes into account the estimation uncertainty embedded in the estimated eigen-system and allows users to have a thorough understanding of where the estimation uncertainty/variation lies so that the choice of a final model and future research plan can be made accordingly. To the best of our knowledge, the theoretical properties we have developed are the first that takes into account the uncertainty of the estimated eigen-components in the resulting parametric estimators and could be adopted by other estimators that use estimated eigenfunctions or eigenvalues. Numerical performances are evaluated in simulations and through a study of the impacts of BMI and SBP readings during adulthood on hypertension status later in life.

email: eli@stat.tamu.edu

**28. ADAPTIVE DESIGNS IN PRACTICE: BENEFITS, RISKS AND CHALLENGES**

**BAYESIAN ADAPTIVE DESIGNS IN MEDICAL DEVICE TRIALS**

Scott M. Berry, Berry Consultants

Adaptive designs--especially Bayesian adaptive designs--have been a hot topic. In this talk I focus on presenting actual designs and results of Bayesian adaptive designs. These features include adaptive sample size selection, dropping of treatment arms, and early success analyses. These features will be demonstrated through actual trial examples. The focus of the talk will be on the applications.

email: scott@berryconsultants.com

# ABSTRACTS

## IMPROVED DOSE RANGING THROUGH ADAPTIVE DOSE ALLOCATION

Judith A. Quinlan, Cytel Inc.

Dose ranging studies are internal decision making studies, and are an area of development where agencies have actively encouraged the use of adaptive designs. In this particular case study, we will see how given a fixed sample size, adaptive allocation out performs the traditional approach, and efficiently maximizes patient allocation to doses of interest. Focus will also be given to the role and benefits of simulations. Both in designing the clinical trial, but also importantly as an example of how simulations were effectively used to minimize drug supply requirements for this adaptive trial.

email: Judith.Quinlan@cytel.com

## DEVELOPING AN ADAPTIVE PHASE 2/3 DESIGN THROUGH TRIAL SIMULATION

Brenda L. Gaydos, Eli Lilly and Company

For some applications, a seamless design can reduce the time to submission while providing increased information at the time of filing. However, careful consideration is needed prior to selecting a seamless approach and finalizing the design. Comparisons to alternative clinical plans and trial designs should be performed. The use of trial simulation is essential throughout the development process. Initially, trial simulation is needed to provide a quantitative assessment of risk/benefit prior to deciding on a seamless approach, and after, to finalize design variants. But the role of trial simulation is not limited to internal decision making, trial documentation and implementation planning. For example, simulation output is needed to clearly communicate trial details to the data monitoring committee, ethical review boards and regulatory reviewers as well as to clinical trialists that will be conducting the study. In this presentation, a case-study will be used to illustrate the statistical challenges in designing and communicating a seamless 2/3 design with emphasis on the trial simulations used to determine the design and evaluate the Type I error rate. The application of trial simulation from internal decision making to regulatory review and implementation will also be described.

email: blg@lilly.com

## 29. OUTCOME DEPENDENT SAMPLING

### ON PLANNING A RETROSPECTIVE, OUTCOME DEPENDENT SAMPLING STUDY FOR LONGITUDINAL BINARY RESPONSE DATA

Jonathan S. Schildcrout\*, Vanderbilt University  
Patrick J. Heagerty, University of Washington

The wide availability of longitudinal data through ongoing cohort studies and other resources permits examination of any number of novel hypotheses. Often, we have all information available to conduct analyses except for a key exposure or confounding variable. If ascertainment costs are high, we must be judicious about who is sampled, and in such circumstances, outcome dependent sampling

designs permit efficient estimation. In this presentation, we introduce a class of designs that sample with probability related to whether or not subject-specific response variability was observed, and we propose maximum conditional likelihood for estimation and inference. However, the focus of the discussion will regard study planning. Estimation efficiency of these designs depends highly on the distribution of the target covariate. We will discuss this dependence and will compare the efficiency of various sampling strategies as a function of the target covariate distribution. We will also propose monte-carlo based power calculations that can be used to examine study feasibility using all available information prior to exposure ascertainment.

email: jonathan.schildcrout@vanderbilt.edu

## PARTIAL LINEAR MODEL FOR DATA FROM AN OUTCOME DEPENDENT SAMPLING DESIGN

Haibo Zhou\*, University of North Carolina at Chapel Hill  
Guoyou Qin, University of North Carolina at Chapel Hill

Outcome-dependent sampling (ODS) has been widely used in biomedical studies because it is a cost effective way to improve study efficiency. However, the models considered in the literature are limited to the framework of linear models due to the challenge in terms of both theory and computation. Partial linear model (PLM) is a powerful inference tool to nonparametrically model the relation between an outcome and exposure variables. In this article, we consider a partial linear model for data from an ODS design. We propose a semiparametric maximum likelihood method to achieve the inference of PLM. We develop the asymptotic properties and simulation studies show that the ODS design can produce more efficient estimate than the traditional simple random sampling design with the same sample size. We demonstrate the proposed method via a real data set analysis.

email: zhou@bios.unc.edu

## LONGITUDINAL STUDIES OF BINARY RESPONSE DATA FOLLOWING CASE-CONTROL AND STRATIFIED CASE-CONTROL SAMPLING: DESIGN AND ANALYSIS

Jonathan S. Schildcrout, Vanderbilt University School of Medicine  
Paul J. Rathouz\*, Department of Health Studies, University of Chicago

We discuss design and analysis of longitudinal studies after case-control sampling, wherein interest is in the relationship between a longitudinal binary response that is related to the sampling (case-control) variable, and a set of covariates. We propose a semiparametric modelling framework based on a longitudinal GEE response model and an ancillary model for subjects' case-control status. In this approach, the analyst must posit the population prevalence of being a case, which is then used to compute an offset term in the ancillary model. Parameter estimates from this model are used to compute offsets for the longitudinal response model. Examining the impact of population prevalence and ancillary model misspecification, we show that time-invariant covariate parameter estimates, other than

the intercept, are reasonably robust, but intercept and time-varying covariate parameter estimates can be sensitive to such misspecification. We study design and analysis issues impacting study efficiency, namely: choice of sampling variable and the strength of its relationship to the response, sample stratification, choice of working covariance weighting, and degree of flexibility of the ancillary model. The research is motivated by a longitudinal study following case-control sampling of the time course of ADHD symptoms.

email: prathouz@uchicago.edu

## THE ANALYSIS OF RETROSPECTIVE FAMILY STUDIES

John Neuhaus\*, University of California-San Francisco  
Alastair Scott, University of Auckland  
Chris Wild, University of Auckland

Genetic epidemiologists often augment an initial case-control sample with responses and covariates gathered from the family members of the initial sample persons (probands) to assess within-family covariate effects and measure familial aggregation (within-family dependence) of the response. Standard designs select both case and control probands and sample family members without regard to their response but investigators often use designs that restrict the proband and family samples to improve efficiency and reduce cost. Several different methods have been proposed to analyze data from these alternative designs but many of these methods do not use all available data or estimate all quantities of interest. This talk presents a profile likelihood approach that applies to a wide variety of case-control family designs and overcomes deficiencies in existing approaches. Using our general approach we can comprehensively assess features such as estimation efficiency of specific covariate effects obtained from alternative designs. We will illustrate our approach using data from several case-control family studies of cancer.

email: john@biostat.ucsf.edu

## 30. NEW STATISTICAL METHODS IN DETECTING EPISTASIS INTERACTIONS IN GENOME-WIDE ASSOCIATION STUDIES

### BAYESIAN DETECTION OF GENE-GENE INTERACTIONS ASSOCIATED WITH TYPE 1 DIABETES WITHIN MHC REGION

Yu Zhang, Penn State University  
Jing Zhang, Harvard University  
Jun S. Liu\*, Harvard University

We propose a Bayesian model for simultaneously inferring haplotype-blocks and selecting SNPs within blocks that are associated with the disease, either singly, or through epistatic interactions with other markers. Simulation results show that this approach is uniformly more powerful than our previous BEAM algorithm for epistasis mapping and other methods. We applied the method to WTCCC type 1 diabetes data, and discovered some interesting two-way interactions within the MHC region on chromosome 6. We found very strong interactions within and between (HLA-) DQ-DR region and its

upstream region from TNXB to BTNL2. Our results indicate that the LD structure in the MHC region may have changed substantially in cases compared with that in controls.

email: jliu@stat.harvard.edu

## GENOME-WIDE STRATEGIES FOR GENE-GENE INTERACTION

Dan L. Nicolae\*, University of Chicago

The development of cost-effective genotyping technologies has made genome-wide association studies the most popular tool for finding genes for complex traits. Although most of the findings come from single-marker analysis, there is a growing recognition that interactions (gene-gene and gene-environment) can play an important role in common disease etiology. The multiple comparison adjustments associated with whole genome studies are even more severe when interactions are investigated, making efficient statistical methods even more important. We discuss new methods for gene-gene interaction testing that are more efficient than classical approaches. The increase in power is achieved using several strategies including reasonable constraints in the two-marker parameter space, and testing prioritization based on the structure and dynamics of the human genome. A global association test that incorporates the evidence of marginal association and interactions will be also discussed.

email: nicolae@galton.uchicago.edu

## BAYESIAN ASSOCIATION MAPPING

Anders Albrechtsen, University of Copenhagen  
Rasmus Nielsen\*, University of California-Berkeley

For most common diseases with heritable components, not a single or a few single-nucleotide polymorphisms (SNPs) explain most of the variance for these disorders. Instead, much of the variance may be caused by interactions (epistasis) among multiple SNPs or interactions with environmental conditions. We present a new powerful statistical model for analyzing and interpreting genomic data that influence multifactorial phenotypic traits with a complex and likely polygenic inheritance. The new method is based on Markov chain Monte Carlo (MCMC) and allows for identification of sets of SNPs and environmental factors that when combined increase disease risk or change the distribution of a quantitative trait. Using simulations, we show that the MCMC method can detect disease association when multiple, interacting SNPs are present in the data. When applying the method on real large-scale data from a Danish population-based cohort, multiple interactions are identified that severely affect serum triglyceride levels in the study individuals. The method is designed for quantitative traits but can also be applied on qualitative traits. It is computationally feasible even for a large number of possible interactions and differs fundamentally from most previous approaches by entertaining nonlinear interactions and by directly addressing the multiple-testing problem.

email: rasmus\_nielsen@berkeley.edu

# ABSTRACTS

## 31. ANALYSIS OF MEDICAL COST DATA: JOINT VENTURE OF HEALTH ECONOMISTS AND STATISTICIANS

### A DECOMPOSITION OF CHANGES IN MEDICAL CARE EXPENDITURE DISTRIBUTION IN THE US HOUSEHOLDS: DO WE FARE BETTER TWENTY YEARS AFTER?

Ya-Chen T. Shih\*, University of Texas M.D. Anderson Cancer Center

**Introduction:** Medical expenditures have risen drastically in the US over the past two decades, so has the concern about inequalities in the allocation of healthcare resources. This study examines changes in the distribution of medical expenditures from 1987 to 2006 and identifies possible sources of variations. **Methods:** The study used the 1987 National Medical Expenditure Survey and the 2006 Medical Expenditure Panel Survey. Both data were nationally representative probability surveys on the financing and utilization of medical services for non-institutionalized individuals in the US. Quantile regression method was combined with Oaxaca's decomposition technique to disentangle factors contributing to the changes in the distribution of medical expenditure. **Results:** Findings suggested that disparities in medical expenditures between households at high and low socioeconomic status increased over time, after controlling for demographic and institutional characteristics. The observed discrepancies were largest at the lower percentiles of the distribution and narrowed at the higher percentiles, suggesting that low utilization of basic care among households at low socioeconomic status may have led to high medical expenditures associated with catastrophic health events.

email: yashih@mdanderson.org

### STOCHASTIC MODELS IN COST-EFFECTIVENESS ANALYSIS

Joseph C. Gardiner\*, Michigan State University  
Zhehui Luo, RTI International

Cost-effectiveness analysis (CEA) is a collection of techniques for structuring comparisons between competing interventions. It can inform decision-making by providing means for optimizing health benefits from a specified budget, or finding the lowest cost strategy for a specified health benefit. Markov processes are useful in modeling the dynamics of patient health outcomes as they unfold over time. States of the process represent health conditions or health states. We use a continuous-time finite-state Markov process to incorporate patient costs as they are incurred during sojourn in health states and in transition from one health state to another. By combining these expenditure streams, the net present value is the discounted expected total cost over a specified time period. Other metrics widely used in CEA such as net health benefit, net health cost and the cost-effectiveness ratio, and measures of health benefit such as life expectancy and quality-adjusted life years are defined as functions of expected values. We outline approaches to estimation of these summary statistics from health outcome and cost data that might be incompletely ascertained in some patients. Regression models are used

to incorporate patient-specific demographic and clinical characteristics and their impact on the metrics used in CEA can be assessed.

email: jgardiner@epi.msu.edu

## SEMI-PARAMETRIC MODELS FOR LONGITUDINAL COST DATA SUBJECT TO INCOMPLETE OBSERVATION

Eleanor M. Pullenayegum\*, McMaster University  
Andrew R. Willan, University of Toronto

The mean cost of a treatment strategy is an important consideration for policy-makers. Thus modelling the mean cost semi-parametrically is an attractive option to avoid the need to transform. Costs are often collected longitudinally, and estimation for semi-parametric mean regression models is typically via inverse-probability weighting to account for censoring in the cost data. Because inverse-probability weighting is inefficient, there has been much work on improving the efficiency of these estimators. Typically semi-parametric models for costs have stratified by time interval. Although efficiency has received much attention, modelling has often been overlooked. This talk will argue that careful modelling can lead to greater improvements in precision than complex estimation techniques.

email: pullena@mcmaster.ca

## A FLEXIBLE TWO-PART RANDOM EFFECTS MODEL FOR CORRELATED MEDICAL COSTS

Lei Liu\*, University of Virginia  
Mark Cowen, Quality Institute, St. Joseph Mercy Health System  
Robert Strawderman, Cornell University  
Tina Shih, M. D. Anderson Cancer Center, University of Texas

In this paper, we propose a 'two-part' random effects model (Olsen and Schafer 2001; Tooze, Grunwald, and Jones 2002) for correlated medical cost data. Typically, medical cost data are right-skewed, involve a substantial proportion of zero values, and may exhibit heteroscedasticity. The proposed model specification consists of two generalized linear mixed models, linked together by correlated random effects. Respectively, and conditionally on the random effects and covariates, we model the odds of cost being positive (Part I) and the mean cost (Part II) given that costs were actually incurred. The model is novel in that Part II assumes that observed costs follow a generalized gamma distribution with a scale parameter allowed to depend on covariates. The class of generalized gamma distributions is very flexible and includes the lognormal, gamma, inverse gamma and Weibull distributions as special cases. We demonstrate how to carry out estimation using the Gaussian quadrature techniques conveniently implemented in SAS Proc NLINMIXED. The proposed model is used to analyze pharmacy cost data on 56,245 adult patients clustered within 239 physicians in a midwestern U.S. managed care organization.

email: liulei@virginia.edu



## **32. GENETIC DIVERSITY, MUTATIONS AND NATURAL SELECTION**

### **DETECTING NATURAL SELECTION ACROSS DEPENDENT POPULATIONS**

Eleanne Solorzano\*, University of New Hampshire  
Hongyu Zhao, Yale University

The study of human genes is critical to the understanding of manifestations of diseases. Determining which genes may be under natural selection is an important step in discovering which genes may be associated with disease. Natural selection is a process that results in the survival and reproductive success of individuals or groups best adjusted to their environment, leading to the perpetuation of genetic qualities best suited to that particular environment. Sabeti's (2002) Extended Haplotype Homozygosity Test is usually used to detect recent positive selection of a particular gene across human populations. The standard way to make inferences across the populations is to assume independence using either a Bonferroni or FDR approach. However, human populations are correlated due to the fact that all humans originate from one common African ancestor. Therefore, to reduce bias, it is necessary to account for this correlation among populations. A new statistical method using haplotypes is developed for detecting natural selection across populations which accounts for such correlations. This test is shown to have higher statistical power than the existing methods to make inferences across populations and also controls the Type I error. The test is illustrated with an example using the lactase gene across 42 populations.

email: eleannes@cisunix.unh.edu

### **HIERARCHICAL BAYESIAN ANALYSIS OF GENETIC DIVERSITY IN GEOGRAPHICALLY STRUCTURED POPULATIONS**

Seongho Song\*, University of Cincinnati  
Dipak K. Dey, University of Connecticut  
Kent E. Holsinger, University of Connecticut

Populations may become differentiated from one another as a result of genetic drift. The amounts and patterns of differentiation at neutral loci are determined by local population sizes, migration rates among populations, and mutation rates. We proposed a hierarchical Bayesian model for inference of Wright's  $F_{ST}$ -statistics in a hierarchy in which we estimate the among-region correlation in allele frequencies by substituting replication across loci for replication across time based on the exact first two moments of a stochastic model for hierarchically structured populations subject to migration, mutation, and drift. As an application, microsatellite human data will be discussed.

email: seongho.song@uc.edu

## **DNA BARCODING: BAYESIAN DISCRETE ORDERED CLASSIFICATION**

Michael P. Anderson\*, Kansas State University  
Suzanne R. Dubnicka, Kansas State University

DNA barcodes are short strands of nucleotide bases taken from the cytochrome c oxidase subunit 1 (COI) of the mitochondrial DNA (mtDNA). Unlike nuclear DNA (nDNA), mtDNA remains largely unchanged as it is passed from mother to offspring, and it has been proposed that these barcodes may be used as a method of differentiating between biological species (Hebert 2003). While this proposal is sharply debated among some taxonomists (Will 2004), it has gained much momentum and attention from biologists. One issue at the heart of the controversy is the use of genetic distance measures as a tool for species differentiation. Current methods of species classification utilize these distance measures that are heavily dependent on both evolutionary model assumptions as well as a clearly defined 'gap' between intra- and interspecies variation (Meyer 2005). We point out the limitations of such distance measures and propose a character-based method of species classification which utilizes an application of Baye's rule to overcome these deficiencies. The proposed method is shown to provide accurate species-level classification as well as answers to important questions not addressable with current methods.

email: mpa@ksu.edu

## **JOINT BAYESIAN ESTIMATION OF PHYLOGENY AND SEQUENCE ALIGNMENT**

Heejung Shim\*, University of Wisconsin-Madison  
Bret Larget, University of Wisconsin-Madison

Phylogeny and sequence alignment are estimated separately in the traditional techniques : first estimate a multiple sequence alignment, and then infer a phylogeny based on the sequence alignment estimated in a previous step. We develop a joint model for co-estimating phylogeny and sequence alignment which avoids biased and exaggerated estimations from the traditional approach. Our indel model allows indel events to involve more than one letter and overlap each other. Since our method doesn't use a dynamic programming, we expect improvement in time complexity. We use a Bayesian approach using MCMC to estimate the posterior distribution of phylogenetic tree and a multiple sequence alignment.

email: shim@stat.wisc.edu

## **A STATISTICAL PERSPECTIVE OF DNA-PROTEIN CROSS-LINKS (DPX) DATA**

Martin Klein\*, University of Maryland, Baltimore County  
Bimal Sinha, University of Maryland, Baltimore County

It is generally believed that the DNA-protein cross-links (DPX) in the nasal mucosa of species is formed by formaldehyde inhalation. DPX data available from the nasal mucosa of rats and rhesus monkeys have been used as a measure of tissue dose in cancer risk assessments for formaldehyde. Using an appropriate modeling of the available DPX data observed on rats and rhesus monkeys, prediction of human nasal

# ABSTRACTS

mucosa DPX resulting from formaldehyde inhalation can be done based on a meaningful extrapolation using species-specific covariates. In this talk several statistical aspects of such an existing model (Hubal et al. (1997), Conolly et al. (2000)) are further explored and improved inference about the model parameters is drawn.

email: mklein1@umbc.edu

## CHARACTERIZATION OF mRNA SECONDARY STRUCTURE USING WEIBULL RANDOM VARIABLE

Fisseha Abebe\*, Clark Atlanta University  
William Seffens, Clark Atlanta University

Characterization of transcriptomes using Weibull random variable was first obtained by Cherkasov, Sui, Brunham, and Jones (2004). We have found that Weibull random variable with its general functional form also approximates, with very high accuracy, the distribution of mRNA folding free energy. The variation in the distribution between genomes is characterized by the Weibull reliability parameter <sup>2</sup>. The results of this work demonstrate that reliability analysis can be modified to provide useful insights to model a number of structural genomic studies.

email: fabebe@cau.edu

## COMPARING BAYESIAN AND FREQUENTIST APPROACHES TO ESTIMATING MUTATION RATES

Qi Zheng\*, School of Rural Public Health, Texas A&M Health Science Center

The concept of mutation lies at the heart of modern biology; measuring mutation rates is a major task in genetics research. The fluctuation experiment, pioneered by Luria and Delbrück in 1943, has been the method of choice for estimating microbial mutation rates. Recent years witnessed vigorous development of new statistical methods for estimating mutation rates in the context of the fluctuation experiment. These methods are strictly based on the likelihood principle. However, a Bayesian approach holds promise for some problems that are not tractable by a frequentist approach. This presentation is based on a pilot study. It first presents Bayesian solutions to some problems that have already been solved by a frequentist approach, it then suggests Bayesian solutions to some open problems to which no solutions exist.

email: qzheng@srph.tamhsc.edu

## 33. ESTIMATION METHODS

### WEIGHTED LIKELIHOOD METHOD FOR A LINEAR MIXED MODEL

Tianyue Zhou\*, Sanofi-aventis

Maximum likelihood is a widely used estimation method in statistics. This method is model dependent and as such is criticized as being non-robust. In this paper we consider using weighted likelihood

method to make robust inferences for linear mixed models where weights are determined at both the subject level and the observation level. This approach is appropriate for problems where maximum likelihood is the basic fitting technique, but a subset of data points is discrepant with the model. It allows us to reduce the impact of outliers without complicating the basic linear mixed model with normal distributed random effects and errors. The weighted likelihood estimators are shown to be robust and asymptotically normal. Our simulation study demonstrates that the weighted estimates are much better than the unweighted ones when a subset of data points is far away from the rest. Its application to the analysis of deglutition apnea duration in normal swallows shows that the differences between the weighted and unweighted estimates are due to large amount of outliers in the data set.

email: tianyue.zhou@sanofi-aventis.com

## ESTIMATOR OF THE INTENSITY OF A MODULATED POISSON PROCESS WITH A GAMMA PRIOR

Benjamin B. Neustifter\*, University of Georgia  
Stephen L. Rathbun, University of Georgia

The works of Rathbun, Shiffman, and Gwaltney (2007) and Waagepetersen (2008) on using modulated Poisson processes to model events based on time-varying covariates are extended to allow for variation among subjects in baseline rates. Estimating functions for covariate parameters are proposed and their large-sample properties are examined, with proof that the resulting estimators are consistent and asymptotically normal. The approach is illustrated using data from an ecological momentary assessment of smoking.

email: benn@uga.edu

## THE USE OF EXTREME ORDER STATISTICS TO ESTIMATE STANDARD DEVIATION

Chand K. Chauhan\*, Indiana Purdue University Fort Wayne Indiana  
Yvonne M. Zubovic, Indiana Purdue University Fort Wayne Indiana

For an unknown standard deviation,  $\bar{A}$ , we investigate an estimate  $S_p$ , which has a probability  $p$ , of being as high as or higher than  $\bar{A}$ , where the value of  $p$  may be selected by an experimenter. To avoid underestimating  $\bar{A}$ , one may pick  $p$  as high as or higher than 95%. In many situations only the highest value,  $Y_n$  and the lowest value,  $Y_1$ , of a sample may be available. Therefore in this paper we propose a linear combination,  $aY_n - bY_1$ , as an estimate. For symmetric distributions such as normal, it is common to divide a sample range by 4 to obtain an estimate of  $\bar{A}$ . However this may not result in a estimate with a high value of  $p$ . The values of  $a$  and  $b$  depend on many factors such as the parent distribution, the sample size used to find the extreme order statistics, the desired value of  $p$ , and other properties of the estimate. The authors discuss the choice of the linear combination,  $aY_n - bY_1$ , under different conditions. Analytical and simulated results of the performance of the proposed estimator are provided.

email: chauhan@ipfw.edu

## THE BIASED-BOOTSTRAP FOR GMM MODELS

Mihai C. Giurcanu\*, University of Louisiana at Lafayette  
Brett D. Presnell, University of Florida

In this talk, I present some theoretical and empirical properties of the uniform and biased-bootstrap distribution estimators for generalized method of moments (GMM) models. This version of the biased-bootstrap is a form of weighted bootstrap with weights chosen to satisfy some constraints imposed by the model. A typical biased-bootstrap resample is obtained by resampling from a member within a pseudo-parametric family of weighted empirical distributions on the sample. Because of its parametric nature, importance sampling can be successfully used when the biased-bootstrap is iterated, by re-weighting the first level bootstrap resamples. The resulting procedure yields an efficient and computationally feasible bootstrap recycling algorithm. I will present some consistency results of both the uniform and biased-bootstrap estimators of the distributions of GMM estimators and the J-test statistic. An application to the bootstrap calibrated confidence intervals shows some empirical results on the finite sample properties of the proposed method.

email: giurcanu@louisiana.edu

## BIOMEDICAL APPLICATIONS OF CONVOLUTIONS OF MIXED DISTRIBUTIONS

Calvin L. Williams, Clemson University  
Charity N. Watson\*, Clemson University

The concept of a convolution is statistical theory can be applied to many situations where the distributions being convolved are mixed densities and mass functions. This theory has been well established in the insurance industry (Lanzenauer and Lundberg(1974), and Brown (1977)). The idea of estimating the parameter(s) from convolution was discussed with some intrepidity by Sclove and Van Ryzin(1969), and Gong and Samaniego, (1981). Sprott (1984) tackled the issue of estimating parameters of convolutions. Moral (et. al, 2001) shows some interesting applications of mixed truncated exponentials whose convolutions show some interesting features. Here, we review these issues along with some interesting biostatistical applications and give some additional examples based on our own insight.

email: calvinw@ces.clemson.edu

## SIEVE TYPE DECONVOLUTION ESTIMATION OF MIXTURE DISTRIBUTIONS WITH BOUNDARY EFFECTS

Mihee Lee\*, University of North Carolina at Chapel Hill  
Peter Hall, The University of Melbourne  
Haipeng Shen, University of North Carolina at Chapel Hill  
Christina Burch, University of North Carolina at Chapel Hill  
Jon Tolle, University of North Carolina at Chapel Hill  
J. S. Marron, University of North Carolina at Chapel Hill

Density estimation in measurement error models has been widely studied. However, most existing methods consider only the case where the target distribution is continuous, hence they cannot be

applied directly to many practical problems, such as the estimation of mutation effects distribution in evolutionary biology, which is a mixture of a discrete atom and a continuous component. In this paper, we propose two sieve type estimators for distributions that are mixtures of a finite number of discrete atoms and continuous distributions under the framework of measurement error models. One major contribution of our paper is correct handling of known boundary effects. Compared to classical Fourier deconvolution, our estimators reduce the boundary problem at known non-smooth boundaries. In addition, the use of penalization improves the smoothness of the resulting estimator and reduces the estimation variance. We establish some asymptotic properties of the proposed estimators, and illustrate their performance via a simulation study and a real application in evolutionary biology.

email: miwing@email.unc.edu

## A METHOD FOR ACCELERATING THE QUADRATIC LOWER-BOUND ALGORITHM

Aiyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health  
Chunling Liu\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health  
Man Lai Tang, Hong Kong Baptist University  
Guo-liang Tian, University of Hong Kong

The quadratic lower-bound (QLB) algorithm is particularly suited to problems such as logistic, multinomial logistic, and Cox's proportional hazards models, where EM-type algorithms cannot be applied because of lacking a missing-data structure. However, to overcome slow convergence, we propose a novel "shrinkage parameter" approach to accelerate the QLB algorithm while maintaining its simple implementation and stable convergence. The uniformly optimal shrinkage parameter was derived. Two real data examples applying to logistic regression and Cox proportional hazard models separately are analyzed and compared to show that the accelerated QLB algorithm is dramatically twice as fast as the traditional QLB and to illustrate the proposed methods.

email: liuc3@mail.nih.gov

## 34. SPATIAL MODELS

### A KRONECKER PRODUCT LINEAR EXPONENT AR(1) FAMILY OF CORRELATION STRUCTURES

Sean L. Simpson\*, Wake Forest University School of Medicine  
Lloyd J. Edwards, University of North Carolina at Chapel Hill  
Keith E. Muller, University of Florida

The linear exponent autoregressive (LEAR) correlation structure, introduced by Simpson et al. (2009), is a flexible two-parameter correlation model that can be applied in situations in which the within subject correlation is believed to decrease exponentially in time or space. It allows for an attenuation or acceleration of the exponential decay rate imposed by the commonly used continuous-time AR(1) structure. We propose a Kronecker product LEAR correlation structure for multivariate repeated measures data in which the

# ABSTRACTS

correlation between measurements for a given subject is induced by two factors. The model allows for an imbalance in both dimensions across subjects. This four-parameter structure is especially attractive for the High Dimension, Low Sample Size cases so common in medical imaging and various kinds of “-omics” data. Excellent analytic and numerical properties make the Kronecker product LEAR model a valuable addition to the suite of parsimonious correlation structures for multivariate repeated measures data. We employ the model in the analysis of a longitudinal imaging data example.

email: slsimpso@wfubmc.edu

## ON THE VORONOI ESTIMATOR FOR INTENSITY OF AN INHOMOGENEOUS PLANAR POISSON PROCESS

Christopher D. Barr\*, Johns Hopkins University  
Frederic P. Schoenberg, University of California, Los Angeles

We investigate the statistical properties of the Voronoi estimator for the intensity of an inhomogeneous Poisson process. The Voronoi estimator may be defined for any location as the inverse of the area of the corresponding Voronoi cell. The estimator is well-known to be unbiased in the case of estimating the intensity of a homogeneous Poisson process, and is shown here to be approximately ratio unbiased in the inhomogeneous case. Simulation studies show the sampling distribution is well approximated by the inverse gamma model, extending similar results from the homogeneous case. Performance of the Voronoi estimator is compared to a kernel estimator using two simulated data sets as well as a dataset consisting of earthquakes within the continental United States.

email: cdbarr@gmail.com

## ZERO-INFLATED BINOMIAL SPATIAL MODELS, WITH APPLICATIONS TO COLON CARCINOGENESIS

Tatiyana V. Apanasovich\*, Thomas Jefferson University  
Marc G. Genton, Texas A&M University  
Raymond J. Carroll, Texas A&M University

When considering regression models for binomial spatial data in experimental colon carcinogenesis data, we have observed distinct regions where background rates of response abruptly become zero. These zero-inflated regions cause difficulties in, for example, low-order basis function modeling, because model fits attempt to reproduce the regions of zeros. We cast this general problem as one in which there are two spatial processes. The first is the underlying smooth regression surface, while the second is the spatial process that leads to regions of zeros. The methods are applied to an experiment involving aberrant crypt foci in colon carcinogenesis experiments.

email: Tatiyana.Apanasovich@jefferson.edu

## SPATIAL MODELING OF AIR POLLUTION AND MORTALITY TIME TRENDS IN THE UNITED STATES

Sonja Greven\*, Johns Hopkins University  
Francesca Dominici, Johns Hopkins University  
Scott Zeger, Johns Hopkins University

We are interested in the association between long-term exposure to particulate matter (PM) and mortality. In cross-sectional comparisons of mean pollution concentrations and mortality between cities, it is difficult to fully control for all potential confounding factors. We instead compare local trends in PM and mortality, with each location acting as its own control, thus minimizing confounding effects. Our data includes PM time series for seven years in 814 locations in the US, as well as individual level data on survival from a location-matched subset of the Medicare cohort. While a survival analysis approach reflects that pollution will likely affect longevity rather than overall mortality rates, the size of the data set with over 3 million deaths makes a direct implementation of a survival model impractical. We use an equivalent Poisson regression model, adjusting for location-specific hazard functions changing smoothly with age. We model potential spatial correlation in the data using penalized splines. To fit this complex model to the high-dimensional data, we develop a suitable backfitting algorithm.

email: sgrevens@jhsph.edu

## SPATIAL CLUSTER DETECTION FOR WEIGHTED OUTCOMES USING CUMULATIVE GEOGRAPHIC RESIDUALS

Andrea J. Cook\*, Group Health Center for Health Studies  
Yi Li, Harvard School of Public Health and The Dana Farber Cancer Institute  
David Arterburn, Group Health Center for Health Studies  
Ram C. Tiwari, CDR, U.S. Food and Drug Administration

Spatial cluster detection is an important methodology to robustly detect spatial clusters of outcomes without making strong model assumptions on the spatial dependence structure. For health outcomes, e.g. body mass index or obesity, the spatial dependence may be difficult to model since its magnitude may be dependent on measures that are difficult to quantify. This talk proposes a robust spatial cluster detection method for point or aggregate data for general outcomes, given the first two moments can be specified, including continuous, binary, and count data. This new method readily incorporates different weighting structures, such as the regional population, to allow the weighting of information on different regions to not be equal, which is key for aggregate data. The proposed method also incorporates the ability for covariate adjustment as there are no previous methods for weighted outcomes with covariate adjustment available. A simulation study is conducted to evaluate the performance of the method. The proposed method is then applied to assess spatial clustering of high Body Mass Index in a HMO population in the Seattle, Washington USA area.

email: cook.aj@ghc.org

## **ADJUSTMENTS FOR LOCAL MULTIPLICITY WITH SCAN STATISTICS**

Ronald E. Gangnon\*, University of Wisconsin

In most cluster detection problems, there are local differences in the extent of the multiplicity problem across the study region. For example, using a fixed maximum geographic radius for clusters, urban areas typically have many overlapping clusters, while rural areas have relatively few. The spatial scan statistic does not account for this local multiplicity problem. We describe two new spatially-varying multiplicity adjustments for spatial cluster detection, one based on a nested Bonferroni adjustment (LASS-B) and one based on a local Gumbel distribution approximation (LASS-G). The performance of the spatial scan statistic, LASS-B and LASS-G in terms of unbiased cluster detection under the null hypothesis is evaluated through simulation. These methods are then applied to both the well-known New York leukemia data and data on breast cancer incidence in Wisconsin.

email: ronald@biostat.wisc.edu

## **IMPROVING DISEASE SURVEILLANCE BY INCORPORATING RESIDENTIAL HISTORY**

Justin Manjourides\*, Harvard School of Public Health  
Marcello Pagano, Harvard School of Public Health

In environmental studies associated with the etiology of a human disease it is often of interest to determine whether there is any relationship between the location of an individual's abode and whether or not the person is diseased. We sometimes attempt to answer this question by studying the spatial distribution of individuals' addresses. But if the disease was triggered at some point in the past, then the current address is almost irrelevant. We propose that a subject's residential history be obtained for such studies and that it be incorporated in the analysis via the incubation distribution for the disease. In this talk we present a method for testing the difference between two spatial histories.

email: jmanjour@fas.harvard.edu

## **35. TOXICOLOGY/DOSE-RESPONSE MODELS**

### **ROBUST STATISTICAL THEORY AND METHODOLOGY FOR NONLINEAR MODELS WITH APPLICATION TO TOXICOLOGY**

Changwon Lim\*, University of North Carolina at Chapel Hill and Biostatistics Branch, NIEHS, NIH  
Pranab K. Sen, University of North Carolina at Chapel Hill  
Shyamal D. Peddada, Biostatistics Branch, NIEHS, NIH

Often toxicologists are interested in investigating the dose-response relationship when animals are exposed to varying doses of a chemical. In some instances a nonlinear regression model such as the Hill model is used to describe the relationship. The standard asymptotic confidence intervals and test procedures based on the ordinary least squares methodology may not be robust to heteroscedasticity and

consequently may produce inaccurate coverage probabilities and Type I error rates. On the other hand, the standard weighted least squares based methodology may be computationally intensive and may not be efficient when the variances are approximately equal across dose groups (homoscedasticity). In practice one generally does not know if the data are homoscedastic or heteroscedastic. Also neither method is robust against outliers or influential observations. Since the performance of a method depends on whether the data are homoscedastic or heteroscedastic, we introduce a simple preliminary test estimation (PTE) procedure that uses robust M-estimators. The methodology is illustrated using a data set obtained by the National Toxicology Program (NTP).

email: changwon@email.unc.edu

## **INCORPORATING HISTORICAL CONTROL INFORMATION INTO QUANTAL BIOASSAY WITH BAYESIAN APPROACH**

Din Chen\*, South Dakota State University

A Bayesian approach with an iterative reweighted least squares is used to incorporate historical control information into quantal bioassays to estimate the dose-response relationship, where the logit of the historical control responses are assumed to have a normal distribution. The parameters from this normal distribution are estimated from both empirical and full Bayesian approaches with a marginal likelihood function being approximated by Laplace's Method. A comparison is made using real data between estimates that include the historical control information and those that do not. It was found that the inclusion of the historical control information improves the efficiency of the estimators. In addition, this logit-normal formulation is compared with the traditional beta-binomial for its improvement in parameter estimates. Consequently the estimated dose-response relationship is used to formulate the point estimator and confidence bands for \$ED(100p)\$ for various values of risk rate \$p\$ and the potency for any dose level.

email: din.chen@sdstate.edu

## **A COMPARATIVE STUDY ON CONSTRUCTING CONFIDENCE BANDS FOR EFFECTIVE DOSES**

Gemechis D. Djira\*, South Dakota State University  
Din Chen, South Dakota State University

In this presentation we conduct a comparative study to construct simultaneous confidence bands for multi-dimensional effective doses in dose-response modeling with a binary response variable and multiple covariates. The methods based on the inversion of Scheffe simultaneous confidence intervals, the delta method and a bootstrap method will be compared through a simulation study.

email: gemechis.djira@sdstate.edu

# ABSTRACTS

## SEMIPARAMETRIC BAYES MULTIPLE TESTING: APPLICATIONS TO TUMOR DATA

Lianming Wang\*, University of South Carolina  
David B. Dunson, Duke University

In National Toxicology Program (NTP) studies, investigators want to assess whether a test agent is carcinogenic overall and specific to certain tumor types, while estimating the dose-response profiles. Because there are potentially correlations among the tumors, joint inference is preferred to separate univariate analysis for each tumor type. In this regard, we propose a random effect logistic model with a matrix of coefficients representing log-odds ratios for the adjacent dose groups for tumors at different sites. We propose appropriate nonparametric priors for these coefficients to characterize the correlations and to allow borrowing of information across different dose groups and tumor types. Global and local hypotheses can be easily evaluated by summarizing the output of a single MCMC chain. Two multiple testing procedures are applied for testing local hypotheses based on the posterior probabilities of local alternatives. Simulation studies are conducted and a NTP tumor data set is analyzed illustrating the proposed approach.

email: wangl@stat.sc.edu

## TESTING FOR SUFFICIENT SIMILARITY IN DOSE- RESPONSE IN COMPLEX CHEMICAL MIXTURES: DO INTERACTION AND DOSE SCALE MATTER?

LeAnna G. Stork\*, Monsanto Co.  
Scott L. Marshall, Virginia Commonwealth University  
Chris Gennings, Virginia Commonwealth University  
Linda K. Teuschler, National Center for Environmental Assessment, U.S. Environmental Protection Agency  
John Lipscomb, National Center for Environmental Assessment, U.S. Environmental Protection Agency  
Mike Devito, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency  
Kevin Crofton, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency

Chemical mixtures in the environment are often the result of a dynamic process. When dose-response data are available on random samples throughout the process, equivalence testing can be used to determine whether the mixtures are sufficiently similar in dose-response based on a pre-specified biologically important similarity region. Now consider a full mixture of  $c$  chemicals and suppose that  $s$  ( $s < c$ ) of them are not dose-responsive individually. It may be reasonable to assume (under the assumption of no interaction) that a subset mixture of only the dose-responsive chemicals (in the same relative proportions) should be sufficiently similar to the same mixture with the addition of the chemicals that are not dose-responsive. The total dose scale of the full mixture is adjusted to account for the proportion of the non-dose-responsive chemicals that are present. A nonlinear mixed-effects model is fit to both mixtures to account for the random variability in the total dose. Equivalence testing logic is then applied to test for sufficient similarity in dose-response. An example using five organophosphorous pesticides is demonstrated.

Partially supported by NIEHS #T32 ES007334 and #R01ES015276. Does not reflect USEPA policy and is not associated with Monsanto.

email: leanna.g.stork@monsanto.com

## INVESTIGATING STATISTICAL DISTANCE AS A SIMILARITY MEASURE FOR DETERMINING SUFFICIENT SIMILARITY IN DOSE-RESPONSE IN CHEMICAL MIXTURES

Scott Marshall\*, Virginia Commonwealth University  
Chris Gennings, Virginia Commonwealth University  
LeAnna G. Stork, Monsanto Co.  
Linda Teuschler, National Center for Environmental Assessment, U.S. Environmental Protection Agency  
Glenn Rice, National Center for Environmental Assessment, U.S. Environmental Protection Agency  
John Lipscomb, National Center for Environmental Assessment, U.S. Environmental Protection Agency  
Mike DeVito, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency  
Kevin Crofton, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency

Chemical mixtures in the environment are often the result of a dynamic process. For the purposes of risk assessment it is of interest to test whether these resulting candidate mixtures are sufficiently similar to a reference mixture when only the mixing ratios are available for the candidate mixtures. Using statistical equivalence testing logic and mixed model theory an approach has been developed, that extends the work of Stork et al (JABES,2008), to define sufficient similarity in dose-response for chemical mixtures containing the same chemicals with different ratios or a subset of chemicals. Total dose of the mixture can be adjusted for chemicals not in the subset. Four similarity measures based on combinations of adjusted total dose and weights based on potencies are described. A simulation study was conducted to assess the "power" of the approach. The current work estimated how often a resulting candidate mixture was sufficiently similar in dose-response to the reference mixture. (Partially supported by NIEHS #T32 ES007334 and # R01ES015276 and does not reflect USEPA policy. This research is not associated with Monsanto).

email: marshalls12@vcu.edu

## PROPORTION OF SIMILAR RESPONSE (PSR) AND RECEIVER OPERATING CHARACTERISTICS (ROC) METHODOLOGIES IN ASSESSING CORRELATES OF PROTECTION FOR VACCINE EFFICACY

Katherine E.D. Giacoletti\*, Merck Research Labs  
Joseph F. Heyse, Merck Research Labs

A question of interest in many vaccine clinical development programs is whether a level of antibody response can be determined that is considered a "protective level." Such a finding has important implications in decisions regarding booster programs, as well as in the interpretation of the persistence or waning of antibody levels in the years following vaccination. Traditionally, analyses to answer this question have been based on modelling the probability of developing

disease as a function of antibody level among vaccinated subjects. Such methods are often underpowered due to high vaccine efficacy resulting in very few cases to use for the model; furthermore, the models require many assumptions regarding the distribution of antibody responses. Two non-parametric approaches will be considered as alternative ways of addressing this question. These methods, PSR and ROC, have advantages over parametric statistical models in terms of interpretability and easy graphical representations and require few, if any, assumptions about the distributional properties of the antibody response data. An example based on a vaccine clinical trial will be presented, as well as more general applications beyond vaccine development programs.

email: katherine\_giacolletti@merck.com

## 36. CLASSIFICATION/MACHINE LEARNING

### CLASSIFICATION OF DATA UNDER AUTOREGRESSIVE CIRCULANT COVARIANCE STRUCTURE

Christopher L. Louden\*, University of Texas at San Antonio  
Anuradha Roy, University of Texas at San Antonio

The problem of classification is an old one that has application in environmental, geophysical, signal processing and in many other fields. There are numerous approaches to this problem using the statistical properties of the populations from which observations are drawn. In applications such as geophysical and signals processing there is a natural structure on the variance-covariance matrix of the observation vectors. The efficacy of classification is generally increased by taking that structure into account. One such structure that is used to model that variance-covariance matrix is the autoregressive circulant (ARC) structure. Classification rules have been developed for data that have an ARC covariance structure. The effectiveness of these rules has been shown by simulating data sets that have such ARC structure and comparing the error rates by using the rule that assumes an ARC structure, a compound symmetric (CS) structure and no structure. The results of these simulations show that the rule based on the correct structure has the lowest error rate and the rule based on the simple CS structure, in some cases, has a higher error rate than the rule based on no structure assumption.

email: clouden1@yahoo.com

### A SUPPORT VECTOR MACHINE APPROACH FOR GENOME-WIDE COPY-NUMBER-VARIATION STUDY

Shyh-Huei Chen\*, National Yunlin University of Science and Technology, Yunlin, Taiwan  
Fang-Chi Hsu, Wake Forest University School of Medicine

Extensions of genome-wide association studies (GWAS) to copy-number variations (CNV) have already suggested that CNV should be accounted for a crucial component of human phenotypic variations. Many CNV-discovery approaches are available in the literature; however, only few studies of CNV analysis associated with risk of common diseases have been reported because such association studies

rely on the discoveries of CNV-containing regions which are hard to be identified. By considering CNV as factors of resulting in human genetic disorders, a classification approach can be adopted to perform CNV association study. In this study, a support vector machine with a heuristics search approach is proposed. This method was applied to explore the association with prostate cancer risk in a case-control genome-wide CNV study, using the Affymetrix 5.0.

email: fhsu@wfubmc.edu

## ON MARGIN-BASED CLASSIFICATION METHODS

Lingsong Zhang\*, Harvard School of Public Health

SVM and DWD are two margin-based classification methods. In this paper, we investigate and integrate these two methods into a more general setting. Some theoretical and empirical properties are explored as well.

e-mail: zhang@hsph.harvard.edu

## GROUPED LASSO-PATTERNSEARCH ALGORITHM

Weiliang Shi\*, GlaxoSmithKline  
Grace Wahba, University of Wisconsin-Madison

The LASSO-Patternsearch algorithm (LPS) is an efficient method that can identify patterns of multiple dichotomous risk factors for outcomes of interest in genomic studies. The method is designed for the case where there is a possibly very large number of candidate patterns but it is believed that only a relatively small number are important. LPS, as a global method can handle very complicated correlations among the predictor variables. However, the number of variables is limited due to limited computer memory. The current trends in genetic epidemiology are to evaluate as many markers as possible. The number can easily exceed ten thousand or more. The Grouped LASSO-Patternsearch algorithm (GLPS) is proposed to tackle this problem. GLPS, as suggested by its name, begins with dividing the covariates into small groups. LPS is run on each group of variables. We collect all the variables that show up in at least one of these runs, either as main effects or in second order patterns. Then LPS is run again on all variables surviving the group step. One big advantage of our algorithm is that the runs in group steps are parallel. We can run them simultaneously on different machines. The computing system Condor provides us an excellent opportunity to do this job.

e-mail: weiliang.2.shi@gsk.com

## PERFORMANCE GUARANTEE FOR INDIVIDUALIZED TREATMENT RULES

Min Qian\*, University of Michigan  
Susan A. Murphy, University of Michigan

An individualized treatment rule is a decision rule that assigns treatment according to patient characteristics. Our goal is to estimate the individualized treatment rule that yields the maximal mean

# ABSTRACTS

response using data from a randomized trial. We consider algorithms based on prediction error minimization and derive an upper bound on the excess mean response of any decision rule in terms of its associated excess prediction error. This upper bound is sharp if a margin type condition holds. To ensure a good approximation and avoid overfitting, LASSO is used to minimize the prediction error. We study the finite sample properties of the LASSO estimator under mild conditions. Combining the above two results, we obtain an oracle inequality of the excess mean response of the estimated individualized treatment rule. Results from simulations and a real example are provided.

e-mail: minqian@umich.edu

## NONPARAMETRIC CLASSIFICATIONS OF TUMORS USING GENE EXPRESSION DATA BASED ON THE TRIANGLE DATA DEPTH

Zhenyu Liu\*, George Washington University  
Reza Modarres, George Washington University

A reliable and precise classification of tumors using gene expression data is essential for successful diagnosis and treatment of cancers. Gene expression data are difficult to analyze using classical multivariate analysis as such data contain a very large number of variables (genes) relative to the number of observations (tumor samples), presenting a ‘large p, small n’ challenge to classical discriminant analyses and classification methods that require  $n > p$ . Notion of data depth provides an alternative way of ordering multivariate observations in high dimensional space; hence, reducing the high dimensional problem to one dimensional problem. Recently, we proposed the triangle data depth and applied it to the classical two-sample testing of equality of distribution functions in high dimensions. The triangle data depth enjoys computational simplicity and efficiency in high dimensions as its time complexity is  $O(n^2)$ , which is independent of p. In this paper, we propose new nonparametric classification methods based on the triangle data depth for high dimensional settings such as gene expression observations. We will use real and simulated data to explore the properties of the proposed methods using the triangle depth and compare them with existing methods for the classification of tumors based on gene expression profiles.

e-mail: zliu@gwu.edu

## CLASSIFICATION OF SELF-MODELING REGRESSIONS WITH UNKNOWN SHAPE FUNCTIONS

Rhonda D. VanDyke\*, Cincinnati Children’s Hospital and Department of Pediatrics, University of Cincinnati  
Kert Viele, University of Kentucky  
Robin L. Cooper, University of Kentucky

Studies can present distinct, yet related functions. In certain settings, self-modeling regressions can be used to provide overall and smoothed individual estimates by combining information across functions. A set of self-modeling functions is defined by the entire set of functions being related through affine transformations of the x- and y-axes to a common function  $g(t)$ . We expand this definition to include the

possibility that a set of functions contains two underlying sets of self-modeling functions, the first related through a common shape  $g_1(t)$  and the second related through a separate shape function  $g_2(t)$ . We propose a method to take data consisting of a set of functions, estimate the two underlying shape functions, and to classify each function as belonging to either the first or second group of self-modeling functions. We estimate the underlying shape functions through Bayesian Adaptive Regression Splines (BARS). We illustrate the methodology through Synaptic Transmission data, where the functions measure electrical current across time and the two self-modeling groups of functions are hypothesized to result from different vesicles within synapses releasing transmitter in qualitatively different manners. The method is further assessed through simulations.

e-mail: rhonda.vandyke@cchmc.org

## 37. CLUSTERED SURVIVAL DATA

## MARGINAL MODELS FOR CLUSTERED TIME TO EVENT DATA WITH COMPETING RISKS USING PSEUDO-VALUES

Brent R. Logan\*, Medical College of Wisconsin  
Mei-Jie Zhang, Medical College of Wisconsin  
John P. Klein, Medical College of Wisconsin

Many time-to-event studies are complicated by the presence of competing risks and by nesting of individuals within a cluster, such as patients in the same center in a multi-center study. Several methods have been proposed for modeling the cumulative incidence function with independent observations. However, when subjects are clustered, one needs to account for the presence of a cluster effect either through frailty modeling of the hazard or subdistribution hazard, or by adjusting for the within-cluster correlation in a marginal model. We propose a method for modeling the marginal cumulative incidence function directly. We compute leave one out pseudo-observations from the cumulative incidence function at several time points. These are used in a generalized estimating equation to model the marginal cumulative incidence curve, and obtain consistent estimates of the model parameters. A sandwich variance estimate is derived to adjust for the within cluster correlation. The method is easy to implement using standard software once the pseudo-values are obtained, and is a generalization of several existing models. Simulation studies show that the method has good operating characteristics. We illustrate the method on a dataset looking at outcomes after bone marrow transplantation.

email: blogan@mcw.edu

## COMPETING RISKS REGRESSION FOR STRATIFIED DATA

Bingqing Zhou\*, University of North Carolina at Chapel Hill

For competing risks data, Fine and Gray (1999) proportional hazards model for subdistribution has gained popularity in its convenience in directly assessing the effect of covariates on the cumulative incidence function. However, in many important applications, proportional hazards may not be satisfied, including multicenter clinical trials,

where the baseline subdistribution hazards may not be common due to varying patient populations. In this article, we consider a stratified competing risks regression, to allow the baseline hazard to vary across levels of the stratification covariate. According to the relative size of the number of strata and strata sizes, two stratification regimes are considered. Using partial likelihood and weighting techniques, we obtain consistent estimators of regression parameters. The corresponding asymptotic distributions are provided for the two regimes separately, along with various estimation techniques. Data from a breast cancer clinical trial and from a bone marrow transplantation registry illustrate the potential utility of the stratified Fine-Gray model.

email: bzhou@bios.unc.edu

## STATISTICAL ANALYSIS OF CLUSTERED CURRENT STATUS DATA

Ping Chen\*, University of Missouri-Columbia  
 Junshan Shen, Beijing University-China  
 Jianguo Sun, University of Missouri-Columbia

This paper discusses regression analysis of clustered current status data which occur if the failure times of interest in a cluster survival study is either left- or right-censored, or each subject is observed only once. Some examples of the areas that often produce such data are cross-sectional studies and tumorigenicity experiments (Keiding, 1990; Sun, 2006). A few methods have been developed if the failure time of interest is only right-censored but there does not seem to exist a method for current status data. For the problem, we present a Cox frailty model and a two-step EM algorithm is developed for parameter estimation. A simulation study is conducted for the evaluation of the proposed methodology and indicates that the approach performs well for practical situations. An illustrative example from a tumorigenicity experiment is provided.

email: pcgv8@missouri.edu

## PARAMETRIC ANALYSIS OF INTERVAL CENSORED DATA WITH INFORMATIVE CLUSTER SIZE

Xinyan Zhang\*, University of Missouri  
 Jianguo Sun, University of Missouri

This paper considers the problem of parameter estimate for interval censored failure time data with informative cluster size. These types of data often arise in biomedical research settings. Ignoring the possible informativeness of cluster size could lead to a biased and misleading result. For right censored data, Cong et al. (2007) and Williamson et al. (2007) proposed weighted score function and within cluster resampling method. However, for this topic, very little research was done on interval censored data because of its complicated nature. In this project, we extend the idea and proposed weighted generalized estimating score equation method and within-cluster resampling approaches to interval censored data. Simulation studies are conducted for the evaluation of the presented approaches and demonstrate that the proposed methods produce unbiased parameter estimates. The proposed methods are also illustrated by application to a lymphatic filariasis survival data.

email: xzmpc@mizzou.edu

## MODELING SURVIVAL DATA WITH ALTERNATING STATES AND A CURE FRACTION USING FRAILTY MODELS

Yimei Li\*, University of Pennsylvania  
 Paul E. Wileyto, University of Pennsylvania  
 Daniel F. Heitjan, University of Pennsylvania

We introduce a cure model for survival data where a common frailty influences both the cure probability and the hazard function given not cured. Data generated from this model have a close-form likelihood, making it straightforward to obtain maximum likelihood estimates (MLEs). We then extend our model to data with multiple events and alternating states, using a Clayton copula to link two gamma frailties, one for each type of event. We illustrate the model with an analysis of data from two smoking cessation trials comparing bupropion and placebo, in which each subject potentially experienced a series of lapse and recovery events. The model suggests that bupropion increases the probability of being abstinent for good, and decreases the hazard of lapse. However, bupropion does not significantly influence the probability of abandoning the quit attempts, nor does it accelerate time to recovery significantly. The data also suggest a positive but not significant association between lapse and recovery. A simulation study suggests that the estimates have little bias and their 95% confidence intervals (CI) have nearly nominal coverage.

email: yimeili@mail.med.upenn.edu

## MODELS WITH MULTIPLE EVENT TYPES AND THEIR PREDICTIONS

Kent R. Bailey\*, Mayo Clinic

Longitudinal follow-up studies often record and analyze more than one event type, for example, myocardial infarction, stroke, and death, leading to multiple right-censored observations per individual. Methods have been proposed to treat these events simultaneously in a multivariate mode, for example, Wei Lin and Weissfeld (1989 JASA). We consider a simple procedure utilizing the proportional hazards framework, in which each event type is treated as a separate observation, leading to  $kN$  observations, where  $k$  is the number of event types, and  $N$  the number of observations. Stratification on event type allows fitting a rich variety of models with various degrees of linkage between the predictive models for the different events. We then develop a method whereby the resulting marginal models' predictions can be combined, by strategic use of the probit transformation, and construction of a multiply right-censored multinormal likelihood. We apply this to a large Percutaneous Coronary Intervention Database to illustrate the features of this approach.

email: baileyk@mayo.edu

## CONSTRAINED SURVIVAL ANALYSIS

Yong Seok Park\*, University of Michigan

When we study survival analysis, we often have prior knowledge about stochastically ordered survival distributions. For example, when we consider time to event of cancer patients from diagnosis of

# ABSTRACTS

cancer for different tumor stages, the survival probability in lower stage group are believed to be larger than that in higher stage group at any time. Researchers have proposed various approaches to restricted survival estimators. Brunk et al(1966) and Dykstra(1982) studied the constrained maximum likelihood estimators of two stochastically ordered distribution from uncensored and censored data. Rojo(1996) proposed and studied a new restricted survival estimators of two-sample case. El Barmi et al extended Rojo's estimators to the k-sample case assuming k populations with survival functions  $S_1 e S_2 e \dots e S_k$ ,  $e S_2$ . In this paper, we propose a criterion to extend Rojo's estimators under more general constraints than El Barmi et al, such as  $S_1 e S_2 e S_4$ ,  $S_1 e S_3 e S_4$ . We also propose an algorithm to obtain estimators under this criterion and study the properties of proposed estimators.

email: yongpark@umich.edu

## 38. MARGINS AND MONITORING OF NON-INFERIORITY CLINICAL TRIALS

### NON-INFERIORITY IN ORPHAN DISEASES: CAN WE IMPROVE UPON EXISTING THERAPIES?

Janet Wittes\*, Statistics Collaborative

The prototypical non-inferiority trial deals with common diseases for which ample sample size is available to define a clinically rational non-inferiority margin. For orphan diseases, especially diseases caused by genetic abnormalities, the population at risk may be very small. The small population leads to approval of drugs on the basis of a single small (under 50 patients) Phase 3 trial. If a new clinically attractive therapy becomes available, it must be compared to the approved therapy; however, if the new therapy is anticipated to be no more beneficial than the approved therapy, a non-inferiority margin is necessary. In such a case, setting the non-inferiority margin becomes very difficult. This talk addresses the problems setting a non-inferiority margin in the context of orphan diseases with Gaucher disease as a specific example.

email: janet@statcollab.com

### NON-INFERIORITY MARGIN IN THE PRESENCE OF CONSTANCY VIOLATION OR DIFFERENT PATIENT POPULATIONS

Sue-Jane Wang\*, U.S. Food and Drug Administration  
H.M. James Hung, U.S. Food and Drug Administration  
Robert T. O'Neill, U.S. Food and Drug Administration

Fixed margin method and synthesis method have been proposed for establishing the efficacy of an experimental therapy through a non-inferiority analysis of a two-arm active controlled trial. Either approach can be very sensitive to the validity of the two critical assumptions: assay sensitivity and constancy of the control effect, although the study objectives may differ. When historical data are available, the current practice in defining the margin generally uses the worst 95% confidence limit, say, of the control effect, and considers preservation of some fraction of the control effect obtained from the historical placebo controlled trials. In this presentation, we consider a statistical approach that predefines the margin anticipating

the presence of constancy assumption violation or different patient populations. The consideration is predicated on anticipated unavoidable differences in trial design due to ethical constraints to performing a placebo controlled trial. In such cases, the implications of a closely matched active controlled trial would call into question, and, it may not be possible to formulate a non-inferiority margin based on the frequently used method described above.

email: suejane.wang@fda.hhs.gov

## MONITORING NON-INFERIORITY TRIALS WITH RECURRENT EVENT OUTCOMES OVER MULTIPLE TREATMENT PERIODS

Richard Cook\*, University of Waterloo  
Grace Yi, University of Waterloo

We describe methods for monitoring non-inferiority trials with recurrent event responses designed to compare an experimental treatment to standard care. Two types of designs are considered under the framework of mixed Poisson processes, but robust methods are developed and advocated at the analysis stage. The first design is a standard parallel group design in which patients receive either the experimental treatment or the control. The second design is a two-period study in which patients receive one treatment in the first period and the alternative treatment in the second period. Semiparametric methods for estimating and testing treatment effects are derived. The robustness and empirical power of the semiparametric approaches are assessed through simulation. Data from a two period cross-over trial of patients with bronchial asthma are analysed for illustration.

email: rjcook@uwaterloo.ca

## 39. STATISTICAL ANALYSIS OF INFORMATIVE MISSING DATA

### EVERY MISSING NOT AT RANDOM MODEL FOR INCOMPLETE DATA HAS GOT A MISSING AT RANDOM COUNTERPART WITH EQUAL FIT

Geert Molenberghs\*, Universiteit Hasselt and Katholieke Universiteit Leuven  
Michael G. Kenward, London School of Hygiene and Tropical Medicine, United Kingdom  
Geert Verbeke, Universiteit Hasselt and Katholieke Universiteit Leuven  
Caroline Beunckens, Universiteit Hasselt and Katholieke Universiteit Leuven  
Cristina Sotto, Universiteit Hasselt and Katholieke Universiteit Leuven

Many models for incomplete data allow for MNAR missingness. Sensitivity to unverifiable modeling assumptions, has initiated a lot of work. A key issue is that an MNAR model is not fully verifiable from the data, rendering the empirical distinction between MNAR and MAR next to impossible, unless one is prepared to accept the posited MNAR model in an unquestioning way. We show that empirical distinction between MAR and MNAR is not possible, since each MNAR model corresponds to exactly one MAR counterpart. Such a

pair will produce different predictions of the unobserved outcomes, given the observed ones. This is true for selection, pattern-mixture, and shared-parameter models. We will focus on the latter. Theoretical considerations are supplemented with illustrations based on a clinical trial in onychomycosis and on the Slovenian Public Opinion survey. The implications for sensitivity analysis are discussed. Missing data can be seen as latent variables. Such a view allows extension of our results to other forms of coarsening, such as grouping and censoring. In addition, the technology applies to random effects models, where a parametric form for the random effects can be replaced by certain other parametric (and non-parametric) form, without distorting the model's fit, latent classes, latent variables, etc.

email: geert.molenberghs@uhasselt.be

## BAYESIAN SEMIPARAMETRIC SELECTION MODELS WITH APPLICATION TO A BREAST CANCER PREVENTION TRIAL

Chenguang Wang, University of Florida  
 Michael Daniels\*, University of Florida  
 Daniel Scharfstein, Johns Hopkins University

We consider inference for the treatment difference of outcomes from longitudinal cancer studies, in which repeatedly measured outcomes may be informatively missing due to drop out (withdraw of consent or loss to follow-up) or protocol-defined events (progression or death). It is known that in the presence of informative missingness, the treatment difference is not identifiable unless unverifiable assumptions are made. We posit a high dimensional semiparametric selection model that uses a semiparametric missing data mechanism (MDM) and a nonparametric (saturated) model for the full-data response distribution. This model is useful for conducting sensitivity analysis and incorporating expert opinion about the missingness. We propose reducing the dimensionality by introducing shrinkage priors for the high order interactions in the full-data response and MDM models. We explore the degree of identification of potential sensitivity parameters by the data and develop a new way to elicit response-specific prior distributions using expert opinions. This modeling approach is applied to an analysis of data from the Breast Cancer Prevention Trial (BCPT).

email: mdaniels@stat.ufl.edu

## CONSTRUCTING AND CALIBRATING INFORMATIVE PRIORS FOR NONIDENTIFIED PARAMETERS IN MODELS FIT TO INCOMPLETE DATA

Joseph W. Hogan\*, Brown University

Consider the problem of estimating a parameter of interest from incomplete data (e.g. population mean, odds ratio, regression coefficient). Two common inferential summaries are bounds and sensitivity analyses. Bounds are attractive because they convey lack of information about the parameter (i.e. a point estimate is not available); sensitivity analyses are useful because they can be used to depict possible inferences across a range of plausible assumptions. A relatively underutilized approach is the use of informative priors

under a Bayesian formulation of the inferential problem. We illustrate with several examples how to encode untestable assumptions with prior distributions, and how to draw sensible posterior inferences. Issues addressed include elicitation and formulation of priors using expert opinion, dynamic updating of priors for longitudinal data, and calibration of ranges for priors and sensitivity analyses.

email: jhogan@stat.brown.edu

## 40. MODEL-BASED CLUSTERING OF HIGH-DIMENSIONAL GENOMIC DATA

### IDENTIFYING CLUSTER STRUCTURE AND RELEVANT VARIABLES IN HIGH-DIMENSIONAL DATA SETS

Mahlet G. Tadesse\*, Georgetown University

In the analysis of high-dimensional data there is often interest in uncovering cluster structure and identifying discriminating variables. For example, the goal may be to use gene expression data to discover new disease subtypes and to determine genes with different expression levels between classes. Another research question that is receiving increased attention is the problem of relating genomic data sets from various sources. For instance, the goal may be to identify subsets of DNA sequence variations from SNP arrays or array CGH that are associated with changes in mRNA transcript abundance in a set of correlated genes. We have proposed mixture models with variable selection to address these problems. I will present the methods and illustrate their applications on various genomic data sets.

email: mgt26@georgetown.edu

## RECURSIVELY PARTITIONED MIXTURE MODELS WITH APPLICATIONS TO DNA METHYLATION ARRAY DATA

E. Andres Houseman\*, Brown University  
 Brock C. Christensen, Brown University  
 Ru-Fang Yeh, University of California San Francisco  
 Carmen J. Marsit, Brown University  
 Margaret R. Karagas, Dartmouth-Hitchcock Medical Center  
 Margaret Wrensch, University of California San Francisco  
 Heather H. Nelson, University of Minnesota School of Public Health  
 Joseph Wiemels, University of California San Francisco  
 John K. Wiencke, University of California San Francisco  
 Karl T. Kelsey, Brown University

Epigenetics is the study of heritable changes in gene function that cannot be explained by changes in DNA sequence, and is one of the modes by which environment is thought to interact with genetics. As such, epigenetic variables constitute an important type of genome-scale variable mediating between environment and disease, and present unique statistical challenges. One of the most commonly studied epigenetic alterations is cytosine methylation, which is a well recognized mechanism of gene silencing that often becomes dysregulated in cancer. Arrays are now being used to study DNA methylation at a large number of loci; for example, the Illumina GoldenGate platform assesses DNA methylation at 1505 loci, while

# ABSTRACTS

the Infinium platform measures DNA methylation at over 27,000 sites. Mixture models have been used to identify DNA methylation subgroups in low-dimensional data sets, but few methods exist for clustering high-dimensional data in a reliable and computationally efficient manner. We present a novel model-based recursive-partitioning algorithm to navigate clusters in a mixture model, and propose methods for inferring associations between resulting nested “methylator phenotypes” and variables (e.g. upstream environmental variables and downstream disease phenotypes and clinical outcomes). We demonstrate our method on normal tissues and various types of tumors.

email: E\_Andres\_Houseman@brown.edu

## A LATENT CLASS MODEL WITH HIDDEN MARKOV DEPENDENCE FOR ARRAY CGH DATA

Stacia M. DeSantis\*, Medical University of South Carolina  
E. Andres Houseman, Brown University  
Brent A. Coull, Harvard School of Public Health  
David N. Louis, Massachusetts General Hospital  
MA 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 supervised Bayesian latent class approach for classification that relies on a hidden Markov model to account for the dependence in the intensity ratios. Supervision means that classification is guided by a clinical endpoint. 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, and differentiates classes by survival much better than unsupervised methods.

email: desantis@musc.edu

## TRANSPOSABLE REGULARIZED COVARIANCE MODELS WITH AN APPLICATION TO HIGH-DIMENSIONAL MISSING DATA IMPUTATION

Genevera Allen\*, Stanford University  
Rob Tibshirani, Stanford University

Missing data is a common concern in high-dimensional problems such as microarrays and user-ratings data. Recent authors have suggested that these examples of matrix data do not necessarily have independent rows or columns and it is not clear whether the row or columns are features. Hence, the data is transposable. To model this, we present a modification of the matrix-variate normal, the mean-restricted matrix-variate normal, in which the rows and columns each have a separate mean vector and covariance matrix. We extend regularized covariance models, which place an additive penalty on the inverse covariance matrix, to this distribution by placing separate

penalties on the covariance matrices of the rows and the columns. These so called transposable regularized covariance models allow for maximum likelihood estimation of the mean and non-singular covariance matrices. Using these models, we formulate EM-type algorithms for missing data imputation in both the multivariate and transposable frameworks. An efficient approximation is also given for transposable imputation. Simulations and results on microarray data and the Netflix data show that these imputation techniques often outperform existing methods and offer a greater degree of flexibility.

email: tibs@stanford.edu

## 41. ISSUES IN COMPLICATED DESIGNS AND SURVIVAL ANALYSIS

### STATISTICAL IDENTIFIABILITY AND THE SURROGATE ENDPOINT PROBLEM WITH APPLICATION TO VACCINE TRIALS

Julian Wolfson\*, University of Washington  
Peter Gilbert, University of Washington and Fred Hutchinson Cancer Research Center

Given a treatment  $Z$ , a clinical outcome  $Y$ , and a biomarker  $S$  measured some time after  $Z$  is administered, we may be interested in knowing whether  $S$  can be used in place of  $Y$  for establishing the effect of  $Z$ . This is the set-up for the surrogate endpoint problem. Several recent proposals for the statistical assessment of surrogate value have been based on the framework of potential outcomes and counterfactuals. The major challenge posed by these approaches is that the resulting estimands, which we refer to as surrogate risks, involve distributions which may not be statistically identifiable from observed data. We describe the statistical identifiability of surrogate risks in the context of vaccine trials, and show how different sets of assumptions affect their identifiability. We then suggest related estimands, which we call predictive risks, which do not quantify surrogate value but nonetheless provide valuable clinical information. A clever data collection scheme (closeout vaccination) and an accompanying set of reasonable assumptions allow predictive risks to be identified; we discuss this scheme and suggest extensions which may permit analogous estimands to be identified in a time-to-event setup. Based on algebraic relationships between surrogate and predictive risks, we propose a sensitivity analysis for assessing surrogate value and show that, in some cases, sensitivity analyses may have little or no power to detect surrogate value, even when the sample size is very large.

email: julianw@u.washington.edu

### MULTIPHASE CASE-CONTROL SAMPLING DESIGNS

Bryan Langholz\*, University of Southern California  
Ly Thomas, University of Southern California  
Rakovski Cyril, Chapman University

Multiphase case-control designs can be useful when there is a relatively inexpensive correlate of the variable of interest. With a focus individually matched case-control studies, a direct link between survey sampling and case-control designs will be described and it is

shown that the finite population properties of the conditional logistic likelihood score function are precisely those corresponding survey sampling Lahiri-Midzuno-Sen ratio estimator. A multiphase nested case-control study of endometrial hyperplasia and endometrial cancer will be described to illustrate the methods.

email: langholz@usc.edu

## SEMIPARAMETRIC EFFICIENT ESTIMATION IN CASE-COHORT STUDY

Donglin Zeng\*, University of North Carolina  
Danyu Lin, University of North Carolina

In case-cohort study, some important but expensive risk covariates for failure time are only measured in a subsample randomly selected from the full cohort. Although many methods have been developed to estimate risk effects in the case-cohort design, most of methods restrict to considering the proportional hazards model. Furthermore, no estimation is semiparametrically efficient especially when the expensive covariates depend on the other confounding variables. In this work, we consider estimating the risk factors in general transformation models. We propose an efficient approach by maximizing a modified likelihood function via the expectation-maximization algorithm. Particularly, the nuisance parameter of conditional densities among covariates is obtained via maximizing a local likelihood function in the M-step. We derive the asymptotic results for the obtained estimators and show that they are consistent and asymptotically efficient. The small-sample performance of the proposed method is illustrated via extensive numerical studies and applications to real data.

email: dzeng@bios.unc.edu

## ESTIMATING THE EFFECT OF A TIME-DEPENDENT THERAPY ON RESTRICTED MEAN LIFETIME USING OBSERVATIONAL DATA

Douglas E. Schaubel\*, University of Michigan  
John D. Kalbfleisch, University of Michigan

In observational studies of survival data, the treatment of interest may be time-dependent. The evaluation of a time-dependent treatment is often attempted through a hazard regression model which features a time-dependent treatment indicator. However, depending on the nature of the data structure, this approach may not be applicable. We develop semiparametric procedures to estimate the effect on restricted mean lifetime of a time-dependent treatment in the presence of the following complicating factors: both an experimental and established form of treatment are available; pre- and post-treatment hazards are non-proportional; subjects may experience periods of treatment ineligibility; treatment assignment is not randomized. The proposed methods involve weighting results from stratified proportional hazards models fitted using a generalization of case-cohort sampling. Asymptotic properties of the proposed estimators are derived, with finite sample properties assessed through simulation. The proposed methods are applied to data from the Scientific Registry of Transplant Recipients (SRTR), to quantify the effect on patient survival of expanded criteria donor (ECD) kidney transplantation.

email: deschau@umich.edu

## 42. STATISTICAL INFERENCE FOR FOREST INVENTORY AND MONITORING USING REMOTELY SENSED DATA

### HIERARCHICAL SPATIAL MODELS WITH REMOTELY SENSED PREDICTORS FOR MAPPING TREE SPECIES ASSEMBLAGES ACROSS LARGE DOMAINS

Andrew O. Finley\*, Michigan State University  
Sudipto Banerjee, University of Minnesota  
Ronald E. McRoberts, Northern Research Station, U.S. Forest Service

Spatially explicit data layers of tree species assemblages, referred to as forest types or forest type groups, are a key component in large-scale assessments of forest sustainability, biodiversity, timber biomass, carbon sinks, and forest health monitoring. This talk explores the utility of coupling georeferenced national forest inventory (NFI) data with readily available and spatially complete environmental and remotely sensed predictor variables through multinomial probit regression with varying levels of random spatial effects to predict forest type groups across large forested landscapes. Hierarchical models with spatially varying coefficients that exploit the spatial proximity of the NFI plot array and non-stationarity of predictor variables are proposed to improve the accuracy and precision of forest type group classification at locations where we have observed predictors but not inventory plots. The richness of these models, however, incurs onerous computational burdens and we need to resort dimension reducing spatial processes without sacrificing the richness in modeling. We illustrate using NFI data from Michigan, USA, where we provide a comprehensive analysis of this large study area and demonstrate improved classification with associated measures of uncertainty.

email: finleya@msu.edu

### A COMBINED DESIGN AND MODEL-BASED DERIVATION OF THE MSE OF ESTIMATED ABOVEGROUND BIOMASS FROM PROFILING AIRBORNE LASER SYSTEM

Timothy G. Gregoire\*, Yale University  
Ross F. Nelson, NASA-Goddard Space Flight Center  
Erik Naesset, Norwegian University of Life Sciences  
Goran Stahl, Swedish University of Agricultural Sciences  
Terje Gobakken, Norwegian University of Life Sciences

Profiling airborne laser altimetry has been used to discern a height profile of forests over an extensive area. When coupled with ground sampling within a double sampling framework, an estimator of aboveground forest biomass is obtained. The estimator is affected by multiple sources of statistical error stemming not only from the design but also from a reliance on statistical models to predict individual tree biomass and to link the remotely sensed data to the ground sample. Accounting for these errors when deriving the design-cum-model based mean square error presents some intriguing challenges which are the focus of this presentation.

email: timothy.gregoire@yale.edu

# ABSTRACTS

## MODEL-BASED INFERENCE FOR NATURAL RESOURCE INVENTORIES

Ronald E. McRoberts\*, Northern Research Station, U.S. Forest Service

Natural resources inventory programs have responded to the traditional user question ‘How much?’ using sample data and probability-based inference. Increasingly, users also are asking ‘Where?’ and request maps depicting the spatial distributions of resources and estimates compatible with the maps. For small area estimation and/or for complex relationships between response and ancillary variables, probability-based inference may be unsuitable. For these situations, model-based inference that relies on inventory sample plot data and ancillary satellite image data has been found to be useful. Although maps consisting of model-predictions may be relatively easy to construct, inference compatible with the maps may be much more difficult. First, whereas probability-based estimates or population parameters are known to be at least asymptotically unbiased, such is not the case for model-based estimates. Thus, bias assessment becomes crucial for model-based approaches to inference. Second, computational intensity for even small areas may become prohibitive because of the large number of pixels and necessity of considering correlations among pixel predictions. The presentation focuses on practical solutions to both problems, bias assessment and computational intensity.

email: rmcroberts@fs.fed.us

## 43. PRE-PROCESSING AND QUALITY CONTROL FOR HIGH-THROUGHPUT GENOMIC TECHNOLOGIES

### BACKGROUND CORRECTION BASED ON THE BOX-COX TRANSFORMATION OF NOISES FOR ILLUMINA BEAD ARRAY DATA

Min Chen\*, Yale University  
Yang Xie, University of Texas Southwestern Medical Center

Abstract Illumina bead array is a relatively new technology and becomes increasingly popular due to many attractive features. One distinction with other platforms is that it has negative control beads containing arbitrary oligonucleotide sequences that are not specific to any target genes in the genome. This design provides a way of directly estimating the distribution of the background noise. In the literature of background correction, the noise is often assumed to be normal. However, we show with real data that the noise can be very skewed, and the correction methods based on the normality assumption can lead to biased gene expression intensities. In this study we propose a noise adjustment method based on a model with a Box-Cox transformation on the noise term. To facilitate the search of MLE estimators, a spline technique is applied to approximate the likelihood function, and we show it can greatly improve the performance of the searching algorithm.

email: min.chen@yale.edu

## BACKGROUND ADJUSTMENT FOR DNA MICROARRAYS USING A DATABASE OF MICROARRAY EXPERIMENTS

Yunxia Sui\*, Brown University  
Zhijin Wu, Brown University  
Xiaoyue Zhao, Bionovo Inc.

Microarrays has become an indispensable technique in biomedical research. The raw measurements from microarrays undergo a number of preprocessing steps before the data are converted to genomic level for further analysis. Background adjustment is an important step in preprocessing. Estimating background noise has been challenging because of limited replication of microarray experiments. Most current methods have used the empirical Bayes approach to borrow information across probes in the same array. These approaches shrink the background estimate for either the entire sample or probes sharing similar sequence structures. In this article we present a probe specific solution in estimating background noise using a database of large number of microarray experiments. Information is borrowed across samples and background noise is estimated for each probe individually. The ability to prove truly probe specific background distribution allows us to extend the dynamic range of gene expression levels. We illustrate the improvement in detecting gene expression variation on two datasets: a Latin Square spike-in experiment publicly available and an Estrogen Receptor experiment with biological replicates.

email: ysui@stat.brown.edu

## STATISTICAL METRICS FOR QUALITY ASSESSMENT OF HIGH DENSITY TILING ARRAY DATA

Hui Tang\*, Mayo Clinic  
Terence Speed, University of California at Berkeley

High density tiling arrays are designed to blanket an entire genomic region of interest using tiled oligonucleotides at very high resolution and are widely used in ChIP-chip, DNA methylation, comparative genomic hybridization and transcriptome mapping studies. Experiments are usually conducted in multiple stages, including chromatin sonication, amplification of DNA fragment mixtures, labeling and hybridizing them onto tiling arrays, in which unwanted variations maybe introduced. As high density tiling arrays become more popular and are adopted by many research labs, it is pressing to develop quality control tools as what people did in DNA expression microarrays. To this aim, we propose a set of statistical quality metrics analogous to those in expression microarray studies with the application in high density tiling array data. We also developed a procedure to estimate the significance of the proposed metrics by using randomization test. Our method is applied to multiple real data sets, including three independent ChIP-chip experiments and one transcriptome mapping study. It showed sensitivity in capturing tiling arrays with high noise levels and filled in a gap in the literature for this topic.

email: h.tang@mayo.edu

## THE EFFECTS OF MISSING IMPUTATION ON VARIOUS DOWN-STREAM ANALYSES IN MICROARRAY EXPERIMENTS

Sunghee Oh\*, University of Pittsburgh  
George C. Tseng, University of Pittsburgh

Amongst the high-throughput technologies, DNA microarray experiments provide enormous quantity of genes and arrays with biological information to disease. Despite advances and the popular usage of microarray, the microarray experiments frequently produce multiple missing values due to many flaw factors. Thus, gene expression data contains some missing entries and a large number of genes may be affected. Many downstream algorithms for gene expression analysis require a complete matrix as an input. For now, there exists no uniformly superior imputation method and the performance depends on the structure and nature of data set. In addition, imputation methods have been mostly compared in terms of variants of RMSEs (Root Mean Squared Error) which compare true expression values to imputed values. The drawback of RMSE-based evaluation is that the measure does not reflect the true biological effect in down-stream analyses. In this study, we investigate how missing value imputation process affects the biological results of differentially expressed genes discovery, clustering and classification. Quantitative measures reflecting the true biological effects in each down-stream analysis will be used to evaluate imputation methods and compared to RMSE-based evaluation.

email: sshshoh1105@gmail.com

## A NOVEL TEST FOR QUALITY CONTROL IN FAMILY-BASED GENOME-WIDE ASSOCIATION STUDIES

David Fardo\*, University of Kentucky  
Iuliana Ionita-Laza, Harvard School of Public Health  
Christoph Lange, Harvard School of Public Health

Allele transmissions in family pedigrees provide an intuitive and unique way to evaluate the genotyping quality of a particular proband in a family-based association study. We propose a transmission test that is based on this feature and that can be used for quality control filtering of genome-wide genotype data for individual probands. The test has one degree of freedom and assesses the average genotyping error rate of the genotyped SNPs for a particular proband. As we show in simulation studies, the test is sufficiently powered to identify probands with unreliable genotyping quality that cannot be detected with standard quality control filters. This feature of the test is further exemplified by an application to a genome-wide association study. The test seems to be ideally suited as the final layer of quality control filters in the cleaning process of genome-wide association studies since it is able to identify probands with poor genotyping quality who have slipped through the standard quality control filtering.

email: david.fardo@uky.edu

## STATISTICAL INFERENCE FOR POOLED SAMPLES IN NEXT GENERATION SEQUENCING

Justin W. Davis\*, University of Missouri

In the last few years, there have been major breakthroughs in DNA sequencing technology. These 'next-generation' sequencers have made sequencing more accessible to researchers due to reduced costs and higher throughput. Even though costs are lower, researchers pool samples from different individuals to further reduce costs and to take advantage of the large number of reads produced in a single experiment. However, such pooling prevents meaningful variance estimates and eliminates the possibility of statistical inference at the appropriate unit of analysis. This talk focuses on problems caused by pooling and presents a possible solutions to extract more meaningful information from pooled data arising from next-generation short-read sequencing technologies. Simulations demonstrate the performance of the proposed method, while application of the method to actual data from a run on a 454 Life Sciences sequencer closes out the presentation.

email: davisjwa@health.missouri.edu

## OPTIMAL SHRINKAGE VARIANCE ESTIMATION AND OUTLIER DETECTION IN MICROARRAY DATA ANALYSIS

Nysia I. George\*, U.S. Food and Drug Administration  
Naisiyin Wang, Texas A&M University

A core goal of microarray analysis is to identify an informative subset of differentially expressed genes under different experimental conditions. Typically, this is done through hypothesis testing, which relies on test statistics that properly summarize and evaluate information in the sample(s). A reliable variance estimator that is applicable to all genes is important for analysis. We often find that genome-scale microarray expression analysis generates large data sets with a small number of replicates for each gene. The widespread statistical limitations due to low replication make it necessary to devise adaptive methods for estimating gene-specific variance. Further complicating variance estimation is the frequent presence of outliers in microarray data. We propose a robust modification of optimal shrinkage variance estimation. Our estimator is uninfluenced by outliers and allows for gene-specific, rather than pooled, estimates of variance. In order to increase power, we estimate a common variance term by grouping standardized data so that information shared by genes post-standardization can be more efficiently utilized. For outlier detection we adopt a technique which is based on the false discovery rate approach. Numerous methodologies for estimating variance are compared via simulation and real data analysis of colon cancer microarray data.

email: nysiainet@gmail.com

# ABSTRACTS

## 44. ASSESSING GENE AND ENVIRONMENT INTERACTIONS IN GENOME-WIDE STUDIES

### DETECTING GENE-GENE INTERACTION VIA OPTIMALLY WEIGHTED MARKERS

Jing He\*, University of Pennsylvania School of Medicine  
Mingyao Li, University of Pennsylvania School of Medicine

Gene-gene interactions play an important role in complex human diseases. The detection of gene-gene interactions has long been a challenge due to its complexity. The standard method that aims at detecting gene-gene interaction via pairwise interactions between genetic markers may be inadequate since it does not model linkage disequilibrium (LD) between markers in the same gene and may lose power under complicated genetic models. To improve power over this simple approach, we propose a gene-based interaction test by combining optimally weighted markers. A unique feature of our method is its ability to incorporate LD information provided by a reference dataset such as the HapMap. We analytically derived the optimal weight for both quantitative and binary traits. Since markers in the same gene are correlated, to reduce the degrees of freedom, we summarized the information and tested the interactions using the principle components of the weighted genotype scores. We then applied our method to both simulated and real datasets to evaluate its performance. The preliminary results show that our method generally has greater power in detecting gene-gene interactions than other methods. We believe our method will provide a useful tool in discovering disease susceptibility genes for complex diseases.

email: jinghe@mail.med.upenn.edu

### A LIKELIHOOD-BASED APPROACH FOR DETECTING GENE-GENE INTERACTION IN A CASE-CONTROL STUDY

Saonli Basu\*, University of Minnesota

Many complex traits of medical relevance such as Diabetes, Asthma, and Alzheimer's disease are controlled by multiple genes. Statistical methods for the detection of gene-gene interactions in a case-control study can be categorized broadly into parametric and nonparametric approaches. Among these different nonparametric approaches, there is a growing popularity of the Multifactor Dimensionality Reduction (MDR) approach and it has been recently extensively used for gene-gene interaction detection in many real studies. The strong point in favor of MDR is that it can detect SNPs associated with a disease. It searches through any level of interaction without considering the significance of the main effects. It is therefore able to detect high-order interactions even when the underlying main effects are statistically not significant. We have implemented similar data reduction strategy of MDR within a likelihood framework, which can be used to assess the statistical significance of a k-order gene-gene interaction. This approach also can estimate the combined effect of a group of high-risk SNPs under our model. By means of simulation studies, we have compared the performance of our proposed model with other existing techniques. The performances of all these methods are also studied on a real dataset using ROC analysis.

email: saonli@umn.edu

## HIGH-RESOLUTION QTL MAPPING VIA SIMULTANEOUS ANALYSIS OF DENSE MARKERS

Nengjun Yi\*, University of Alabama at Birmingham

Dense sets of polymorphic markers have been widely used in the molecular dissection of complex traits in experimental organisms and human populations. However, strong statistical correlation or linkage disequilibrium (LD) between densely distributed markers makes it difficult to resolve genetic effects into sufficiently small intervals, especially when analyzing one locus at a time. We propose a high-resolution mapping method using hierarchical generalized linear models that simultaneously analyzes all effects, thereby accommodating relationship among markers. The key to our approach is the use of continuous prior distributions on effects that favor sparseness in the fitted model, enabling us to distinguish causative variants from neighboring correlated noncausative loci. We show that even with a typical inbred F2 cross we can map a causative locus that accounts for 5% of the phenotypic variation to within 1 cM. Our method can handle various continuous or discrete phenotypes, and accommodate covariates and interactions, and is generally applicable to large-scale genetic linkage and association studies in animal, plant and human.

email: nyi@ms.soph.uab.edu

### TESTING FOR GENETIC MAIN EFFECTS IN PRESENCE OF GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS IN GENOME-WIDE ASSOCIATION STUDIES

Arnab Maity\*, Harvard School of Public Health  
Xihong Lin, Harvard School of Public Health

There is growing evidence that gene-gene interaction or epistasis ubiquitously contributes to complex traits. Our interest lies in developing testing procedures of effects of a gene, or a group of genes, in the presence of possible gene-gene interactions. In general, testing for the effect of a particular gene in the presence of gene-gene interaction generally requires testing for the corresponding interaction effect of the gene under consideration with other genes. However, in genome-wide association studies (GWAS), a very high number of genetic markers are genotyped and the number of possible interaction combinations among the genotyped markers is astronomical. Testing for all possible interactions in this situation is likely to lead to loss of power and require a strong parametric assumption. In this paper, we propose a kernel machine based method to model the effects of high-dimensional genetic variants by incorporating gene-gene interactions, and develop statistical procedures to test for the effects of a gene or a group of genes in presence of gene-gene interactions. We will use a garrote kernel method to construct statistical score based tests for gene effects that do not require testing for the coefficients of all the interaction terms allowing us to obtain more powerful tests using less degrees of freedom. We will investigate the asymptotic properties of our proposed test and evaluate its performance via simulation studies. We will demonstrate our methodology by applying them to the Framingham Heart Study GWAS data.

email: amaity@hsph.harvard.edu

## **LOCATE COMPLEX DISEASE LOCI BY INVESTIGATING GENE AND ENVIRONMENT INTERACTION FOR GENOME-WIDE ASSOCIATION STUDIES**

Jin Zheng\*, University of Michigan  
Goncalo R. Abecasis, University of Michigan

Most of complex diseases result from the interaction of gene and environmental factors. But we still do not know the best way to locate complex disease susceptibility loci by exploiting the gene environment interaction, especially at the genome-wide associate studies framework, in which hundreds of thousands markers are scanned. The traditional model is the logistic model, with covariates of genotypic and exposure status, and the interaction term of them, testing how gene and environment affect disease interactively. Another way is to detect the gene environment interaction in case-only data. Kraft et al (2007) proposed a two-degrees-of-freedom test for both interaction effect and genotypic effect. Recently, Murcary et al develop a two-step method. They discover the marginal correlation of gene and environment first and then use the traditional model at the second step, and claimed it is more powerful compared to the traditional method. In this paper, we would compare performance of several methods for simulated data under genome-wide association structure. And we propose a one-step method, which considers the correlation of gene and environment at both case and control groups. We speculate this would gain power at some situations.

email: jzhy@umich.edu

## **A GENERAL FRAMEWORK FOR ESTIMATING GENETIC EFFECTS AND GENE-ENVIRONMENT INTERACTIONS WITH MISSING DATA**

Yijuan Hu\*, University of North Carolina-Chapel Hill  
Danyu Lin, University of North Carolina-Chapel Hill  
Donglin Zeng, University of North Carolina-Chapel Hill

Missing data arise in genetic association studies when genotypes are unknown or when haplotypes are of direct interest. We provide a general likelihood-based framework for estimating genetic effects and gene-environment interactions with such missing data. We allow genetic and environmental variables to be correlated while leaving the distribution of environmental variables completely unspecified. We consider three major study designs ---cross-sectional, case-control, and cohort designs---and construct appropriate likelihood functions for all common phenotypes (e.g., case-control status, quantitative traits, and potentially censored ages at onset of disease). The likelihood functions involve both finite- and infinite-dimensional parameters. The maximum likelihood estimators are shown to be consistent, asymptotically normal, and asymptotically efficient. Fast and stable numerical algorithms are developed to implement the corresponding inference procedures. Extensive simulation studies demonstrate that the proposed inferential and numerical methods perform well in practical settings. Illustration with a genomewide association study of schizophrenia is provided.

email: yhu@bios.unc.edu

## **RISK EFFECT ESTIMATION FOR MULTIPLE PHENOTYPES AND GENE-ENVIRONMENT INTERACTION: A CONDITIONAL LIKELIHOOD APPROACH**

Arpita Ghosh\*, University of North Carolina-Chapel Hill  
Fei Zou, University of North Carolina-Chapel Hill  
Fred A. Wright, University of North Carolina-Chapel Hill

The use of genome-wide testing thresholds in association scans is known to inflate estimates of genetic risk among significant SNPs (the “winner’s curse”). We have recently reported an approximate conditional likelihood approach to correct for this bias, using the estimated risk effect and its standard error as reported by standard statistical software. A similar problem arises when risk estimation is performed for secondary effects, such as secondary phenotypes or gene-environment interactions, when the secondary analysis is restricted to SNPs that are significant for the primary phenotype. Such secondary bias can be substantial, and we describe an extension of our conditional likelihood approach to the multivariate setting where multiple effect coefficients are simultaneously estimated. The results have considerable importance for the proper analysis of secondary effects, and in the design of follow-up studies.

email: aghosh@bios.unc.edu

## **45. HYPOTHESIS TESTING**

### **PENALIZED LIKELIHOOD RATIO TEST WHEN SOME PARAMETERS ARE PRESENT ONLY UNDER THE ALTERNATIVE**

Chongzhi Di\*, Johns Hopkins University  
Kung-Yee Liang, Johns Hopkins University  
Ciprian M. Crainiceanu, Johns Hopkins University

We consider hypothesis testing problems where some parameters are present only under the alternative hypothesis. Examples include testing the existence of a change point, testing the number of components in finite mixture models, and testing linkage in genetic epidemiology. The likelihood ratio test (LRT) statistic does not follow a conventional chi-square distribution in these problems, due to nonidentifiability. In this paper, we extend the modified/penalized likelihood ratio test, proposed by Chen et al. (2001) in finite mixture models, to a wider class of problems and obtain its asymptotic null distribution and power. Simulation studies suggest that asymptotic results are accurate in small and moderate samples and are insensitive to the magnitude and the functional form of the penalty. We apply the proposed method to a data set testing the null hypothesis of linearity versus a nonlinear alternative.

email: cdi@jhsph.edu

# ABSTRACTS

## LIKELIHOOD RATIO TEST FOR QUALITATIVE INTERACTIONS

Qing Pan\*, George Washington University

Often, the factor of interest has qualitatively different effects across subject subgroups. Two interesting cases motivate the need for a formal statistical test targeting at detecting cross-over patterns in treatment effects. In the Diabetes Prevention Program, the Metformin treatment was shown to reduce the risk of type II diabetes. What is of concern is whether Metformin works negatively (increase the risk) for certain types of patients. In the second motivating example, the state of Illinois was sued by the Black members over the fairness in the promotion process. Due to the limited sample size, the odds ratio of promotion between the protected and the majority groups would be calculated across all ranks as long as no qualitative (instead of quantitative) differences exist among the race effects in different ranks. Gail and Simon (1985) proposed a likelihood ratio test (LRT) for qualitative interactions by comparing the unrestricted MLE to the parameter values under the null with the smallest test statistics value. This paper proposes another LRT where the projection of the unrestricted MLE to the null region is employed. The asymptotic mixed Chi-square distribution is derived. Size and power are verified under logistic and proportional hazards models. Finally, this new test is applied to the two motivating cases.

email: qpan@gwu.edu

## USING MULTIPLE CONTROL GROUPS AS EVIDENCE ABOUT UNOBSERVED BIASES IN AN OBSERVATIONAL STUDY OF TREATMENTS FOR MELANOMA

Frank B. Yoon, University of Pennsylvania  
Phyllis A. Gimotty, University of Pennsylvania  
DuPont Guerry, University of Pennsylvania  
Paul R. Rosenbaum, University of Pennsylvania

In an observational study of treatments effects, treated and control subjects may differ systematically prior to treatment, and there is invariably concern that some important covariates were not measured, so that adjustments such as matching may fail to render the groups comparable. We present a simple example from an observational study of a surgical treatment for melanoma that uses two control groups to provide some evidence about unmeasured biases. We illustrate a procedure for decomposing complex hypotheses into simpler hypotheses that are tested in order of priority. This use of a second control group is “without cost” to the investigator, in the specific sense that inferences about the first control group are completed without loss of power to corrections for multiple testing; then, only if these initial results are promising, are further inferences made using the second control group.

email: fby@wharton.upenn.edu

## AN EFFICIENT RANK-BASED TEST FOR THE GENERALIZED NONPARAMETRIC BEHRENS-FISHER PROBLEM

Kai F. Yu\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development

Qizhai Li, National Cancer Institute

Aiyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development

Kai Yu, National Cancer Institute

For a generalized Behrens-Fisher hypothesis problem for comparison of multiple outcomes commonly encountered in biomedical research, Huang et al. (2005, Biometrics, 61, 532-539) improved O'Brien's (1984, Biometrics, 40}, 1079-1087) rank-sum tests with replacement of the ad hoc variance by the asymptotic variance of the test statistics. The improved tests control the Type error at the desired level and gain power when the differences in each individual outcome variable fall into the same direction. However, they may lose power when the differences are in different directions ( e.g. some are positive, and some are negative). We propose a more efficient test statistic, taking the maximum of individual rank-sum statistics, that controls the type I error and maintains satisfactory power regardless of the directions of differences. Data from two studies are used to illustrate the application of the proposed test.

email: yukf@mail.nih.gov

## NONPARAMETRIC PROCEDURES FOR COMPARING CORRELATED MULTIPLE ENDPOINTS WITH APPLICATIONS TO OXIDATIVE STRESS BIOMARKERS

Aiyi Liu\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health  
Chunling Liu, Harvard School of Public Health  
Enrique Schisterman, Harvard School of Public Health

Multiple endpoints are common in clinical and epidemiologic studies. To compare multiple endpoints between two independent groups, O'Brien (1984) and Huang et al. (2005) proposed rank-sum based nonparametric testing procedures. In this talk we extend their methods to comparing multiple endpoints between two dependent groups. We further obtain a more powerful test based on optimizing the linear combination of ran-sum statistics. The methods are illustrated using data from the BioCycle Study to compare the levels of oxidative stress biomarkers between a woman's two consecutive menstrual cycles.

email: liua@mail.nih.gov

## ON GLOBAL P-VALUE CALCULATION IN MULTI-STAGE DESIGNS

Shanhong Guan\*, Merck & Co.

In multi-stage clinical trials, test statistic can be constructed based on combination of the stage-wise p-values, assuming the condition of p-clud property is satisfied. In this paper, methods based on



conditional error principle, such as direct combination of p-values and inverse-normal p-values, and methods based on significance levels of individual stages, are introduced. Numerical studies are conducted to evaluate their properties and some examples are provided to illustrate the application of these methods to multi-stage adaptive design.

email: shanhong\_guan@merck.com

## INSIGHTS INTO P-VALUES AND BAYES FACTORS VIA FALSE POSITIVE AND FALSE NEGATIVE BAYES FACTORS

Hormuzd A. Katki\*, National Cancer Institute

The Bayes Factor and likelihood ratio have been shown to have a stronger theoretical justification than p-values for quantifying statistical evidence. However, when the goal of a study is solely hypothesis testing, the Bayes Factor is a black box that does not yield insight about false positive versus false negative results. I introduce the False Positive Bayes Factor and the False Negative Bayes Factor and show that they are approximately the two components of the Bayes Factor. In analogy with diagnostic testing, decomposing the Bayes Factor into the False Positive/Negative Bayes Factors provides additional insight not obvious from the Bayes Factor. The False Positive/Negative Bayes Factors rely on only the p-value and the power under an alternative hypothesis, forging a new link of the p-value to the Bayes Factor. This link can be exploited in data analysis to understand any potentially contradictory inferences drawn by Bayes Factors versus p-values. The False Positive/Negative Bayes Factors provide insight in a genome-wide association study of prostate cancer by helping to reveal the two SNPs declared positive by p-values and Bayes Factors that with future data turned out to be false positives.

email: hkatki@gmail.com

## 46. VARIABLE SELECTION METHODS

### VARIABLE SELECTION FOR IDENTIFYING ENVIRONMENTAL CONTAMINANTS ASSOCIATED WITH HUMAN FECUNDITY

Sungduk Kim\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health  
Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health  
Germaine M. Louis, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

A question of considerable interest is in identifying polychlorinated biphenyl (PCB) and lifestyle variables that are associated with human fecundity. Human fecundity is quantitatively measured through time-to-pregnancy, the number of cycles needed to get pregnant, a censored discrete survival time. Furthermore, PCBs measured are typically highly correlated. So in order to assess their association with human fecundity, requires variable selection in high-dimensional setup with outcome of interest being a discrete right-censored survival time. Such problems are also of interest in dealing with microarray data analysis. However, attention has typically focused on censored continuous survival time. Here, we consider a discrete survival time model. In

this paper, we propose a Bayesian variable selection approach, which allows the identification of relevant chemical by jointly assessing sets of chemicals. The proposed method provides a unified procedure for the selection of relevant chemicals and the prediction of survivor functions. This is accomplished by building a stochastic search variable selection method within discrete survival model. We will investigate our methods through simulations and analysis of real example involving fecundity.

email: kims2@mail.nih.gov

## BAYESIAN VARIABLE SELECTION FOR LATENT CLASS MODELS

Joyee Ghosh\*, University of North Carolina at Chapel Hill  
Amy H. Herring, University of North Carolina at Chapel Hill

Weight gain during pregnancy is believed to have an effect on various maternal and child health outcomes like child birth weight, maternal postpartum weight retention, gestational diabetes etc. Pregnancy weight gain is considered to be a modifiable risk factor and hence this is an important and active area of research. One of our goals is to investigate whether maternal characteristics such as age, gender, diet etc. are predictive of weight gain adequacy, defined based on Institute of Medicine criteria as inadequate, adequate or excessive. We use a Bayesian multinomial logistic model with Bayesian variable selection to assess the importance of predictors of weight gain. Our approach can be extended to latent class models with class probabilities depending on subject-specific predictors of risk, while allowing uncertainty regarding which predictors to include.

email: joyee123in@gmail.com

## REGULARIZATION PARAMETER SELECTIONS VIA GENERALIZED INFORMATION CRITERION

Yiyun Zhang\*, Penn State University  
Runze Li, Penn State University  
Chih-Ling Tsai, University of California, Davis

We apply the nonconcave penalized likelihood approach to obtain variable selections as well as shrinkage estimators. This approach relies heavily on the choice of regularization parameter, which controls the model complexity. In this paper, we propose employing the generalized information criterion, encompassing the commonly used Akaike information criterion (AIC) and Bayesian information criterion (BIC), for selecting the regularization parameter. Our proposal makes a connection between the classical variable selection criteria and the regularization parameter selections for the nonconcave penalized likelihood approaches. We show that the BIC-type selector enables identification of the true model consistently, and the resulting estimator possesses the oracle property in the terminology of Fan and Li (2001). In contrast, the AIC-type selector tends to overfit. However, similar as the loss efficiency of Li (1987) and Shao (1997), we further showed that under appropriate conditions, AIC selector enjoys an asymptotic loss efficiency which BIC-type selectors do not possess. Our simulation results confirm these theoretical findings. An application in breast cancer mammography assessment is presented.

email: yuz115@psu.edu

# ABSTRACTS

## PENALIZED ESTIMATING EQUATIONS FOR SEMIPARAMETRIC LINEAR TRANSFORMATION MODELS

Hao Zhang, North Carolina State University  
Wenbin Lu\*, North Carolina State University  
Hansheng Wang, Peking University

Semiparametric linear transformation models have received much attention due to its high flexibility in modeling survival data. However, the problem of variable selection for linear transformation models is less studied since a convenient loss function is not available. In this paper, we propose a penalized estimating equation (PEE) approach for joint parameter estimation and variable selection for linear transformation models. The new procedure consists of computing the ‘profiled score’ from the estimating equations for the finite dimensional parameter, constructing the variance weighted L<sub>2</sub>-distance based on the profile score, and minimizing the distance subject to some shrinkage penalty. The resulting estimator is shown to be consistent in both parameter estimation and variable selection, and asymptotically normal with improved efficiency. An efficient one-step algorithm is further proposed, which makes it possible to obtain the entire solution path of the estimate. Numerical studies show that the PEE performs competitively with other likelihood based methods.

email: lu@stat.ncsu.edu

## REGULARIZED ESTIMATION IN AFT MODELS WITH HIGH-DIMENSIONAL COVARIATES

Liping Huang\*, University of Kentucky  
Mai Zhou, University of Kentucky  
Arne C. Bathke, University of Kentucky

Several recent researches have focused on the use of AFT models to predict survival times of future cancer patients by investigating their gene expression profiles based on microarray analysis. We investigate use of the elastic net regularization approach to estimation and variable selection in the accelerated failure time model with high-dimensional covariates based on the so called inverse probability of censor weighting method. Huang, Ma and Xie (2006) studied a similar setting using LASSO. However, after ordering the survival times, if the last patient is censored, the current weighting method for that patient is problematic especially so for data with high censoring. We propose and investigate some modified weighting methods for the last patient in this study and show that some modification has improved the prediction performance. We use V-fold cross-validation for tuning parameter selection. The proposed method is evaluated using simulations and applied to real data.

email: liping.huang@uky.edu

## VARIABLE SELECTION FOR THE COX REGRESSION MODEL WITH COVARIATES MISSING AT RANDOM

Ramon Garcia\*, University of North Carolina

We consider variable selection in the Cox Regression model (Cox, 1972, 1975) with covariates missing at random. We investigate the smoothly clipped absolute deviation penalty and adaptive LASSO penalty, and propose a unified model selection and estimation procedure. A computationally attractive algorithm is developed which simultaneously optimizes the penalized likelihood function and penalty parameters. We also optimize a model selection criterion, called the IC<sub>Q</sub> statistic, to estimate the penalty parameters and show that it consistently selects all important covariates. Simulations are performed to evaluate the finite sample performance of the penalty estimates. Also, a lung cancer dataset is analyzed to demonstrate the proposed methodology.

email: rigarcia@email.unc.edu

## BAYESIAN SEMIPARAMETRIC FRAILTY SELECTION IN MULTIVARIATE EVENT TIME DATA

Bo Cai\*, University of South Carolina

Biomedical studies often collect multiple event time data from multiple clusters (either individual subjects or groups of subjects) within each of which event times for subjects are correlated and the correlation may vary in different classes. In such survival analyses, heterogeneity among clusters for overall and specific classes can be accommodated by incorporating parametric gamma frailty or log-normal frailty terms into the model. In this article, we propose a Bayesian approach to relax the parametric distribution assumption for overall and subject-specific frailties by using a Dirichlet process prior while also allowing for the uncertainty of heterogeneity for different classes. Subject-specific frailty selection relies on variable selection-type mixture priors by applying mixtures of point masses at zero and inverse gamma distributions to the log-frailty variances. This selection allows frailty with zero variance to effectively drop out of the model. A reparameterization of log frailty is performed to reduce dependence among the parameters resulting in faster MCMC convergence. Simulated data examples are presented and a real data example is also used for illustration.

email: bocai@gwm.sc.edu

## 47. LONGITUDINAL DATA ANALYSIS

### THREE-LEVEL MIXED EFFECTS LOCATION SCALE MODEL

Eisuke Segawa\*, University of Illinois at Chicago  
Donald Hedeker, University of Illinois at Chicago  
Robin J. Mermelstein, University of Illinois at Chicago

Hedeker et. al. (2008) provided an example of a two-level mixed-effects location scale model. In that application, each subject was observed repeatedly up to 40 times; several times within a day over

several days. The mixed-effects location scale model allows the variance of random intercept (scale) to vary stochastically over subjects as well as the intercept (location). However, this two-level model does not distinguish within-day variation from between-day variation: they both comprise the within-subjects variance term. In order to separate the within-day and between-day variation (both are within-subjects), the two-level mixed-effects location scale model is extended to three-levels, treating observations nested within days nested within subjects.

email: esegawa@uic.edu

## NONPARAMETRIC MODELING OF SEMI-CONTINUOUS DATA WITH APPLICATION TO MEDICAL COST DATA ANALYSIS

Pang Du\*, Virginia Tech  
Lei Liu, University of Virginia  
Anna Liu, University of Massachusetts-Amherst

Semi-continuous data arise in longitudinal studies when the repeated measures contain a large portion of zero values, as well as right skewness and heteroscedasticity for non-zero positive values, e.g., monthly medical costs, daily drinking records, or number of days hospitalized within a year. We propose a nonparametric random effects two-part model to analyze this type of data. One part of the model is for the odds of being positive, the other is for the positive measurements, and the two parts are joined by the correlated random effects. Nonparametric smooth estimates of the pattern functions in both parts are obtained through the minimization of a penalized joint likelihood. The model is then applied to the analysis of longitudinal monthly medical costs of chronic heart failure patients from the clinical data repository (CDR) at the University of Virginia.

email: pangdu@vt.edu

## SEMIPARAMETRIC ANALYSIS OF MULTIVARIATE RECURRENT AND TERMINAL EVENTS

Liang Zhu\*, St. Jude Children's Research Hospital  
Jianguo Sun, University of Missouri-Columbia

Recurrent event data occur in many clinical and observational studies and in this type of data, it is often the case that there also exists a terminal event such as death that is related to the recurrent event of interest. In addition, sometimes there may exist more than one type of recurrent events, that is, one faces multivariate recurrent event data with some correlated terminal event. It is apparent that in these situations, one has to take into account the dependence both among different types of recurrent events and between the recurrent and terminal events. In this paper, we propose two approaches for regression analysis of such data, a joint modelling approach and a marginal model approach. Both finite and asymptotic properties of the proposed estimates are established. The methods are applied to a set of bivariate recurrent event data arising from a study of the patients with the end-stage renal disease.

email: liang.zhu@stjude.org

## DETERMINING WHEN TIME RESPONSE CURVES DIFFER IN THE PRESENCE OF CENSORSHIP WITH APPLICATION TO A RHEUMATOID ARTHRITIS BIOMARKER STUDY

Ann A. Lazar\*, Harvard School of Public Health & Dana-Farber Cancer Institute  
Gary O. Zerbe, University of Colorado at Denver

We propose a method to determine the significance regions of time response curves in random coefficient models with censored data. Here, significance region refers to the set of times for which the curves differ, and censored data arise when all that is known is that the data are less than or more than a known value. An explicit solution provides the set of times for which the two curves (population average or subject specific) generated from the models significantly differ with adjustment of the resulting significance level via the Scheffé method for multiple comparisons. The application and motivation for this methodology is from a longitudinal biomarker study of subjects with rheumatoid arthritis. For example, it is useful to determine when left censored autoantibody levels (such as rheumatoid factor) differ between the cases and controls during the pre-diagnosis study period.

email: alazar@hsph.harvard.edu

## SAS/IML FOR PARAMETER ESTIMATION OF LOGISTIC REGRESSION FOR TRANSITION, REVERSE TRANSITION AND REPEATED TRANSITION FROM FOLLOW-UP DATA

Rafiqul I. Chowdhury\*, Kuwait University, Kuwait  
M. A. Islam, Dhaka University, Bangladesh  
Shahriar S. Huda, Kuwait University, Kuwait

In the past, most of the works on Markov models dealt with estimation of transition probabilities for first or higher orders and it appears to be restricted due to over-parameterization though several attempts have been made to simplify. Muenz and Rubinstein (1985) employed logistic regression models to analyze the transition probabilities from one state to another for first order and the model was extended for higher order by Islam and Chowdhury (2006). Islam and Chowdhury (2007), using the Chapman-Kolmogorov equations, introduced an improvement over the previous methods in handling runs of events by expressing the conditional probabilities in terms of the transition probabilities generated from Markovian assumptions. They introduced three sets of models namely transition, reverse transition and repeated transition to take account of unequal intervals in the occurrence of events. To estimate the parameters of the models proposed by Islam and Chowdhury (2007), extensive pre-processing and computations are needed to prepare the data before one can use the standard available procedures in existing statistical software. In this paper we demonstrate a program developed using SAS/IML to estimate the parameters of the proposed model. The program has been demonstrated using follow-up data on Health and Retirement Survey (HRS) from USA.

email: rafiq@hsc.edu.kw

# ABSTRACTS

## A COMPOSITE LIKELIHOOD APPROACH TO THE ANALYSIS OF LONGITUDINAL CLONAL DATA ON MULTITYPE CELLULAR SYSTEMS UNDER AN AGE-DEPENDENT BRANCHING PROCESS

Rui Chen\*, University of Rochester Medical Center  
Olivier Hyrien, University of Rochester Medical Center

The theory of age-dependent branching processes provides an appealing statistical framework for drawing inference on cell kinetics from clones observed longitudinally at discrete time points. Likelihood inference being difficult in this context, we propose an alternative composite likelihood approach, where the estimation function is defined from the marginal or conditional distributions for the number of cells of each observable cell type. These distributions have generally no closed-form expressions but can be approximated using simulations. We construct a bias-corrected version of the estimating function, which also offers computational advantages. The composite likelihood estimator is proven to be consistent and asymptotically normal, and its finite-sample properties are investigated in simulation studies. It is shown that the proposed approach outperforms the existing methods. Finally an application to the analysis of the effect of neurotrophin-3 on the generation of oligodendrocytes from oligodendrocyte type-2 astrocyte progenitor cells cultured in vitro is presented.

email: rui\_chen@urmc.rochester.edu

## THE UNIVARIATE APPROACH TO REPEATED MEASURES ANOVA FOR HIGH DIMENSION, LOW SAMPLE SIZE

Yueh-Yun Chi\*, University of Florida  
Keith E. Muller, University of Florida

The “univariate approach” to Gaussian repeated measures (UNIREP) extends naturally to High Dimension, Low Sample Size (HDLSS) data. The UNIREP test requires only orthonormal invariance and therefore can be computed with HDLSS data, in contrast to multivariate techniques. However, the simulations demonstrate that traditional tests fail badly in controlling type I error rate. We describe a sphericity parameter estimator and associated UNIREP test for HDLSS data that controls type I error rate well for HDLSS data with any population covariance matrix. The method appropriately accounts for the repeated measures or multivariate nature of HDLSS data such as seen in metabolomics, imaging, and genomics.

email: yychi@biostat.ufl.edu

## 48. MULTIPLE TESTING IN HIGH-DIMENSIONAL DATA

### ESTIMATION OF FALSE DISCOVERY RATE USING PERMUTATION P-VALUES WITH DIFFERENT DISCRETE DISTRIBUTIONS

Tim Bancroft\*, Iowa State University  
Dan Nettleton, Iowa State University

The false discovery rate (FDR) is a multiple testing error rate which describes the expected proportion of type I errors to the total number of rejected hypotheses. Benjamini and Hochberg introduced this quantity and provided an estimator that is conservative when the number of true null hypotheses,  $m_0$ , is smaller than the number of tests,  $m$ . Replacing  $m$  with  $m_0$  in Benjamini and Hochberg's procedure reduces the conservative bias, but requires estimation as  $m_0$  is unknown. Methods exist to estimate  $m_0$  when the  $m$  p-values are distributed continuous  $U(0,1)$  under  $H_0$ . This talk discusses how to estimate  $m_0$  and therefore FDR when the  $m$  p-values are from a mixture of different discrete distributions resulting from permutation testing for data with many zeros. The method will be demonstrated through an analysis of proteomics data.

email: timmyb@iastate.edu

## CONTROLLING FALSE DISCOVERIES IN MULTIDIMENSIONAL DIRECTIONAL DECISIONS, WITH APPLICATIONS TO GENE EXPRESSION DATA ON ORDERED CATEGORIES

Wenge Guo\*, National Institute of Environmental Health Sciences  
Sanat K. Sarkar, Temple University  
Shyamal D. Peddada, National Institute of Environmental Health Sciences

Time-course or dose-response microarrays gene expression studies are common in biomedical research. A goal of such studies is to identify gene expression patterns over time, dose, or tumor stage etc. In this project, we formulated this problem as a multiple testing problem where for each gene the null hypothesis of no difference between the successive mean gene expressions are tested and further directional decisions are made if it is rejected. Much of the existing multiple testing procedures are devised for controlling the usual FDR rather than the mixed directional FDR (mdFDR), the expected proportion of Type I and directional errors among all rejections. In this project, we considered the problem of controlling the mdFDR involving multidimensional parameters. To deal with this problem, we developed a procedure extending Benjamini and Yekutieli (2005)'s one-dimensional directional BH procedure based on the Bonferroni test for each gene. We proved that the proposed procedure controls the mdFDR when the underlying test statistics are independent across the genes. We also applied the proposed methodology to a time-course microarray data and obtained several biologically interesting results.

email: wenge.guo@gmail.com

## SIMULTANEOUS TESTING OF GROUPED HYPOTHESES: FINDING NEEDLES IN MULTIPLE HAYSTACKS

Wenguang Sun\*, North Carolina State University  
Tony Cai, University of Pennsylvania

In large-scale multiple testing problems, hypotheses are often collected from heterogeneous sources and hence form into groups that exhibit different characteristics. Conventional approaches, including the pooled and separate analyses, fail to efficiently utilize the external grouping information. We develop a compound decision theoretic

framework for testing grouped hypotheses and introduce an oracle procedure that minimizes the false non-discovery rate subject to a constraint on the false discovery rate. It is shown that both the pooled and separate analyses can be uniformly improved by the oracle procedure. We then propose a data-driven procedure that is shown to be asymptotically optimal. A numerical study shows that our procedures enjoy superior performance and yield the most accurate results in comparison with both the pooled and separate procedures. The results demonstrate that exploiting external information of the sample can greatly improve the efficiency of a testing procedure, and provide additional insights on how to optimally combine the simultaneous inferences made for multiple groups.

email: kenkingsun@hotmail.com

## FINDING CRITICAL VALUE FOR t-TESTS IN VERY HIGH DIMENSIONS

Hongyuan Cao\*, University of North Carolina-Chapel Hill  
Michael R. Kosorok, University of North Carolina-Chapel Hill

In micro array studies, image analysis, high throughput molecular screening, astronomy, and in many other high dimensional statistical problems, hypothesis testing is applied in a simultaneous way. Popular alternatives to familywise error rate for calibration of multiplicity include k-familywise error rate (k-FWER), false discovery rate (FDR) and false discovery proportion (FDP). Most procedures in the literature are based on the assumption that the p-values of tests are known or that the distributions under null and alternative hypothesis are known. However, these assumptions are usually not realistic. In this paper, we focus on one-sample and two-sample t-statistics and develop a procedure to find cut-off values to control k-FWER, FDR and tail probability of FDP (FDTP). Our approach doesn't require any distributional assumptions on the population and is robust as long as the population has finite third moment plus some very general conditions on the mean and variance. We also develop a new estimator for the proportion of alternative hypotheses. Simulation results and a real data example show that our approach is generally better than the Benjamini and Hochberg (1995)'s procedure.

email: hycao@email.unc.edu

## MULTIPLICITY-CALIBRATED BAYESIAN HYPOTHESIS TESTS

Mengye Guo\*, University of Pennsylvania  
Daniel F. Heitjan, University of Pennsylvania

When testing multiple hypotheses simultaneously, there is a need to adjust the levels of the individual tests to effect control of the family-wise-error-rate (FWER). Standard frequentist adjustments effectively control the error rate but are typically conservative and oblivious to prior information. We propose a Bayesian testing approach --- Multiplicity-Calibrated Bayesian Hypothesis Testing (MCBHT) --- that sets individual critical values to reflect the prior information while controlling the FWER via the Bonferroni inequality. If the prior information is specified correctly, in the sense that those null hypotheses considered most likely to be false in fact are false, the power of our method is substantially greater than that of standard

frequentist approaches. We illustrate our method using data from a preclinical cancer study and a pharmacogenetic trial. We demonstrate its error rate control and power advantage by simulation.

email: mengyego@mail.med.upenn.edu

## SPIKE AND SLAB DIRICHLET PRIOR FOR BAYESIAN MULTIPLE TESTING IN RANDOM EFFECTS MODELS

Sinae Kim\*, University of Michigan  
David B. Dahl, Texas A&M University  
Marina Vannucci, Rice University

We propose a method for multiple hypothesis testing in random effects models that uses Dirichlet process (DP) priors for a nonparametric treatment of the random effects distribution. We consider a general model formulation which accommodate a variety of multiple treatment conditions. A key feature of our method is the use of a product of "spike and slab" distributions as the centering distribution for the DP prior. Adopting spike and slab centering priors readily accommodates sharp null hypotheses and allows for the estimation of the posterior probabilities of such hypotheses. We demonstrate via a simulation study that our method yields increased sensitivity in multiple testing hypothesis and produces a lower proportion of false discoveries than other competitive methods. In our application, the modeling framework allows simultaneously inference on the parameters governing differential expression and inference on the clustering of genes. We use experimental data on the transcriptional response to oxidative stress in mouse heart muscle and compare the results from our procedure with existing Bayesian methods.

email: sinae@umich.edu

## COMPUTATION OF EXACT P-VALUES FOR NONPARAMETRIC TEST

Yuanhui Xiao\*, Georgia State University

As large sets of high-throughput data in genomics and proteomics become more readily available, there is a growing need for fast algorithms designed to compute the exact p-values of distribution-free tests. In this talk we present some issues regarding exact p-values as well some ideas about computing exact p-values for a class of distribution-free tests.

email: matyxx@langate.gsu.edu

# ABSTRACTS

**MONDAY, MARCH 16, 2009**  
**3:45-5:30 PM**

## 49. ROLE OF META-ANALYSIS IN DRUG DEVELOPMENT

### THE USE OF CUMULATIVE META ANALYSIS IN DRUG DEVELOPMENT

Kuang-Kuo G. Lan\*, Johnson & Johnson PRD

Meta analysis is an extremely useful method to summarize accrued information. However, due to heterogeneity of various study background, it is difficult to provide reliable inference from the pooled data. The problem becomes more serious in cumulative meta analysis since (i) the between-study variation cannot be estimated accurately when there are only a few studies under consideration, and (ii) the characteristics of future studies are hard to predict. In general, the parameters under consideration may not follow a simple parametric distribution. Motivated by the Law of Iterated Logarithm, we propose adding a multiplicative penalty factor to the cumulative test statistic. This will introduce a conservative approach to show efficacy for a new compound, or a new treatment procedure.

email: glan@its.jnj.com

### UTILITY AND PITFALLS OF META ANALYSIS FOR DESIGNING NON-INFERIORITY TRIAL

H.M. James Hung\*, U.S. Food and Drug Administration

As ethical reasons prohibit enrolling study patients into a placebo arm, non-inferiority trial designs are employed more frequently in clinical research. The classical non-inferiority trial methodology usually requires a non-inferiority margin be pre-specified at the trial design stage to clearly define the clinical/statistical hypothesis for testing. In the absence of a placebo arm in the trial, the margin can only be selected relying on a guidance from some kind of meta-analysis of the existing relevant placebo controlled trials for estimating the effect of the selected active control. In this presentation, I shall discuss the roles of meta analysis in terms of the extent of useful evidence or information that may be carved out by such a analysis.

email: hsienming.hung@fda.hhs.gov

### ARE THINGS REALLY AS UN-ROSI AS THEY APPEAR?

Michael A. Proschan\*, National Institute of Allergy and Infectious Diseases, National Institutes of Health

A recent meta-analysis implicated the diabetes drug rosiglitazone in an excess of myocardial infarctions and cardiovascular deaths compared to control. Many of the trials in the meta-analysis had very few events, and a naïve approach of combining all control arms and comparing to the combined rosiglitazone group actually showed a higher proportion of events on control. Is the evidence against rosiglitazone really convincing? What is the appropriate analysis when several trials

have very few events? What sort of sensitivity analyses can be done to either bolster or cast doubt on the results? These are some of the topics covered in this talk.

email: ProschaM@mail.nih.gov

## 50. ANALYSIS OF HIGH-DIMENSIONAL DATA WITH BIOLOGICAL APPLICATIONS

### MAXIMUM LIKELIHOOD ESTIMATION OF A MULTIDIMENSIONAL LOG-CONCAVE DENSITY

Richard Samworth\*, University of Cambridge  
Madeleine Cule, University of Cambridge  
Robert Gramacy, University of Cambridge  
Michael Stewart, University of Sydney

We show that if  $X_1, \dots, X_n$  are a random sample from a log-concave density  $f$  in  $\mathbb{R}^d$ , then with probability one there exists a unique maximum likelihood estimator  $\hat{f}_n$  of  $f$ . The use of this estimator is attractive because, unlike kernel density estimation, the estimator is fully automatic, with no smoothing parameters to choose. We exhibit an iterative algorithm for computing the estimator and show how the method can be combined with the EM algorithm to fit finite mixtures of log-concave densities. The talk will be illustrated with pictures from the R package LogConcDEAD. I also hope to discuss some results on the theoretical performance of the estimator.

email: rjs57@hermes.cam.ac.uk

### FORWARD-LASSO ADAPTIVE SHRINKAGE

Gareth James\*, University of Southern California  
Peter Radchenko, University of Southern California

Both Forward Selection (FS) and the Lasso can perform variable selection in high dimensional regression problems. Although the Lasso is the solution to an optimization problem while FS is purely algorithmic, the two methods utilize surprisingly similar approaches. Both add to the model the variable which has the highest correlation with the residual vector, the only difference being in the level of shrinkage. We propose a new method, Forward-Lasso Adaptive SHrinkage (FLASH), which incorporates both FS and the Lasso as special cases. FLASH is fitted using a variant of the LARS algorithm and works well in situations where neither Lasso nor FS succeeds. We prove that it can be formulated as the solution to a weighted Lasso optimization problem. We also demonstrate, on an extensive set of simulations and a real world data set, that FLASH generally outperforms many competing approaches.

email: gareth@usc.edu

## PARTIAL CORRELATION ESTIMATION BY JOINT SPARSE REGRESSION MODELS

Ji Zhu\*, University of Michigan  
Jie Peng, University of California, Davis  
Pei Wang, Fred Hutchinson Cancer Center  
Nengfeng Zhou, University of Michigan

In this talk, we propose a computationally efficient approach for selecting non-zero partial correlations under the high-dimension-low-sample-size setting. This method assumes the overall sparsity of the partial correlation matrix and employs sparse regression techniques for model fitting. We illustrate the performance of our method by extensive simulation studies. It is shown that our method performs well in both non-zero partial correlation selection and the identification of hub variables, and also outperforms two existing methods. We then apply our method to a microarray breast cancer data set and identify a set of “hub genes” which may provide important insights on genetic regulatory networks. Finally, we prove that, under a set of suitable assumptions, the proposed procedure is asymptotically consistent in terms of model selection and parameter estimation.

email: jizhu@umich.edu

## ESTIMATION IN ADDITIVE MODELS WITH HIGHLY CORRELATED COVARIATES

Jiancheng Jiang, University of North Carolina at Charlotte  
Yingying Fan\*, University of Southern California  
Jianqing Fan, Princeton University

Motivated by normalizing Affymetrix array data, we explore nonparametric estimation of highly correlated confounding effects while leaving the high-dimensional treatment effects as nuisance parameters. A simple difference operation reduces the problem to additive models with highly correlated covariates. We introduce two novel approaches for estimating the nonparametric components, integration estimation and pooled backfitting estimation. The former is designed for highly correlated intensity effects, and the latter is useful for non-highly correlated intensity effects. Asymptotic normalities of the proposed estimators are established. Simulations are conducted to demonstrate finite sample behaviors of the proposed methods.

email: fanyingy@marshall.usc.edu

## 51. ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE OBSERVATION AND DROPOUT PROCESSES

## REGRESSION ANALYSIS OF LONGITUDINAL DATA WITH DEPENDENT OBSERVATION PROCESS

(Tony) Jianguo Sun\*, University of Missouri-Columbia

Longitudinal data frequently occur in many studies such as longitudinal follow-up studies. To develop statistical methods and theory for the analysis of them, independent or noninformative

observation and follow-up times are typically assumed, which naturally leads to inference procedures conditional on observation and follow-up times. In many situations, however, this may not be true or realistic. That is, observation times may depend on or be related with the longitudinal responses and the same could be true for the follow-up time. This talk discusses the analysis of such longitudinal data and a joint modeling approach that uses some latent variables to characterize the correlations is presented.

email: sunj@missouri.edu

## ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE DROPOUT TIME FROM AN EXTENDED HAZARDS MODEL

Yi-Kuan Tseng, National Central University, Taiwan  
Meng Mao, University of California, Davis  
Jane-Ling Wang\*, University of California, Davis

In medical follow-up studies, patients may drop out of the study due to a disease related cause or may die during the study. Such events lead to informative dropout of the longitudinal measurements. An effective way to deal with this informative dropout is to model the longitudinal process jointly with the dropout process. In this talk, we model the dropout process with a new extended hazards model that includes both the Cox proportional hazards model and the accelerated failure time (AFT) model. We illustrate how to implement the joint modeling approach by maximizing a pseudo joint likelihood function where random effects from the longitudinal process are treated as missing data. A Monte Carlo EM algorithm is employed to estimate the unknown parameters, including the unknown nonparametric baseline hazard function of the dropout time. Identifiability issues and statistical inference will be discussed. One advantage of the extended hazards model is to facilitate model selection between the Cox and AFT model, for which we propose a nonparametric likelihood ratio test.

email: wang@wald.ucdavis.edu

## MARGINAL ANALYSIS OF LONGITUDINAL DATA WITH BOTH RESPONSE AND COVARIATES SUBJECT TO MISSINGNESS

Grace Y. Yi\*, University of Waterloo  
Baojiang Chen, University of Waterloo  
Richard Cook, University of Waterloo

Data from longitudinal studies often feature both missing responses and missing covariates. When the response is missing, the probability a particular covariate is missing is often higher, as a result of a positive association between the missingness for the response and covariates at each follow-up assessment. The impact of missing data in these settings depends on the frequency data are missing and the strength of the association among the missing data processes and response process. In the setting of incomplete response and covariate data, it is important to take the association between these processes into account when analysing data. Inverse probability weighted generalized estimating equations offer a method for doing this and we develop this here. Empirical results demonstrate that the proposed method yields

# ABSTRACTS

consistent estimators, and is more efficient than alternative methods which ignore the association between the missing data processes.

email: yyi@uwaterloo.ca

## SEMIPARAMETRIC REGRESSION ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE OBSERVATION TIMES

Jianguo (Tony) Sun, University of Missouri-Columbia  
Do-Hwan Park\*, University of Maryland-Baltimore County  
Liuquan Sun, Chinese Academy of Sciences  
Xingqiu Zhao, Hong Kong Polytechnic University

Statistical analysis of longitudinal data is an important topic faced in a number of applied fields including epidemiology, public health and medicine. In general, the information contained in longitudinal data can be divided into two parts. One is the set of observation times that can be regarded as realizations of an observation process and the other is the set of actually observed values of the response variable of interest that can be seen as realizations of a longitudinal or response process. For their analysis, a number of methods have been proposed and most of them assume that the two processes are independent. This greatly simplifies the analysis since one can rely on conditional inference procedures given the observation times. However, the assumption may not be true in some applications. We will consider situations where the assumption does not hold and propose a semiparametric regression model that allows the dependence between the observation and response processes. Inference procedures are proposed based on the estimating equation approach and the asymptotic properties of the method are established. The results of simulation studies will be reported and the method is applied to a bladder cancer study.

email: dopark@unr.edu

## 52. ADDRESSING KEY STATISTICAL ISSUES IN ENVIRONMENTAL EPIDEMIOLOGY

### MULTISTAGE SAMPLING FOR LATENT VARIABLE MODELS IN ENVIRONMENTAL AND GENETIC EPIDEMIOLOGY

Duncan C. Thomas\*, University of Southern California

Multistage sampling schemes will be discussed for latent variable models where surrogate measurements of the latent variables are made on a subset of subjects. Such models arise when detailed exposure measurements are combined with exposure predictors to assign exposures to unmeasured subjects, when biomarkers are used to assess an unobserved disease process, or in various other situations. Analytic calculations of the optimal design are possible when all variables are binary, when all are normally distributed, or when the latent variable and its measurement are normally distributed but the outcome is binary. In these scenarios, it is often possible to improve the cost-efficiency of the design considerably by appropriate selection of the sampling fractions. More complex situations arise when exposures are spatially correlated. Applications to the Children's Health Study of the health effects of air pollution and candidate gene interactions will be used to illustrate sampling designs for an analysis involving

measurements of local variation in air pollution levels on one subsample of homes and of biomarkers of the oxidative stress and inflammatory pathways on an overlapping sample of individuals.

email: dthomas@usc.edu

## ADJUSTMENT UNCERTAINTY IN EFFECT ESTIMATION

Ciprian M. Crainiceanu\*, Johns Hopkins University

Often there is substantial uncertainty in the selection of confounders when estimating the association between an exposure and health. We define this type of uncertainty as 'adjustment uncertainty'. We propose a general statistical framework for handling adjustment uncertainty in exposure effect estimation for a large number of confounders, we describe a specific implementation, and we develop associated visualization tools. Theoretical results and simulation studies show that the proposed method provides consistent estimators of the exposure effect and its variance. We also show that, when the goal is to estimate an exposure effect accounting for adjustment uncertainty, Bayesian model averaging with posterior model probabilities approximated using information criteria can fail to estimate the exposure effect and can over- or underestimate its variance. We compare our approach to Bayesian model averaging using time series data on levels of fine particulate matter and mortality.

email: ccrainic@jhsph.edu

## BIAS AND SPATIAL SCALE IN MODELS WITH SPATIAL CONFOUNDING

Christopher J. Paciorek\*, Harvard School of Public Health

Increasingly, regression models are being used when residuals are spatially correlated. The spatial residual may be induced by an unmeasured confounder. In this setting, the hope is that models that account for the spatial structure will reduce or eliminate the bias from confounding. I show that regression models with spatial random effects and closely-related models such as kriging and penalized splines are biased. I provide analytic and simulation results showing how the bias depends on the spatial scales of the covariate of interest and the residual, with bias reduced substantially only when the scale of the covariate is small and that of the residual larger than the covariate. The use of fixed effects to capture residual spatial structure effectively reduces bias but with a bias-variance tradeoff. I illustrate the results with an example of the effects of black carbon traffic pollution on birthweight. Time provided, I will also discuss the effects of residual spatial correlation on the precision of the exposure estimator.

email: paciorek@hsph.harvard.edu

## **53. RECENT DEVELOPMENT OF QUANTILE REGRESSION METHODS FOR SURVIVAL DATA**

### **QUANTILE REGRESSION FOR DOUBLY CENSORED DATA**

Guixian Lin, University of Illinois  
Xuming He\*, University of Illinois  
Stephen Portnoy, University of Illinois

Quantile regression offers a semiparametric approach for analyzing data with possible heterogeneity. It is particularly powerful for censored responses, where the conditional mean functions are unidentifiable without strict parametric assumptions on the distributions. Recent work by Portnoy (2003) and by Peng and Huang (2008) demonstrated how the Kaplan-Meier estimator and the Nelson-Aaron estimator for the univariate samples can be generalized for estimating the conditional quantile functions with right censored data. We propose a new algorithm for quantile regression when the response variable is doubly censored. The algorithm distributes probability mass of each censored point to its left or right in a self-consistent manner, taking the idea of Turnbull (1976) to a broader platform. The algorithm is insensitive to starting values, and can be used to estimate a set of quantile functions with a small number of iterations. Computational and Asymptotic properties of the method will be summarized and issues concerning implementation and inference will be discussed. Extensions to more general forms of censoring will also be explored.

email: sportnoy@uiuc.edu

### **LOCALLY WEIGHTED CENSORED QUANTILE REGRESSION**

Huixia Judy Wang\*, North Carolina State University  
Lan Wang, University of Minnesota

Censored quantile regression offers a valuable supplement to Cox proportional hazards model for survival analysis. Existing work in the literature often requires stringent assumptions, such as unconditional independence of the survival time and the censoring variable or global linearity at all quantile levels. Moreover, some of the work uses recursive algorithms which makes it challenging to derive asymptotic normality. To overcome these drawbacks, we propose a novel locally weighted censored quantile regression approach. The new approach adopts the redistribution-of-mass idea and employs a local reweighting scheme. Its validity only requires conditional independence of the survival time and the censoring variable given the covariates, and linearity at the particular quantile level of interest. Our method leads to a simple algorithm that can be conveniently implemented with R software. Applying recent theory of M-estimation with infinite dimensional parameters, we rigorously establish the consistency and asymptotic normality of the proposed estimator. The proposal method is studied via simulations and the analysis of an acute myocardial infarction dataset.

email: wang@stat.ncsu.edu

## **QUANTILE REGRESSION WITH CENSORED DATA**

Yijian Huang\*, Emory University

Quantile regression has developed into a primary statistical methodology to investigate functional relationship between a response and covariates. This technique has a long history in econometric applications. More recently, it has also been advocated for the analysis of survival data to assess evolving covariate effects. However, with censored data, existing methods typically require strong assumptions on the censoring mechanism or entail complication-plagued algorithms. In this talk, I will discuss several recent advances on this problem and motivate a new estimation procedure based on a set of differential equations. A fairly efficient algorithm has been developed for the computation. The proposed estimator is uniformly consistent and converges weakly to a Gaussian process. Inference procedures are suggested. Simulations show good numerical and statistical performance. The proposal is illustrated in the application to a clinical study.

email: yhuang5@emory.edu

### **COMPETING RISKS QUANTILE REGRESSION**

Limin Peng\*, Rollins School of Public Health, Emory University  
Jason P. Fine, University of North Carolina at Chapel Hill

Quantile regression has emerged as a significant extension of traditional linear models and its potential in survival applications has recently been recognized. In this paper we study quantile regression with competing risks data, formulating the model based on conditional quantiles defined using the cumulative incidence function, which includes as a special case an analog to the usual accelerated failure time model. The proposed competing risks quantile regression model provides meaningful physical interpretations of covariate effects and moreover relaxes the constancy constraint on regression coefficients, thereby providing a useful, perhaps more flexible, alternative to the popular subdistribution proportional hazards model. We derive an unbiased monotone estimating equation for regression parameters in the quantile model. The uniform consistency and weak convergence of the resulting estimators are established across a quantile continuum. We develop inferences, including covariance estimation, second-stage exploration, and model diagnostics, which can be stably implemented using standard statistical software without involving smoothing or resampling. Our proposals are illustrated via simulation studies and an application to a breast cancer clinical trial.

email: lpeng@sph.emory.edu

# ABSTRACTS

## 54. ADAPTIVE DESIGN IN CLINICAL TRIALS

### ADAPTIVE GROUP SEQUENTIAL DESIGN IN CLINICAL TRIALS WITH CHANGING PATIENT POPULATIONS

Huaibao Feng\*, Johnson & Johnson  
Qing Liu, Johnson & Johnson  
Jun Shao, University of Wisconsin-Madison

Standard group sequential test assumes that the treatment effects are homogeneous over time. In practice, however, the assumption may be violated. Often, this occurs when treatment effects are heterogeneous in patients with different prognostic groups, which are not evenly distributed over the time course of the group sequential trial. In this talk, we consider a setting where the inclusion/exclusion criteria for patient entry are relaxed at interim analyses. This triggers heterogeneous treatment effects over the enlarged patient population. In particular, we assume that the population change relates to some baseline covariates. We propose a set of linear regression models. With these models, we make inference on the target population based on additional data from the changed populations. The effect of the changing patient population on the available information is studied and the group sequential boundary values are adjusted accordingly in order to control overall type I error. We propose to carry out the group sequential procedure with these revised boundary values. Simulation results show that the type I error probability of this procedure is close to the desired level when the patient populations are changing across stages. Results on the power of our proposed group sequential design are also presented.

email: hfeng12@its.jnj.com

### ADAPTIVE PENALIZED D-OPTIMAL DESIGNS FOR DOSE FINDING FOR CONTINUOUS BIVARIATE OUTCOMES

Krishna Padmanabhan\*, Wyeth Research  
Francis Hsuan, Temple University  
Vladimir Dragalin, Wyeth Research

When a new drug is under development, a conventional dose-finding study involves learning about the dose-response curve in order to bring forward right doses of the drug to late-stage development. We propose an alternative adaptive design for dose-finding in clinical trials in the presence of both (continuous) efficacy and toxicity endpoints. We use the principles of optimal experimental designs for bivariate continuous endpoints. However, instead of using the traditional D-optimal design, which favors collective ethics but neglects the individual ethics, we consider the penalized D-optimal design that achieves an appropriate balance between the efficient treatment of patients in the trial and the precise estimation of the model parameters to be used in the identification of the target dose. We also show how to incorporate these penalty functions into the D-optimality criteria to build penalized optimal designs. This is compared with the traditional fixed allocation design in terms of allocation of subjects and precision of the identified dose-response curve and selection of the target dose.

email: padmans@wyeth.com

### DOSE FINDING BY JOINTLY MODELING TOXICITY AND EFFICACY AS TIME-TO-EVENT OUTCOMES

Ying Yuan, M. D. Anderson Cancer Center  
Guosheng Yin\*, M. D. Anderson Cancer Center

In traditional phase I and II clinical trial designs, toxicity and efficacy are often modeled as binary outcomes. This method ignores information on when the outcome event occurs (experiencing toxicity or achieving cure/remission), and also has difficulty accommodating a high accrual rate under which toxicity and efficacy cannot be observed timely, which results in treatment assignment delays. To address these issues, we propose a Bayesian adaptive phase I/II design that jointly models toxicity and efficacy as time-to-event outcomes. At each decision-making time, patients who have not experienced toxicity or efficacy are naturally censored. We apply the marginal cure rate model to explicitly account for patients insusceptible to efficacy due to drug resistance. The correlation between the bivariate time-to-toxicity and -efficacy outcomes is properly adjusted through the Clayton model. After screening out the excessively toxic or futile doses, we adaptively assign each new patient to the most appropriate dose based on the ratio of the areas under the predicted survival curves corresponding to toxicity and efficacy. We conducted extensive simulation studies to examine the operating characteristics of the proposed method. Our design selects the target dose with a high probability and treats most patients at the desirable dose.

email: gsyin@mdanderson.org

### BAYESIAN ADAPTIVE RANDOMIZATION DESIGNS VERSUS FREQUENTIST DESIGNS FOR TARGETED AGENT DEVELOPMENT

Xuemin Gu\*, M.D. Anderson Cancer Center  
Suyu Liu, M.D. Anderson Cancer Center  
Jack J. Lee, M.D. Anderson Cancer Center

With the advent of searching for predictive markers in guiding targeted agent development, many new designs have been proposed recently to increase study efficiency. These include the efficient targeted design,<sup>1</sup> adaptive signature design,<sup>2</sup> marker by treatment interaction design,<sup>3</sup> and biomarker adaptive threshold design,<sup>4</sup> etc. In contrast to the frequentist selection or equal randomization designs, we had reported a novel Bayesian adaptive randomization design to allow treating more patients with effective treatments, stopping ineffective treatments early, and increasing efficiency while controlling type I and type II errors.<sup>5</sup> Similar Bayesian designs were proposed in this study. The new designs incorporate rational learning from the interim data to guide the study conduct in terms of randomization. By comparing with previously published designs, the proposed design can be efficient and ethical and is also more flexible in the study conduct. The statistical properties for various designs are evaluated via simulation studies.

email: xuegu@mdanderson.org

## **STOPPING BOUNDARIES OF FLEXIBLE SAMPLE SIZE DESIGN WITH FLEXIBLE TRIAL MONITORING - A UNIFIED APPROACH**

Yi He, Sanofi-aventis  
Zhenming Shun, Sanofi-aventis  
Yijia Feng\*, Penn State University

In the group sequential (GS) approach with a fixed sample size design, the type I error is controlled by the additivity of exit spending values. However, in a flexible sample size design where the sample size will be re-calculated using the interim data, the overall type I error rate can be inflated. Therefore, the pre-defined the GS stopping boundaries have to be adjusted to maintain the type I error level at each interim analysis and the overall level. The modified  $\pm$ - spending function adjusted for sample size re-estimation (SSR) is proposed to maintain the type I error level. We use unified approach and mathematically quantify the type I error with and without sample size adjustment constraints. As a result, stopping boundaries can be obtained by inversely solving the exact type I error functions. This unified approach, using Brownian motion theory, can be applied to normal, survival, and binary endpoints. Extensive simulations show the stopping boundaries can control the type I error at each analysis and the overall level.

email: yijia@psu.edu

## **A SURROGATE: PRIMARY REPLACEMENT ALGORITHM FOR RESPONSE-ADAPTIVE RANDOMIZATION IN STROKE CLINICAL TRIALS**

Amy Nowacki\*, Cleveland Clinic Foundation

Response-adaptive randomization (RAR) offers clinical investigators ethical benefit by modifying the treatment allocation probabilities to assign more participants to the putatively better performing treatment. A misconception surrounding RAR is that it is limited to clinical trials where the primary outcome is obtained instantaneously. Delayed primary outcomes and their effect on RAR have been studied in the literature; however, the incorporation of surrogate outcomes has not been studied. We explore the benefits and limitations of surrogate outcome utilization in RAR in the context of acute stroke clinical trials. We propose a novel surrogate-primary (S-P) replacement algorithm method where the parameter estimates are based on the surrogate outcomes only until the primary outcome becomes available. A simulation study investigates the effects of: (1) the enrollment period; (2) the underlying population distribution of the surrogate and primary outcomes; and (3) the correlation among the surrogate and primary outcomes. When the primary outcome is delayed, the S-P replacement algorithm method outperforms standard RAR by reducing the variability of the treatment allocation probabilities and stabilizing the treatment allocation sooner. The enrollment period affects the degree to which the S-P replacement algorithm method has an advantage; with a short enrollment period showing significant advantage which decreases as the enrollment period lengthens. The correlation affects the rate of convergence to the target allocation which is of importance in sequential trials. To illustrate, we apply the proposed method to the NINDS rt-PA Stroke Study data set.

email: nowacka@ccf.org

## **BAYESIAN NONPARAMETRIC Emax MODEL**

Haoda Fu\*, Eli Lilly and Company

NDLM (Normal Dynamic Linear Model) has been used widely in Bayesian adaptive design. Our new method improved the NDLM in terms of its ability to control the shape of the curve, estimating Emax, ED50, and obtaining a monotone smoothed dose response curve. This method can be looked at as a monotone nonparametric curve fitting method. Several examples on using this new method on adaptive design will be provided.

email: fuhaoda@gmail.com

## **55. GENE SELECTION IN DNA MICROARRAY STUDIES**

### **AN INTEGRATIVE ANALYSIS APPROACH FOR IDENTIFICATION OF GENES ASSOCIATED WITH MULTIPLE CANCERS**

Shuangge Ma\*, Yale University

Genomic markers identified from expression studies can be used to improve diagnosis, prognosis and prediction in cancer clinical studies. Biomedical studies have suggested that development of multiple cancers may share common genomic basis. A complete description of the associations between genes and cancers amounts to identification of not only multiple genes associated with a single type of cancer, but also multiple cancers that a specific gene is associated with. For such a purpose, we propose an integrative analysis approach capable of analyzing multiple cancer microarray studies conducted on different cancers. The proposed approach is the first regularized approach to conduct "two-dimensional" selection of genes in the joint modeling of multiple gene effects. Using the proposed approach, we analyze seven microarray studies investigating development of seven different types of cancers. Genes associated with one or more of the seven cancers are identified. Many identified genes have sound biological basis.

email: shuangge.ma@yale.edu

## **ASSOCIATION PATTERN TESTING: A POWERFUL STATISTICAL TOOL TO IDENTIFY BIOLOGICALLY INTERESTING GENOMIC VARIABLES**

Stanley B. Pounds\*, St. Jude Children's Research Hospital  
Cheng Cheng, St. Jude Children's Research Hospital  
Xueyuan Cao, St. Jude Children's Research Hospital  
James R. Downing, St. Jude Children's Research Hospital  
Raul C. Ribeiro, St. Jude Children's Research Hospital  
Kristine R. Crews, St. Jude Children's Research Hospital  
Jatinder Lamba, University of Minnesota

The association pattern test (APT) is proposed as a general procedure to identify genomic variables that exhibit a specific biologically interesting pattern of association with multiple phenotype variables. Prior biological knowledge is used to specify a pattern of interest. Next, the pattern of interest is used to define a statistic that measures the evidence that a genomic variable exhibits this pattern of

# ABSTRACTS

association with the phenotypes. Statistical significance is determined via permutation. In contrast to classical multivariate procedures, APT can easily handle different types of phenotype variables (such as categorical and survival-type endpoints). Simulation studies show that APT has greater power to detect the specified biologically interesting pattern than other methods. In an example application, APT is used to identify genes with expression levels that show an interesting pattern of association with two pharmacokinetic endpoints, two pharmacodynamic endpoints, and three clinical endpoints. A number of biologically interesting genes were identified that would otherwise have been overlooked, including oncogenes, regulators of pharmacologically relevant genes, and cell-cycle genes.

email: stanley.pounds@stjude.org

## INCORPORATING GENE EFFECTS INTO PARAMETRIC EMPIRICAL BAYES METHODS FOR MICROARRAYS

Steven P. Lund\*, Iowa State University  
Dr. Dan Nettleton, Iowa State University

The log-normal normal (LNN) and gamma-gamma (GG) models (Kendziorski CM, Newton MA, Lan H, et al., (2003), *Statistics in Medicine*, 22:3899-3914) are two popular parametric empirical Bayes approaches used to identify differentially expressed genes among multiple treatment groups from gene expression profiles. An assumption of these models is that observed expression levels between treatment groups with different means are independent, even within genes. If false, this assumption will bias conclusions towards the null hypothesis of equivalent expression. We introduce a gene effect to form a log-normal normal model and show that it identifies differential expression as well as or better than the LNN and GG models through a variety of simulation studies.

email: lunds@iastate.edu

## NETWORK-BASED SUPPORT VECTOR MACHINE FOR CLASSIFICATION OF MICROARRAY SAMPLES

Yanni Zhu\*, Xiaotong Shen, University of Minnesota  
Wei Pan, University of Minnesota

The importance of network-based approach to identifying biological markers for diagnostic classification and prognostic assessment in the context of microarrays has been increasingly recognized. To our knowledge, there have been few, if any, statistical tools that explicitly incorporate the prior information of gene networks into classifier building. The main idea of this paper is to take full advantage of the biological observation that neighboring genes in a network tend to function together in biological processes and to embed this information into a formal statistical framework. We propose a network-based support vector machine for binary classification problems by constructing a penalty term from the F-infinity norm being applied to pairwise gene neighbors with the hope to improve predictive performance and gene selection. Simulation studies in both low- and high-dimensional data settings as well as two real microarray applications indicate that the proposed method is able to identify more clinically relevant genes while maintaining a sparse model

with either similar or higher prediction accuracy compared with the standard and the L1 penalized support vector machines. The proposed network-based support vector machine has the potential to be a practically useful classification tool for microarrays.

email: zhux0130@umn.edu

## DualKS: DEFINING GENE SETS WITH TISSUE SET ENRICHMENT ANALYSIS

Eric J. Kort, Van Andel Research Institute  
Yarong Yang\*, Northern Illinois University  
Zhongfa Zhang, Van Andel Research Institute  
Bin Teh, Van Andel Research Institute  
Nader Ebrahimi, Northern Illinois University

Gene set enrichment analysis (GSEA) is an analytic approach which simultaneously reduces the dimensionality of microarray data and enables ready inference of the biological meaning of observed gene expression patterns. Here we invert the GSEA process to identify class-specific gene signatures enabling tissue diagnosis. This can be conceptualized as “tissue-set enrichment analysis”. Because our approach uses the KS approach both to define class specific signatures and to classify samples using those signatures, we have termed this methodology “Dual-KS” (DKS). The optimum gene signature identified by the DKS algorithm was smaller than other methods to which it was compared in 5 out of 10 datasets. The estimated error rate of DKS using the optimum gene signature was smaller than the estimated error rate of the random forest method in 4 out of the 10 datasets, and was equivalent in two additional datasets. DualKS outperformed other benchmarked algorithms to a similar degree. DKS is an efficient analytic methodology that can identify highly parsimonious gene signatures useful for classification in the context of microarray studies.

email: yarongyang78@yahoo.com

## EVALUATION OF A CLASSIFIER PERFORMANCE AT VARIOUS CUTOFFS OF GENE SELECTION IN MICROARRAY DATA WITH TIME-TO-EVENT ENDPOINT

Dung-Tsa Chen\*, Moffitt Cancer Center & Research Institute, University of South Florida  
Ying-Lin Hsu, National Chung Hsing University, Taichung, Taiwan  
Tzu-Hsin Liu, National Chung Hsing University, Taichung, Taiwan  
James J. Chen, National Center for Toxicological Research, U.S. Food and Drug Administration  
Timothy Yeatman, Moffitt Cancer Center & Research Institute, University of South Florida

Statistical methods have been widely used to analyze microarray data with time-to-event endpoint by reducing gene expression data from a high-dimensional space to a manageable low-dimensional space for the follow-up survival analysis. Basically, the methods start with exploring an optimal cutoff to subset genes. The selected cutoff will lead to determine the number of genes (and which genes). With the cutoff, we can employ survival analysis methods to build a classifier for prediction purpose. Often we do not have the luxury to have a

test dataset ready for validation. To evaluate how good the classifier is, random split scheme has been used to divide the whole data into training and test sets and repeat the process many times for evaluation. In this study, we examine performance of a classifier at various cutoffs by utilizing the random split scheme. We compare performance of a classifier in the test set at a pre-specified optimal cutoff versus other cutoffs. We also examine correlation of performance of the classifier at the training set versus at the test set. Several random split schemes are also compared for their impact on performance. A set of real data and simulation data are used for illustration.

email: dung-tsa.chen@moffitt.org

## DOES A GENE EXPRESSION CLASSIFIER HAVE A CLINICAL VALUE?

Samir Lababidi\*, U.S. Food and Drug Administration

The early papers in microarray gene expression promised better prediction in cancer outcome classification than what have been done before for prognostic and diagnostic medicine. However, it became clear later that the actual gain in predictive ability due to the use of gene expression classifiers may have been sometimes exaggerated and in need of careful evaluation. In fact, a statistical significant classification rate obtained by, e.g., validation in an independent dataset, does not necessarily imply that the classifier would have the clinical value of interest. Here in this talk, we will present statistical analysis methods to establish the clinical value of a classifier and compare its performance to models using clinical covariates alone.

email: samir.lababidi@fda.hhs.gov

## 56. IMAGE ANALYSIS

### ON THE MERITS OF VOXEL-BASED MORPHOMETRIC PATH-ANALYSIS FOR INVESTIGATING VOLUMETRIC MEDIATION OF A TOXICANT'S INFLUENCE ON COGNITIVE FUNCTION

Shu-chih Su,\* Merck & Co., Inc.

Brian Caffo, Johns Hopkins Bloomberg School of Public Health

Previous study showed that lifetime cumulative lead dose was associated with persistent and progressive declines in cognitive function and with decreases in MRI-based brain volumes in former lead workers. Moreover, larger region-specific brain volumes were associated with better cognitive function. These findings motivated us to explore a novel application of path analysis to evaluate effect mediation, whether the association of lead dose with cognitive function is mediated through brain volumes, on a voxel-wise basis. Application of these methods to the former lead worker data demonstrated potential limitations in this approach. Moreover, a complimentary analysis using anatomically-derived regions of interest (ROI) volumes yielded opposing results, suggesting evidence of mediation. In the ROI-based approach, there was evidence that the association of tibia lead with function in three cognitive domains was mediated through the volumes of total brain, frontal gray matter, and possibly cingulate. A simulation study was conducted to investigate

whether the voxel-wise results arose from an absence of localized mediation, or more subtle defects in the methodology. Both the lead worker data results and the simulation study suggest that a null-bias in voxel-wise path analysis limits its inferential utility for producing confirmatory results.

email: shsu@jhsph.edu

## INTRINSIC REGRESSION MODELS FOR MEDIAL REPRESENTATION OF SUBCORTICAL STRUCTURES

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  
Faming Liang, Texas A&M University  
Martin Styner, University of North Carolina at Chapel Hill

The aim of this paper is to develop a statistical framework for describing the variability of the medial representation (m-rep) of subcortical structures and its association with covariates in Euclidean space. Because an m-rep does not form a vector space, applying classical multivariate regression techniques may be inadequate in establishing the association between an m-rep and covariates of interest in real applications. Our proposed regression model as a semiparametric model avoids specifying a probability distribution on a Riemannian manifold. We develop an estimation procedure based on the annealing evolutionary stochastic approximation Monte Carlo (AESAMC) algorithm to obtain parameter estimates and establish their limiting distributions. We use Wald statistics to test linear hypotheses of unknown parameters. Simulation studies are used to evaluate the accuracy of our parameter estimates and examine the finite sample performance of the Wald statistics. We apply our methods to the detection of the difference in the morphological changes of the left and right hippocampi between schizophrenia patients and healthy controls using medial shape description.

email: xyshi@email.unc.edu

## MULTISCALE ADAPTIVE REGRESSION MODELS FOR IMAGING DATA

Yimei Li\*, University of North Carolina-Chapel Hill  
Hongtu Zhu, University of North Carolina-Chapel Hill  
Joseph G. Ibrahim, University of North Carolina-Chapel Hill  
Dinggang Shen, University of North Carolina-Chapel Hill

We develop a multiscale adaptive regression model (MARM) for spatial and adaptive analysis of imaging data. The primary motivation and application of the proposed methodology is statistical analysis of imaging data on the two-dimensional (2D) surface or in the 3D volume for neuroimaging studies. The key idea of the MARM is to successively increase the radius of a spherical neighborhood around each voxel and combine all the data in a given radius of each voxel with appropriate weights to adaptively calculate parameter estimates and test statistics. We establish consistency and asymptotic normality of the adaptive estimates and the asymptotic distributions of the adaptive test statistics. Particularly, we show theoretically that the MARM outperforms classical voxel-wise approach. Simulation studies

# ABSTRACTS

are used to demonstrate the methodology and examine the finite sample performance of the MARM. We apply our methods to the detection of spatial patterns of brain atrophy in a neuroimaging study of Alzheimers disease. Our simulation studies and real data analysis demonstrate that the MARM significantly outperforms the voxel-wise methods.

email: liyimei@email.unc.edu

## CONNECTIVITY ANALYSIS BASED ON fMRI AND DTI BRAIN IMAGING DATA

Shuo Chen\*, Rollins School of Public Health, Emory University  
DuBois Bowman, Rollins School of Public Health, Emory University  
Gordana Derado, Rollins School of Public Health, Emory University

Recent development has shown promising advantage of using multimodality brain imaging data to build connectivity between regions of the human brain. However, it still remains challenging to combine different types of imaging data effectively and appropriately. One goal of our study is to stabilize the connectivity discovery from fMRI data by utilizing the information of tractography probability from DTI data. We proposed a novel approach based on mixture criteria of likelihood and stability of connectivity structure. The tuning parameter between likelihood and stability is determined by the variability of connectivity structure of different subjects.

email: schen33@emory.edu

## MODELING THE SPATIAL AND TEMPORAL DEPENDENCE IN fMRI DATA

Gordana Derado\*, Emory University

Functional neuroimaging is important for investigating behavior-related neural processing linked to substance abuse disorders and associated treatment interventions. Functional magnetic resonance imaging (fMRI) data sets are large and are characterized by complex dependence structures, driven by highly sophisticated neurophysiology and aspects of typical experimental designs. These complex dependence structures pose analytical challenges for statistical modeling. Typical analyses investigating task-related changes in measured brain activity proceed using a two stage procedure in which the first stage involves subject-specific models relating neural processing to experimental tasks and the second-stage specifies group (or population) level parameters. Customarily, the first-level accounts for temporal correlations between the serial scans acquired during one session or even under one experimental stimulus. Despite accounts for these correlations, fMRI studies often include multiple experimental conditions and/or sessions (e.g., before and after treatment), and temporal dependencies may persist between the corresponding estimates of mean neural activity. Further, spatial correlations between brain activity measurements in different locations are often not accounted for in statistical modeling and estimation. Bowman (2005) proposed an extended two-stage model for the estimation and testing of localized activity in which the second stage accounts for spatial dependencies between voxels within the same neural processing cluster (defined by a data-driven cluster analysis). This model, however, did not account for repeated measures type associations between

the multiple experimental effects for each subject. We propose an extended two-stage, spatio-temporal, simultaneous autoregressive model which accounts for both spatial dependencies between voxels within the same anatomical region and for temporal dependencies between the multiple experimental effects/sessions for a subject. We develop an algorithm that leverages the special structure of our covariance model, enabling relatively fast and efficient estimation. We apply our method to fMRI data from a cocaine addiction study.

email: gderado@emory.edu

## APPROXIMATION OF THE GEISSER-GREENHOUSE SPHERICITY ESTIMATOR AND ITS APPLICATION TO ANALYZING DIFFUSION TENSOR IMAGING DATA

Meagan E. Clement\*, Rho, Inc.  
David Couper, University of North Carolina-Chapel Hill  
Keith E. Muller, University of Florida  
Hongtu Zhu, University of North Carolina-Chapel Hill

Recent protocol innovation with magnetic resonance imaging has resulted in diffusion tensor imaging (DTI). The approach holds tremendous promise for improving our understanding of neural pathways, especially in the brain. The DTI protocol highlights the distribution of water molecules, in three dimensions. In a medium with free water motion, the diffusion of water molecules is expected to be isotropic, the same in all directions. With water embedded in nonhomogeneous tissue, motion is expected to be anisotropic, not the same in all directions, and might show preferred directions of mobility. DTI fully characterizes diffusion anisotropy locally in space, thus providing rich detail about tissue microstructure. However, little has been done to define metrics or describe credible statistical methods for analyzing DTI data. Our research addresses these research issues. First, we show that fractional anisotropy values for given regions of interest are functions of the Geisser-Greenhouse (GG) sphericity estimator. Next, we demonstrate that the GG sphericity estimator can be approximated by a squared beta distribution. Finally, noise is added to show these approximations also work for simulated diffusion tensors. Thus, using these approximate distributions, one can then avoid the "curse of dimensionality".

email: meagan\_clement@rhoworld.com

## 57. SURVEY RESEARCH

### BAYESIAN INFERENCE OF FINITE POPULATION DISTRIBUTION FUNCTIONS AND QUANTILES FROM UNEQUAL PROBABILITY SAMPLES

Qixuan Chen\*, University of Michigan School of Public Health  
Michael R. Elliott, University of Michigan School of Public Health  
Roderick J.A. Little, University of Michigan School of Public Health

This paper develops two robust Bayesian model-based estimators of finite population distribution functions and associated quantiles for continuous variables in the setting of unequal probability sampling, where inferences are based on the posterior predictive distribution of the non-sampled values. The first method fits a multinomial ordinal

probit regression model of the distribution function evaluated at multiple values on a penalized spline of the selection probabilities. Finite population quantiles are then obtained by inverting the distribution function. However, heavy computation is involved in inverting the distribution function; therefore we consider the second method that posits a smoothly-varying relationship between the continuous outcome and the selection probabilities by modeling both the mean function and the variance function using penalized splines. Simulation studies show that both methods yield estimators that are more efficient with closer to the nominal level credible intervals than the design-based estimators.

email: qixuan@umich.edu

## **APPLICATION OF NONPARAMETRIC PERCENTILE REGRESSION TO BODY MASS INDEX PERCENTILE CURVES FROM SURVEY DATA**

Yan Li\*, National Cancer Institute  
 Barry I. Graubard, National Cancer Institute  
 Edward L. Korn, National Cancer Institute

Increasing rates of overweight among children in the U.S. stimulated interest in obtaining national percentile curves of body size to serve as a benchmark in assessing growth development in clinical and population settings. In 2000, the Centers for Disease Control and Prevention (CDC) developed conditional percentile curves for Body mass index (BMI) for ages 2-20 years. The 2000 CDC BMI-for-age curves are partially parametric and only partially incorporated the survey sample weights in the curve estimation. As a result, they may not fully reflect the underlying pattern of BMI-for-age in the population. This motivated us to develop a nonparametric double-kernel-based method and automatic bandwidth selection procedure. We include sample weights in the bandwidth selection, conduct median correction to reduce small-sample smoothing bias, and rescale the bandwidth to make it scale-invariant. Using this procedure we re-estimate the national percentile BMI-for-age curves and the prevalence of high-BMI children in the U.S.

email: lisherry@mail.nih.gov

## **OPTIMAL COEFFICIENTS FOR SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT USING GODAMBE'S GENERAL LINEAR ESTIMATOR**

Ruitao Zhang\*, Department of Public Health, University of Massachusetts-Amherst  
 Ed Stanek, Department of Public Health, University of Massachusetts-Amherst

Godambe (1955) proved that a best linear unbiased estimator does not exist based on sampling. Recently, we have developed unique optimal coefficients for estimating the population total when sampling without replacement of and using Godambe's general linear class of estimators. The solution to the estimating equations requires that all subject's values are not equal to zero. We discuss solutions to this problem, and extend solutions to the populations with one or two zero values or ties when sampling without replacement of and. Other extensions are developed to the problem where n=3 from N=4 and n=4 from N=5

following the same methods with no zeros. Then we will generalize to N-1 from N. The estimator will also be extended to include values with at least one zero values or ties. This basic research is important since it explores conceptual issues in formally integrating sampling into statistical inference, identifying promising avenues to be explored.

email: ruitaozhang@hotmail.com

## **SHOULD AUXILIARY VARIABLES WITH MEASUREMENT ERROR BE USED IN THE ESTIMATION OF POPULATION MEAN BASED ON SURVEY SAMPLES?**

Wenjun Li\*, University of Massachusetts Medical School  
 Edward J. Stanek III, Department of Public Health, University of Massachusetts-Amherst

Auxiliary information is commonly used to improve the precision of estimators of population quantities in sample surveys. However, auxiliary variables may be measured with error. We extend the random permutation model proposed by Stanek, Singer and Lencina (2004) to obtain best linear unbiased estimators of a finite population mean in situations where auxiliary information is available but measured with error in the sample under simple random without replacement sampling (SRS). The variance of the estimator is inflated by measurement error, in particular among scenarios where the correlation coefficient between the outcome and auxiliary covariates are relatively strong and the sampling fraction is high. For a modest reliability of auxiliary variates (0.75) and a strong correlation, the variance inflation is nearly 2.5 fold. When compared to simple estimator, the potential variance reduction due to adjustment diminishes quickly with lower reliability of the auxiliary variable. When reliability is lower than 0.5, meaningful variance reduction is unlikely even when the correlation between outcome and the auxiliary variable is strong ( $\rho=0.9$ ), in particular when sampling fraction is high. The results of this analysis are used to guide the choice of auxiliary variables in the estimation of community walkability in a study of Boston neighborhoods.

email: Wenjun.Li@umassmed.edu

## **PROXY PATTERN-MIXTURE ANALYSIS FOR SURVEY NONRESPONSE**

Rebecca R. Andridge\*  
 Roderick J. Little, University of Michigan

We consider assessment of nonresponse bias for the mean of a survey variable Y subject to nonresponse. We assume that there are a set of covariates observed for nonrespondents and respondents. To reduce dimensionality and for simplicity we reduce the covariates to a proxy variable X that has the highest correlation with Y, estimated from a regression analysis of respondent data. We consider adjusted estimators of the mean of Y that are maximum likelihood for a pattern-mixture model with different mean and covariance matrix of Y and X for respondents and nonrespondents, assuming missingness is an arbitrary function of a known linear combination of X and Y. We propose a taxonomy for the evidence concerning bias based on the strength of the proxy and the deviation of the mean of X for respondents from

# ABSTRACTS

its overall mean, propose a sensitivity analysis, and describe Bayesian versions of this approach. We propose using the fraction of missing information from multiple imputation under the pattern-mixture model as a measure of nonresponse bias. Methods are demonstrated through simulation and data from the third National Health and Nutrition Examination Survey (NHANES III).

email: fedarko@umich.edu

## MULTIPLE IMPUTATION METHODS FOR DISCLOSURE LIMITATION IN LONGITUDINAL DATA

Di An\*, Merck Research Laboratories, Merck & Co., Inc.  
Roderick J.A. Little, University of Michigan  
James W. McNally, University of Michigan

Disclosure limitation is an important consideration in the release of public use data sets. It is particularly challenging for longitudinal data sets, since information about an individual accumulates over time. We consider problems created by high ages in cohort studies. Because of the risk of disclosure, ages of very old respondents can often not be released, as stipulated by the Health Insurance Portability and Accountability Act (HIPAA). Top-coding of individuals beyond a certain age is a standard way of dealing with this issue, but it has severe limitations in longitudinal studies. We propose and evaluate an alternative to top-coding for this situation based on multiple imputation (MI). This MI method is applied to a survival analysis of simulated data and data from the Charleston Heart Study, and is shown to work well in preserving the relationship between hazard and covariates.

email: di\_an@merck.com

## IMPACT OF MULTI-LEVEL MEASUREMENT ERRORS IN SURVEY DATA

Jianjun Gan\*, School of Public Health, University of South Carolina  
Hongmei Zhang, School of Public Health, University of South Carolina

It is widely accepted that both environmental and genetic factors are related to the causation of NTDs (neural tube defects). Many Studies have been contributed to this area and demonstrated the effect of Folic Acid from daily supplement and the effect folate from food on NTD risk reduction. Recent studies also provided evidence of potential contribution from other micronutrients, for instances, dietary botaine and myo-inositol. However, the results vary dramatically from one study to another. Although factors related to experimental designs may account for some of the variations, we expect that much of the variation is at least in part due to measurement errors, because survey questionnaires are usually the only instrument used in these studies and responses to survey questions are likely to be biased. In survey questionnaires, typically there are two possible levels of measurement errors. One level is formed by participants' recall bias when filling the questionnaires, and the other level is when a system summarizes the responses. Unfortunately, little attention has been paid on multi-level measurement error modeling. In this paper, we develop a multi-level measurement error model and further incorporate it into a Logistic

regression model. Simulations are used to demonstrate the methods, evaluate the impact of measurement errors on the inferences of factor effects , and study the effect of different choices of random effect distributions. Finally, we apply the method to data sets of CDC questionnaires and FFQ to evaluate the impact of measurement errors at different levels during the data collection process. We expect the findings will self-explain the importance of adjusting for measurement errors and thus benefit future data collection effort.

email: ganj@mailbox.sc.edu

## 58. MEASUREMENT ERROR MODELS

### DICHOTOMIZED MISMEASURED PREDICTORS IN REGRESSION MODELS

Loki Natarajan\*, University of California at San Diego

In epidemiologic studies, interest focuses on estimating exposure-disease associations. In some cases, a continuous exposure (e.g., intake of dietary fat) may be dichotomized (e.g., fat intake >30% of calories) if threshold levels of the predictor are of primary public health interest. Exposure-disease risk estimates will be biased when the exposure is mismeasured. In order to correct for these biases, a validation substudy may be conducted where the "true" and imprecise exposures are observed on a small random subsample. Regression models then incorporate validation substudy data, to obtain less biased exposure-disease risk estimates. In this presentation, we will focus on biases associated with dichotomization of a mismeasured continuous exposure. The amount of bias in relation to mismeasurement in the imprecise continuous predictor, and choice of dichotomization cut-point will be discussed. Measurement error correction methods will be developed for this scenario in the validation substudy setting, and compared to naively using the dichotomized mismeasured predictor in exposure-disease models. Properties of the measurement error correction methods (i.e., bias, mean-squared error) will be presented. The proposed methods will be applied to data on blood pressure and dietary intake collected as part of an ongoing epidemiologic study.

email: lnatarajan@ucsd.edu

### AUGMENTING INSTRUMENTAL VARIABLES ESTIMATORS IN A TWO-STAGE DESIGN

Tor D. Tosteson\*, Dartmouth Medical School

We consider an example of an epidemiologic study relating two continuous exposure measurements for arsenic, toenail (T) and tap water (W) concentrations, to a continuous measure of gene expression serving as a potential intermediate marker of cancer incidence. Because of budget concerns, gene expression has been evaluated on a relatively small subsample of controls in a large case-control study which has toenail and water concentrations for most participants. Both exposure measurements are subject to measurement error, but the toenail is thought to be unbiased for the true biological exposure. In a study in which one observes (Y,W,T) and a simple linear model applies, the instrumental variables method of moments estimator for the slope coefficient can be expressed as  $\text{cov}(Y,W)/\text{cov}(T,W)$ . Because there is only a small sample, it appears that a more accurate estimate might be

obtained by augmenting the data available for determining  $\text{cov}(T, W)$  with data from the main study. We compare the performance of the ordinary instrumental variable estimate, the augmented instrumental variable estimate, and a maximum likelihood estimate under normality assumptions with asymptotic methods and small sample simulations. The results are applied to data from the arsenic case-control study.

email: tor.tosteson@dartmouth.edu

## REGRESSION ANALYSIS ON A COVARIATE WITH HETEROSEDASTIC MEASUREMENT ERROR

Ying Guo\*, University of Michigan  
Roderick Little, University of Michigan

We consider the problem of estimating the regression of an outcome D on a covariate X, where X is unobserved, but a variable Y which measures X with error is observed. A calibration sample that measures pairs of values of X and Y is also available; we also consider the case where the calibration sample includes values of D. The standard approach for estimating the regression of D on X is to estimate the calibration curve of X on Y and then replace unknown values of X by predictions from this curve. An alternative approach is to multiply impute the missing values of X given Y and D based on an imputation model, and then use multiple imputation (MI) combining rules for inferences about the coefficients of the regression of D on X. A recent paper by Friedman et al (2008) compares these two approaches. However, their work assumes the measurement error of Y has a constant variance, whereas in many situations, the measurement error variance varies as a function of X. We consider modifications of the calibration prediction method and the multiple imputation method that allow for non-constant measurement error variance, and compare these methods by simulation. The multiple imputation model is shown to provide better inferences in this setting.

email: guoy@umich.edu

## COX MODELS WITH SMOOTH FUNCTIONAL EFFECT OF COVARIATES MEASURED WITH ERROR

Yu-Jen Cheng\*, Johns Hopkins University  
Ciprian Crainiceanu, Johns Hopkins University

We propose, develop and implement a fully Bayesian inferential approach for the Cox model when the log hazard function contains unknown smooth functions of the variables measured with error. Our approach is to model nonparametrically both the log-baseline hazard and the smooth components of the log-hazard functions using low-rank penalized splines. The likelihood of the Cox model is coupled with the likelihood of the measurement error process. Careful implementation of the Bayesian inferential machinery is shown to produce remarkably better results than the naive approach. Our methodology was motivated by and applied to the study of progression time to chronic kidney disease (CKD) as a function of baseline kidney function and applied to the Atherosclerosis Risk in Communities (ARIC) study, a large epidemiological cohort study.

email: ycheng3@jhsph.edu

## UNDERREPORTING IN THE GENERALIZED POISSON REGRESSION MODEL

Mavis Pararai\*, Indiana University of Pennsylvania

The generalized Poisson regression model has been used to model equi-over- and under-dispersed count data. In many of these situations the assumption is that the response, the count, is reported without error. It is possible that the count maybe underreported or overreported. The Poisson regression model and the negative binomial regression model have been modified and used in modeling count data that is underreported. In this paper, the generalized Poisson regression model for underreported counts is developed. The parameters of the proposed model are estimated by the maximum likelihood method. We propose a score test to determine whether there is significant underreporting in the data in order to use of the generalized Poisson regression model for underreported counts as opposed to the ordinary generalized Poisson regression model. The generalized Poisson regression model is applied to data on number of sexual partners.

email: pararaim@iup.edu

## MEASUREMENT ERROR IN LONGITUDINAL DATA WITHOUT VALIDATION SAMPLES

Ruifeng Xu\*, Merck & Co., Inc.  
Jun Shao, University of Wisconsin-Madison  
Mari Palta, University of Wisconsin-Madison  
Zhiguo Xiao, School of Management, Fudan University

Measurement error in covariates has received considerable attention in the biostatistics literature. Longitudinal data allow correction for errors in covariates in linear models, even when true replicate measurements or validation samples are not available. Wansbeek (2001) proposed a generalized method of moments (GMM) framework for analyzing longitudinal data with measurement error in a single regressor. We generalize his method to the case of uneven length of follow-up and the case where more than one covariate are measured with errors. The methods are applied to the Wisconsin Sleep Cohort Study.

email: xu\_ruifeng@yahoo.com

## MODELING HEAPING IN LONGITUDINAL 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. Heaping in daily reported cigarette counts compounds the problem by affecting estimates of change in consumption in addition to marginal mean counts. We present a model to describe longitudinal heaped count data. 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 latent variables that are at best partially

# ABSTRACTS

observed. To account for correlations in longitudinal cigarette consumption, our method specifies separately the correlation of true cigarette smoking and the reporting behavior within a subject, but models them simultaneously.

email: haow@mail.med.upenn.edu

## 59. MIXTURE MODELS

### CONFOUNDING AND BIAS FROM INTERMEDIATE VARIABLES, AND A JOINT MODEL FOR BIRTHWEIGHT AND GESTATIONAL AGE ADDRESSING THEM

Scott L. Schwartz\*, Duke University  
Alan E. Gelfand, Duke University  
Marie L. Miranda, Duke University

A frequently studied birth outcome is birthweight (BW), evidently, strongly associated with gestational age (GA). Customary modeling for BW is conditional on GA. However, adjusting for intermediate variables (1) entails the loss of “causal effect” estimation and (2) may produce “spurious effect estimates” due to back-door criterion violation. Joint modeling provides an alternative means to address the relationship between BW and GA while at the same time offering increased flexibility and interpretation in studying these outcomes. We introduce a mixture of bivariate regressions’ model for the joint distribution of BW and GA which, (1) enables us to capture the clearly nonGaussian distributional shapes observed in histograms, (2) addresses the interval censoring commonly associated with gestational age (to the nearest completed week) through a latent specification, and (3) facilitates interpretation by providing a bivariate regression structure for well-established risk factors. This approach explicitly avoids adjusting for the intermediate variable GA, the consequences of which are explained.

email: scott.schwartz@stat.duke.edu

### USING FINITE MULTIVARIATE MIXTURES TO MODEL ADVERSE BIRTH OUTCOMES

Matthew W. Wheeler\*, University of North Carolina-Chapel Hill  
Amy Herring, University of North Carolina-Chapel Hill  
Eric Kalendra, North Carolina State University  
Montse Fuentes, North Carolina State University  
Brian Reich, North Carolina State University

Gestational age and birth weight are continuous pregnancy outcomes that are frequently investigated in the literature. As the observed distribution of both outcomes is a complex multivariate distribution investigators frequently dichotomize both responses into adverse response categories. These dichotomized variables are then analyzed independently. Such analyses obviously ignore possible correlation between the pregnancy outcomes, and create cut points that may lose important features of data. In attempt to better characterize risk of both pre-term delivery and low birth weight we study both of these outcomes jointly using finite mixture models. Here a bivariate normal distribution is used for each bin, and, given a bin, the mean age and

birth weights are modeled through regression using a vector of risk factors. Further, as the probability of group membership may also be dependent of these and other risk factors, we model the probability of falling in a specific bin through logistic regression. Thus risk factors are used to describe the risk of an adverse outcome, and explain the changes in the mean gestational age and birth weight given the assignment to a specific bin in order to better characterize the observed distribution of birth outcomes.

email: mwheeler@bios.unc.edu

### A MIXTURE MODEL FOR THE ANALYSIS OF CORRELATED BINOMIAL DATA

N. Rao Chaganty\*, Old Dominion University  
Yihao Deng, Indiana University-Purdue University Fort Wayne  
Roy Sabo, Virginia Commonwealth University

Analysis of correlated binary data has been the topic of numerous papers during the last two decades. However, not much attention was given to the analysis of correlated binomial data. The analysis is complicated because there are many parameters in the usual multivariate models. Further the ranges of the correlation parameters, being functions of the marginal means, are subject to unmanageable constraints. In this talk we will discuss a mixture model, which circumvents those complications, allows a wide range of dependence and requires only specification of the first two moments for the mixing distribution. We will discuss estimation of the parameters and an illustrative example.

email: rchagant@odu.edu

### LATENT TRANSITION MODELS TO STUDY CHANGE IN DIETARY PATTERNS OVER TIME

Daniela Sotres-Alvarez\*, University of North Carolina-Chapel Hill  
Amy H. Herring, University of North Carolina-Chapel Hill  
Anna Maria Siega-Riz, University of North Carolina-Chapel Hill

Latent class models (LCM) have been shown empirically to be more appropriate to derive dietary patterns (DP) than cluster analysis since they allow different outcome distributions, correlated measurement errors, and adjustment for energy intake and covariates. The latent transition model (LTM) might be useful to study change as characterized by the movement between discrete DP. In practice, LTM have been mostly used in the social sciences and for applications with few nominal outcomes, and have not been used to study movement between DP. In addition to the problem on how to determine the number of classes, there are several challenges particular to DP analysis: large (>80) number of food-items, non-standard mixture distributions (continuous with a mass point at zero for non-consumption), and typical assumptions (conditional independence given the class and timepoint, time-invariant conditional responses, and invariant transition probabilities) may not be realistic. We review the LTM and illustrate a model selection strategy using an example from nutritional epidemiology. We investigate the implications in interpretation of the classic and identifiability assumptions, and provide guidance for potentially problematic situations (small sample size or intermediate item-response probabilities). We estimate LTM

using the free SAS procedure LTA, and a software-package for latent models, Mplus.

email: dsotres@bios.unc.edu

## CONDITIONAL ASSESSMENT OF ZERO-INFLATED MIXTURE MODELS

Yan Yang\*, Arizona State University

Doug G. Simpson, University of Illinois at Urbana-Champaign

The class of zero-inflated mixture models has been widely used to analyze data bounded below by zero with an excess of zero observations. However, little attention has been focused on assessing the adequacy of these models. We propose a conditional decomposition approach that separately evaluates the fit for values at the boundary and the fit for values exceeding it. The model-based conditional mean and quantiles for values greater than the lower bound and marginal mean and quantiles for all values are derived. Confidence intervals with delta method standard errors are then implemented for the probability of the boundary event, conditional mean and conditional quantiles to assess inflated mixture models. A simulation study is conducted to investigate the finite-sample behavior of the intervals. The usefulness of the proposed methods is illustrated for data from an ultrasound safety study and from a measles vaccine study, where conditional evaluations help suggest a reason for the lack of fit of current models and thus lead to improved ones.

email: yy@math.asu.edu

## BAYESIAN ESTIMATION OF MULTILEVEL MIXTURE MODELS

Tihomir Asparouhov\*, Mplus

Bengt Muthén, UCLA

We describe a general multilevel mixture model where latent class variable appear not only on the individual-level but also on the cluster-level. A fully Bayesian approach is used for model estimation using the Gibbs sampler. We contrast this estimation method with the maximum-likelihood estimation method. We illustrate the technique with multilevel analysis of achievement data with classification of both students and schools.

email: asparouhov@hotmail.com

## SMOOTH DENSITY ESTIMATION WITH MOMENT CONSTRAINTS USING MIXTURE DENSITIES

Ani Eloyan\*, North Carolina State University

Sujit K. Ghosh, North Carolina State University

One of the common issues in statistical analysis is the estimation of a density function based on a sample of observations drawn from an unknown density. There are a number of approaches to solving this problem using both nonparametric and parametric methods. One of the nonparametric methods is the adaptation of P-spline smoothing techniques to estimate a density. The motivation is that any

continuous density can be approximated by a mixture of densities with appropriately chosen moments and weights. In many problems (e.g., random effects) we may have specific information about the moments of the density such as restrictions on the mean and variance. A novel method based on EM-algorithm is proposed for estimating the weights of the mixture density under the constraints on the moments of the density. In addition, the proposed method also obtains an estimate of the number of components in the mixture density needed for optimal approximation. The proposed method is compared with the usual Kernel-based density estimation using simulated data and it is shown that the proposed estimate outperforms the Kernel-based method in terms of minimizing the Kullback-Leibler divergence. The proposed method is illustrated by applying it to several real data examples.

email: aeloyan@ncsu.edu

## TUESDAY, MARCH 17, 2009 8:30-10:15 AM

### 60. STATISTICAL ISSUES IN BRAIN IMAGING

#### ON COMBINING AND CONTRASTING BRAINS

Nicole A. Lazar\*, University of Georgia

A challenging problem in the statistical analysis of human brain function via functional magnetic resonance imaging (fMRI) is that of comparing activation across groups of subjects. In the first part of this talk, I will discuss methods for “combining” brains, that is, creating a map that summarizes the overall activity pattern of a group of subjects. This can be analogized to the old problem of combining information from independent studies, and I draw on techniques historically used for that problem, to solve the current one. Once a map has been created for a single group of subjects, we can think about “contrasting”, or comparing, the maps for multiple groups. While group comparisons are often accomplished via such standard techniques as the random effect linear model, I will argue that this approach is potentially over-conservative, impairing the ability to detect differences of interest, which may be differences of extent, of magnitude, or both. Instead, I propose extending various of the methods used in the first part of the talk for making group maps, via a combination of statistical distribution theory and computational procedures (bootstrap and permutation). In the second part of this talk, I will discuss some of the issues that arise in extending group maps in this way, and some possible solutions.

email: nlazar@stat.uga.edu

### ANALYZING fMRI DATA WITH UNKNOWN BRAIN ACTIVATION PROFILES

Martin A. Lindquist\*, Columbia University

Lucy F. Robinson, Columbia University

Tor D. Wager, Columbia University

Most statistical analyses of fMRI data assume that the exact nature, timing and duration of the psychological processes being studied are known. However, in many areas of psychological inquiry (e.g.

# ABSTRACTS

studies on memory, motivation, emotion and drug uptake), it is hard to specify this information a priori. In this talk we discuss a spatio-temporal model that can be used to analyze this type of data. The approach allows for the estimation of voxel-specific distributions of onset times and durations from the fMRI response assuming no functional form (e.g., no assumed neural or hemodynamic response), and allowing for the possibility that some subjects may show no response. The distributions can be used to estimate the probability that a voxel is activated as a function of time, and to cluster voxels based on characteristics of their onset, duration, and anatomical location.

email: martin@stat.columbia.edu

## STATISTICAL ANALYSIS OF BRAIN MORPHOMETRIC MEASURES ON RIEMANNIAN MANIFOLD

Hongtu Zhu\*, University of North Carolina-Chapel Hill  
Joseph G. Ibrahim, University of North Carolina-Chapel Hill  
Yimei Li, University of North Carolina-Chapel Hill  
Weili Lin, University of North Carolina-Chapel Hill  
Yasheng Cheng, University of North Carolina-Chapel Hill

The aim of this talk is to develop an intrinsic regression model for the analysis of brain morphometric measures as responses in a Riemannian manifold and their association with a set of covariates, such as age and gender, in a Euclidean space. The primary motivation and application of the proposed methodology is in medical imaging. Because some brain morphometric measures do not form a vector space, applying classical multivariate regression to modeling those measures may undermine their association with covariates of interest, such as age and gender, in real applications. Our intrinsic regression model, as a semiparametric model, uses a link function to map from the Euclidean space of covariates to Riemannian manifold. We develop an estimation procedure to calculate parameter estimates and establish their limiting distribution. We develop score statistics to test linear hypotheses of unknown parameters and develop a test procedure based on resampling method to simultaneously assess the statistical significance of linear hypotheses across a large region of interest.

email: hzhu@bios.unc.edu

## TILING MANIFOLDS WITH ORTHONORMAL BASIS

Moo K. Chung\*, University of Wisconsin-Madison

One main obstacle in building a sophisticated parametric model along an arbitrary manifold is the lack of an easily available orthonormal basis. Although there are at least two numerical techniques available for constructing an orthonormal basis such as the Laplacian eigenfunction approach and the Gram-Smidt orthogonalization, they are computationally not so trivial and costly. We present a relatively simpler method for constructing an orthonormal basis for an arbitrary manifold by the concept of pullback operation on a preconstructed basis. As an application, we construct an orthonormal basis on amygdala surface of the brain. Then using the basis, we present a new

variance reducing shape representation of amygdala compared to the traditional Fourier representation.

email: mkchung@wisc.edu

## 61. CAUSAL INFERENCE WITH INSTRUMENTAL VARIABLES METHODS

### A NONPARAMETRIC APPROACH TO INSTRUMENTAL VARIABLES ANALYSIS WITH BINARY OUTCOMES

Michael Baiocchi\*, University of Pennsylvania  
Paul Rosenbaum, University of Pennsylvania  
Dylan Small, University of Pennsylvania

An instrumental variable is a variable that is effectively randomly assigned and influences the treatment but affects the outcome only through its influence on the treatment. Instrumental variables are useful for estimating the causal effect of a treatment in an observational study in which not all confounders can be measured. When the outcome is binary, a number of parametric models, such as the multivariate probit, have been developed for using instrumental variables to estimate the causal effect of a treatment. We demonstrate that these methods are sensitive to the parametric assumptions and develop a nonparametric, permutation inference approach. Our approach uses matching methods to account for exogenous covariates. We apply our approach to estimate the effects of attending a regional perinatal center versus a local hospital on mortality of prematurely born children using the mother's excess distance from the regional perinatal center compared to the local hospital as an instrumental variable.

email: mbaiocch@wharton.upenn.edu

### SEMIPARAMETRIC ESTIMATION AND INFERENCE FOR DISTRIBUTIONAL AND GENERAL TREATMENT CAUSAL EFFECTS

Jing Cheng\*, University of Florida College of Medicine  
Jing Qin, National Institute of Allergy and Infectious Diseases,  
National Institutes of Health  
Biao Zhang, University of Toledo

This talk considers evaluating the treatment effects on the outcome distribution and its general functions in randomized trials with noncompliance. For distributional treatment effects, fully nonparametric and fully parametric approaches have been proposed, where the fully nonparametric approach could be inefficient and the fully parametric approach could be non-robust to the violation of distribution assumptions. In this work, we develop a semiparametric instrumental variable approach by using empirical likelihood method with a density ratio model, under which the underlying densities of the latent compliance classes are linked together by exponential tilts, however, the baseline density is left unspecified. Our method can be applied to general outcomes and general functions of outcome distributions, and allows us to predict a subject's latent compliance

class based on an observed outcome value in observed assignment and treatment received groups. Asymptotic results for the estimators and likelihood ratio statistic are derived, and finite sample performance is examined by a simulation study. The method is 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

## EXTENDED INSTRUMENTAL VARIABLES ESTIMATION FOR OVERALL EFFECTS

Marshall M. Joffe\*, University of Pennsylvania  
Dylan Small, University of Pennsylvania  
Thomas Ten Have, University of Pennsylvania  
Steven Brunelli, University of Pennsylvania  
Harold I. Feldman, University of Pennsylvania

We consider a method for extending instrumental variables methods in order to estimate the overall effect of a treatment or exposure. The approach is designed for settings in which the instrument influences both the treatment of interest and a secondary treatment also influenced by the primary treatment. We demonstrate that, while instrumental variables methods may be used to estimate the joint effects of the primary and secondary treatments, they cannot by themselves be used to estimate the overall effect of the primary treatment. However, instrumental variables methods may be used in conjunction with approaches for estimating the effect of the primary on the secondary treatment to estimate the overall effect of the primary treatment. We consider extending the proposed methods to deal with confounding of the effect of the instrument, mediation of the effect of the instrument by other variables, failure-time outcomes, and time-varying secondary treatments. We motivate our discussion by considering estimation of the overall effect of the type of vascular access among hemodialysis patients. We also consider application of these methods and variants to estimation of the joint effects of multiple phenotypes.

email: mjoffe@mail.med.upenn.edu

## 62. FROM THEORY TO PRACTICE: EXAMPLES AND DISCUSSION OF SOFTWARE DEVELOPMENT FOR WIDE DISSEMINATION OF STATISTICAL METHODOLOGY

### SOFTWARE DEVELOPMENT FOR GEE AND ORTH

John S. Preisser\*, University of North Carolina at Chapel Hill  
Bahjat F. Qaqish, University of North Carolina at Chapel Hill

Although they received some early attention, the generalized estimating equations (GEE) observation- and cluster-deletion diagnostics of Preisser and Qaqish (1996) had yet seen broad use in practice up to their incorporation in the GENMOD procedure of SAS version 9.2 released in 2008. We consider our experience in the early lack of promotion of this methodology and in the late development of accompanying software. We apply lessons learned to the recent dissemination of computational tools (R and SAS) for a new statistical

methodology, orthogonalized residuals (ORTH), that includes the alternating logistic regressions procedure. We discuss incentives, disincentives, and rewards for pursuing software development in an academic setting.

email: jpreisse@bios.unc.edu

## SAS MACRO QIF: TRANSITION OF METHODOLOGY RESEARCH TO APPLICATION

Peter X. Song\*, University of Michigan School of Public Health

Quadratic inference function (QIF) proposed by Qu et al. (2000) is getting increasingly popular because this method is shown to have some desirable properties that the widely used method of generalized estimating equation (GEE) lacks of. To pass this new powerful method to the hands of practitioners, we developed a user-friendly SAS macro. In this talk, I will focus on the development of SAS MACRO QIF and its usage tips. I will also give a live demonstration of running this macro for the analysis of real world longitudinal data. Some comparisons of the QIF to the GEE will be discussed based on the results produced by the SAS software package.

email: pxsong@umich.edu

## DERIVATION AND SOFTWARE IMPLEMENTATION OF THE CANONICAL NEGATIVE BINOMIAL MODEL

Joseph M. Hilbe\*, Arizona State University

The negative binomial regression model can be parameterized in several ways. I discuss the software algorithms used to model the negative binomial for each of these parameterizations, and how each is to be interpreted for a given data situation. The canonical negative binomial, NB-C, developed by the presenter, is described in detail and related to other parameterizations of the negative binomial; eg NB-1, NB-2, NB-H, NB-P, and ZIP/ZINB. The presenter is author of /Negative Binomial Regression/ (2007, Cambridge University Press), has been Software Reviews Editor of The American Statistician since 1997, and was the first to implement the negative binomial into commercial generalized linear models software.

email: jhilbe@aol.com

## 63. BAYESIAN METHODS IN PROTEIN BIOINFORMATICS

### MODELING THE JOINT DISTRIBUTION OF PAIRS OF DIHEDRAL ANGLES FOR PROTEIN STRUCTURE PREDICTION

David B. Dahl\*, Texas A&M University  
Ryan Day, University of the Pacific  
Jerry W. Tsai, University of the Pacific

There is considerable interest in the prediction of a protein's structure, particularly its backbone from its underlying amino acid sequence. A description of a protein's backbone can be described by the torsion

# ABSTRACTS

angle pairs of each of the protein's residues. Most methods for torsion angles focus on the bivariate angles for a given residue. In this talk, I will present a method to jointly model the angles at several positions and demonstrate its use for protein structure prediction.

email: dahl@stat.tamu.edu

## BAYESIAN NONPARAMETRIC ANALYSIS OF SITE-SPECIFIC SELECTION EFFECTS IN SERIALLY DNA SEQUENCES

Daniel Merl\*, Duke University  
Raquel Prado, University of California, Santa Cruz  
Athanasios Kottas, University of California, Santa Cruz

We present a novel statistical model for investigating the effects of natural selection in serially sampled protein coding sequence alignments. The major methodological development here is the use of a combined Dirichlet process and variable selection prior for the distribution of site-specific selection effects in order to accomplish simultaneous variable selection and clustering. The potential of these modelling assumptions for use in the problem of inferring site-specific selection effects is clear: variation at many sites will be adequately explained by the baseline effects associated with neutral selection, while those sites that are found to have non-zero selection effects will be subsequently clustered according to association with some number of unique selection profiles. We demonstrate the method using sequence data collected to measure the effects of nevirapine treatment on the HIV1 polymerase gene.

email: dan@stat.duke.edu

## MODEL-BASED VALIDATION OF PROTEIN-PROTEIN INTERACTIONS IN LARGE-SCALE AFFINITY PURIFICATION-MASS SPECTROMETRY EXPERIMENTS

Hyungwon Choi\*, University of Michigan  
Ashton Breitkreutz, Samuel Lunenfeld Research Institute, Mount Sinai Hospital Toronto  
Brett Larsen, Samuel Lunenfeld Research Institute, Mount Sinai Hospital Toronto  
Anne-Claude Gingras, Samuel Lunenfeld Research Institute, Mount Sinai Hospital Toronto  
Mike Tyers, Wellcome Trust Centre for Cell Biology, University of Edinburgh, UK  
Zhaohui S. Qin, University of Michigan  
Alexey I. Nesvizhskii, University of Michigan

Large-scale affinity purification (AP) and mass spectrometry (MS) have facilitated the identification of protein-protein interaction (PPI) networks on the global scale. However, elucidating the complex and transient nature of the signaling network is challenging if the identification step results in many false positives. In this work, we present a computational approach for model-based assessment of significance of observed interactions using semi-quantitative measures of absolute protein abundance based on spectral counts. We describe an advanced probability model-based method, called Significance Analysis of Interactome (SAINT), for selecting high

confidence interactions from experimental noise in the AP-MS data. The proposed model was applied to a large-scale AP-MS dataset of PPIs in budding yeast whole cell lysate, with a specific focus on the protein kinases. The reconstructed kinase network from multiple tag experiments is composed of close to 900 proteins with more than 1500 high-probability interactions. As expected in a signaling network, the structure of the identified network clearly reveals a notably high degree of connectivity among kinases.

email: hwchoi912@gmail.com

## BAYESIAN ANALYSIS OF MOLECULAR FORCEFIELDS

Scott C. Schmidler\*, Duke University

Computer simulation, especially molecular dynamics simulation, is an important and widely used tool in the study of biomolecular systems. The behavior of molecular systems is typically described by a potential (forcefield) involving many physicochemical parameters. However, for macromolecular systems such as proteins and nucleic acids and their complexes, large scale validation against experimental data have been lacking, and formal statistical procedures for parameter estimation have not been developed. We focus on two aspects of the problem: the difficulty of evaluating simulations using experimental measurements due to ensemble averaging, and the challenges of likelihood-based parameter estimation due to the well-known problem of intractable normalizing constants in Gibbs random fields. We describe the application of standard ideas from statistical learning and convergence of stochastic processes to address the former, and describe a Monte Carlo scheme for parameter estimation in general Gibbs fields. We demonstrate the results on simulations of helical peptides.

email: schmidler@stat.duke.edu

## 64. ADAPTIVE DESIGNS FOR EARLY-PHASE CLINICAL TRIALS

### SEQUENTIAL IMPLEMENTATION OF STEPWISE PROCEDURES FOR IDENTIFYING THE MAXIMUM TOLERATED DOSE

Ken Cheung\*, Columbia University

In this talk, I address the dose-finding problem in phase I clinical trials by a multiple test approach: step-down tests are used in an escalation stage, and step-up tests in a de-escalation stage in order to allow sequential dose assignments for ethical purposes. By formulating the estimation problem as a testing problem, the proposed procedures formally control the error probability of selecting an unsafe dose. In addition, we can control the probability of correctly selecting the maximum tolerated dose (MTD) under a parameter subspace where no toxicity probability lies in an interval bracketed by the target toxicity rate and an unacceptably high toxicity rate, the so-called "indifference zone". This frequentist property, which is currently lacking in the conduct of dose-finding trials in humans, is appealing from a regulatory viewpoint. From a practical viewpoint, stepwise tests are simple and easy to understand, and the sequential implementation operates in a similar manner to the traditional algorithm familiarized

by the clinicians. Extensive simulations illustrate that our methods yield good and competitive operating characteristics under a wide range of scenarios with realistic sample size, and performs well even in situations other existing methods may fail, namely, when the dose-toxicity curve is flat up to the targeted MTD.

email: yc632@columbia.edu

## **ADAPTIVE RANDOMIZATION FOR MULTI-ARM COMPARATIVE CLINICAL TRIALS BASED ON JOINT EFFICACY/TOXICITY OUTCOME**

B. Nebiyou Bekele\*, M. D. Anderson Cancer Center  
Yuan Ji, M.D. Anderson Cancer Center

An outcome-adaptive randomization scheme for comparative clinical trials in which the primary endpoint is a joint efficacy/toxicity outcome is presented. Under the proposed scheme, the randomization probabilities are unbalanced adaptively in favor of treatments with superior joint outcomes characterized by higher efficacy and lower toxicity. This type of scheme is advantageous from the patients' perspective since on average, more patients are randomized to superior treatments. We extend the approximate Bayesian time-to-event model in Cheung and Thall (2002) to model the joint efficacy/toxicity outcomes and perform posterior computation based on a latent variable approach. Consequently, this allows us to incorporate essential information about patients with incomplete follow-up. Based on the computed posterior probabilities, we propose an adaptive randomization scheme that favors the treatments with larger joint probabilities of efficacy and no toxicity. We illustrate our methodology with a leukemia trial that compares three treatments in terms of their 52-week molecular remission rates and 52-week toxicity rates.

email: bbekele@mdanderson.org

## **FINDING THE DOSE WITH THE BEST EFFICACY/TOLERABILITY PROFILE**

Anastasia Ivanova\*, University of North Carolina at Chapel Hill

The goal of a phase II dose-finding study is to find the best dose to investigate in subsequent trials. Such a dose should have good efficacy, which is often established by comparing the efficacy of the drug with placebo and/or active control. Doses with high efficacy might have higher rates of adverse events, hence both outcomes should be taken into consideration for dose selection. We present an efficient adaptive dose-finding strategy to estimate the dose which has the best efficacy/tolerability profile.

email: aivanova@bios.unc.edu

## **65. MEMORIAL SESSION: THE LIFE OF DAVID BEATTY DUNCAN: BIOSTATISTICIAN, MENTOR, AND GENTLEMAN**

## **MEMORIAL SESSION: THE LIFE OF DAVID BEATTY DUNCAN: BIOSTATISTICIAN, MENTOR, AND GENTLEMAN**

Jay Herson, Independent Consultant and Johns Hopkins University  
Gene A. Pennello\*, Food and Drug Administration

Dennis O. Dixon, National Institute of Allergy and Infectious Diseases

Kenneth M. Rice, University of Washington

David Duncan (1916-2006), distinguished Professor of Biostatistics at Johns Hopkins Bloomberg School of Public Health from 1960-1984, had a profound influence in the US and in Australia on the field of biostatistics. His contributions include the discovery of logistic regression, the development of recursive estimation methods (e.g., dynamic estimation of the regression equation, i.e., the Kalman filter), and the development of multiple comparison procedures, his career long interest. His 1955 Biometrics paper, "Multiple Range and Multiple F tests", is one of most highly cited in all of Statistics. It also laid the groundwork for his pioneering, Bayesian decision theoretic approach to multiple comparisons. The approach was honed in doctoral thesis work by many of his Ph.D. students to produce multiple comparison procedures for a variety of problems. David is remembered for his genuine concern for the training and career development of junior colleagues. His energy, enthusiasm, and engaging personality are missed. In this session, a retrospective will be given on David's contributions to the field of statistics and how his contributions relate to contemporary statistical research. Speakers and audience members are also invited to reflect on personal remembrances and anecdotes about David.

email: gene.pennello@fda.hhs.gov

## **66. SPATIAL ANALYSES OF ALCOHOL, ILLEGAL DRUGS, VIOLENCE AND RACE**

### **GEOSPATIAL MODELS OF ALCOHOL, DRUGS AND VIOLENCE**

Dennis M. Gorman\*, School of Rural Public Health, Texas A&M Health Science Center

Ecologic studies have shown a relationship between alcohol outlet densities and violence and between the location of illicit drug markets and violence. Most of the former studies have been conducted by researchers in the fields of alcohol studies and public health and have employed some geographic space (such as a census block group or tract) as the unit of analysis. The major theoretical models used by researchers in this area include social disorganization theory, social capital theory, and collective efficacy theory. The second group of studies has mainly been conducted by criminologists and typically uses point data pertaining to specific places (or clusters of places) as the unit of analysis. The major theoretical models used by researchers in this area include routine activities theory and hot-spot theory. Both sets of studies present theoretical, methodological and statistical challenges. These are reviewed in this introductory presentation so as to provide a context for the analysis of spatial models of alcohol and drug availability and violence. Particular emphasis is placed on examining the extent to which the theories that are used in this research are truly spatial and ecologic.

email: gorman@srph.tamhsc.edu

# ABSTRACTS

## MODELING THE METHAMPHETAMINE EPIDEMIC IN CALIFORNIA

Paul J. Gruenewald\*, Prevention Research Center, Pacific Institute for Research and Evaluation

William R. Ponicki, Prevention Research Center, Pacific Institute for Research and Evaluation

Rates of methamphetamine abuse and dependence have reached epidemic levels in California over the past three decades, with rates increasing by about 200% over the past 8 years and currently increasing by 20% per year. Three central questions are of concern in studies of the epidemic: Are there natural limits to epidemic growth? Are there correlated patterns of epidemic growth over space and time? Are specific population frailties related to rates of growth? This paper presents the results of Bayesian space-time disease models applied to the assessment of geographic patterns of methamphetamine abuse and dependence across California cities over 27 years. Results of these modeling exercises illuminate the difficulties of assessing limits to growth of complex evolving epidemics, identify geographically correlated patterns of epidemic growth and decline, and suggest some population characteristics that accelerate rates of spatially correlated growth across human populations. Some concluding observations are offered regarding modeling needs for predicting the development of future drug epidemics.

email: paul@prev.org

## SPATIAL RELATIONSHIPS BETWEEN THE SUBSTANCE USE ENVIRONMENT AND CHILD MALTREATMENT

Bridget J. Freisthler\*, UCLA

Substance use has long been studied as both a contributor to and a consequence of child maltreatment. For example, studies of clinical or convenience samples of families already involved with the child welfare system found a positive relationship between child maltreatment and substance use. Substance-abusing parents are also more likely to be reported multiple times to the child welfare system for child maltreatment. Yet the social mechanisms that produce this relationship remain largely unarticulated and understudied. For example, already weakened (or "frail") neighborhood structure may lack the appropriate social capital to absorb the negative effects related to high densities of alcohol outlets in their community while the density of substance abuse services may mitigate that risk. Without greater attention to understanding how community systems affect patterns of child maltreatment, effective prevention programs cannot be developed. Through a discussion of advanced spatial modeling techniques, this presentation will delineate ecological pathways by which the substance use environment (a) increases the likelihood of child maltreatment; and (b) how the density of social services may moderate this relationship. It is argued that specific family dynamics are affected by parent's substance use and use of community systems for acquiring and using alcohol and drugs.

email: freisthler@spa.ucla.edu

## SOLVING THE MISALIGNMENT PROBLEM IN THE ANALYSIS OF DRUG ISSUES

Li Zhu\*, Texas A&M Health Science Center

Lance Waller, Emory University

Paul Gruenewald, Prevention Research Center

It is well known that the choice of geographic unit can affect the interpretation of maps and the results of spatial analysis, a phenomenon known as the modifiable areal unit problem. A model-based hierarchical structure is developed to analyze data whose geographic units are misaligned in both space and time. We extend the multivariate Gaussian models to generalized linear mixed models that incorporate several forms of discrete data. The approach is illustrated with a dataset relating illicit drug use, a range of sociodemographic variables, geography, and time in Tracy, CA. Here geography is indicated with ZIP codes which change over a period of eleven years. Computing is implemented via a carefully tailored Metropolis-Hastings algorithm, with map summaries created using a geographic information system.

email: lizhu@srph.tamhsc.edu

## VARYING PARAMETER MODELS IN SPACE AND TIME

William R. Ponicki\*, Prevention Research Center, Pacific Institute for Research and Evaluation

Lance A. Waller, Rollins School of Public Health, Emory University

Unlike contagious diseases, rates of problems related to alcohol and drug abuse are strongly conditioned by well understood psychological characteristics (e.g., impulsivity), behavioral habits (e.g., daily routine activities), economic circumstances (e.g., access to drugs), and societal structures (e.g., friendship networks). Human populations exhibit a broad range of heterogeneous risks for such problems, and these risks are often conditional upon one another and correlated with both social strata and the locations of neighborhoods in which people live. It is thus expected that correlations between individual risk factors and abuse / dependence outcomes will also be heterogeneously distributed across geographic areas. This presentation examines whether the relationships between measures of abuse and dependence and a number of exogenous variables vary in a spatially-correlated manner. Bayesian space-time models are used to analyze arrests and hospital discharges related to methamphetamine in a large sample of California cities over the years 1980 through 2006. Varying parameters are implemented using conditional-autoregressive random effects in a manner analogous to geographically weighted regressions. The results demonstrate that effects related to exogenous measures vary geographically and over time across areas of California.

email: bponicki@prev.org



## **HIERARCHICAL ADDITIVE MODELING OF NONLINEAR ASSOCIATION WITH SPATIAL CORRELATIONS - AN APPLICATION TO RELATE ALCOHOL OUTLET DENSITY AND NEIGHBORHOOD ASSAULT RATES**

Qingzhao Yu\*, Louisiana State University  
Bin Li, Louisiana State University  
Richard Scribner, Louisiana State University

Previous studies have suggested a link between alcohol outlets and assaults. In this talk, we explore the effects of alcohol availability on assaults at the census tract level over time. Several features of the data raise statistical challenges: (1) the association between covariates and the assault rates may be complex and therefore cannot be described using a linear model without covariates transformation; (2) the covariates may be highly correlated with each other; (3) there are a number of observations that have missing inputs; and (4) there is spatial association in assault rates at the census tract level. We propose a hierarchical additive model, where the nonlinear correlations and the complex interaction effects are modeled using the multiple additive regression trees and the residual spatial association in the assault rates that cannot be explained in the model are smoothed using a Conditional Autoregressive (CAR) method. We develop a two-stage algorithm that connects the non-parametric trees with CAR to look for important covariates associated with the assault rates, while taking into account the spatial association of assault rates in adjacent census tracts. The proposed method is applied to the Los Angeles assault data. The method is compared with a hierarchical linear model.

email: qyu@lsuhsc.edu

## **SPATIAL ASSOCIATION BETWEEN RACIAL AND SOCIAL FACTORS IN DETERMINING LOW BIRTH WEIGHT AND PRETERM BIRTH, NEW YORK, 1995-2005**

Eric Kalendra\*, North Carolina State University  
Montserrat Fuentes, North Carolina State University  
Brian Reich, North Carolina State University  
Amy Herring, University of North Carolina  
Matthew Wheeler, University of North Carolina

Regardless of whether low birth weight (2500 grammes or less at birth) is caused by intra-uterine growth retardation or pre-term birth (before 37 weeks of gestation), the condition remains one of the strongest predictors of neonatal and infant mortality. Using data from singleton live births in New York City from 1995-2003, in which over 92% of vital records were matched to maternal and infant hospital discharge data, this study examines the potential impact that racial and social factors have in determining low birth weight and preterm birth. The analysis controls for a number of potentially confounding factors, including mother's age and census-based neighborhood socioeconomic and demographic characteristics of the environment in which these women live. The analysis is carried out using a spatially varying coefficients model for race. Spatial smoothing is done with a conditionally autoregressive (CAR) model. By using different weighting functions in the CAR model we are able to control the amount of smoothing. In our model we are able to explain local differences that are averaged over when we use a non-spatially varying

coefficients model. Our study shows that the census level information is not enough to explain the within racial variation, that the spatially varying coefficients model is able to characterize.

email: eric.kalendra@gmail.com

## **67. INCORPORATING EXTERNAL KNOWLEDGE IN THE ANALYSIS OF GENOMIC DATA**

### **INCORPORATION OF PRIOR KNOWLEDGE FROM LINKAGE STUDIES INTO GENETIC ASSOCIATION STUDIES**

Brooke L. Fridley\*, Mayo Clinic  
Daniel Serie, Mayo Clinic  
Kristin L. White, Mayo Clinic  
Gregory Jenkins, Mayo Clinic  
William Bamlet, Mayo Clinic  
Ellen L. Goode, Mayo Clinic

In the last decade, numerous genome-wide linkage and association studies of complex diseases have been completed. These datasets and/or results are often available to investigators through collaborations or dbGaP. Critical questions remain regarding how best to use this valuable information to inform current, on-going or future genetic association studies to improve study design and statistical analysis. One promising approach to incorporating prior knowledge from linkage scans or other information is to up- or down-weight p-values resulting from genetic association study in either a frequentist or Bayesian approach. An alternative approach would use a fully Bayesian analysis for the genetic association study, where the specification of the prior distributions is based on the prior knowledge. We propose a Bayesian mixture model for analysis of genetic association studies to incorporate prior knowledge based on biology or linkage results. We illustrate the proposed method using a genetic association study of colon polyps and a genome-wide linkage study of colon cancer. In addition to application to genetic studies of colon cancer, we present results from a variety of simulation studies.

email: fridley.brooke@mayo.edu

## **TESTING UNTYPED SNPs IN CASE-CONTROL ASSOCIATION STUDIES WITH RELATED INDIVIDUALS**

Zuoheng Wang\*, University of Chicago  
Mary Sara McPeek, University of Chicago

Genome-wide association study has been a popular tool for identifying genetic factors influencing human complex disease. However only a subset of the total variations across the genome are genotyped in the recent high-throughput genotyping platforms. We introduce a new approach to testing for association between untyped variants and a binary trait, using linkage disequilibrium (LD) information between typed markers and untyped markers. This information can often be obtained from an appropriate reference panel such as HapMap. We construct a 1-d.f. incomplete-data quasi-likelihood score (IQLS) test based on an IQLS function. Because the external LD information is only used to form the alternative mean model, our method maintains

# ABSTRACTS

the nominal single-SNP type I error rate even when the external LD information is misspecified. As a result, our method is robust to an inappropriate choice of the reference samples for testing untyped SNPs. Furthermore, the IQLS function was previously proposed to address both dependent and partially-observed data, our 1-d.f. IQLS test appropriately accommodates the problem of haplotype-phase ambiguity in genotype data, and can also be applied to case-control association studies in which some sampled individuals are related, with known relationships. We apply the method to test for association with type 2 diabetes in the Framingham Heart Study.

email: zwang@galton.uchicago.edu

## MULTIPLE SNP-BASED APPROACH FOR GENOME-WIDE CASE-CONTROL ASSOCIATION STUDY

Min Zhang\*, Purdue University  
Yanzhu Lin, Purdue University  
Libo Wang, Purdue University  
Vitara Pungpapong, Purdue University  
James C. Fleet, Purdue University  
Dabao Zhang, Purdue University

Genome-wide association study is challenged by the collection of a large number of SNPs from a relatively small number of individuals, and therefore, one SNP is investigated at a time. However, such univariate association study ignores the multigenic nature of common complex diseases as well as the linkage disequilibrium between SNPs. Here we propose a method to analyze all SNPs in the same linkage group simultaneously by implementing the linear discriminant analysis through a penalized orthogonal-components regression, which is a newly developed variable selection approach for high dimensional data. The proposed method is applied to real genome-wide association study data.

email: minzhang@stat.purdue.edu

## MINING PATHWAY-BASED SNP SETS IN GWAS STUDY WITH SPARSE LOGISTIC REGRESSION

Lin S. Chen\*, Fred Hutchinson Cancer Research Center  
Ulrike Peters, Fred Hutchinson Cancer Research Center  
Li Hsu, Fred Hutchinson Cancer Research Center

Genome-wide association study (GWAS) conducts marginal tests for association between disease status and markers across the entire genome. The huge number of SNP markers and often the high dependence among them make it challenging to detect genetic variations contributing to the complex disease of interest, and to interpret the identified significant SNPs with sound and convincing biological evidence. To more efficiently mining and interpreting SNP information across the entire genome, we propose to map SNPs to the gene level, and explore gene-gene and/or gene-environment interaction within a priori defined disease-related pathways or gene sets. Specifically, we perform principal component analysis on SNPs mapped to each gene to construct eigen-SNPs, each of which is a pseudo SNP that capture independent genetic structure within the gene. Applying lasso logistic regression to each gene, we are able to flexibly select disease associated eigen-SNPs based on gene structure.

We further apply group lasso logistic regression to the selected eigen-SNPs to characterize gene-gene and gene-environment interaction within a pathway. As an example, we present an application to a colon cancer GWAS.

email: lche2@fhcrc.org

## BAYESIAN MODELING OF PHARMACOGENETICS DATA

Donatello Telesca\*, University of Texas M.D. Anderson Cancer Center  
Gary L. Rosner, University of Texas M.D. Anderson Cancer Center  
Peter Muller, University of Texas M.D. Anderson Cancer Center

We describe a general framework for the exploration of the relationships between pharmacokinetic pathways and polymorphisms in genes associated with the metabolism of a compound of interest. We integrate a population pharmacokinetics model with a simple sampling model of genetic mutation via a latent conditional dependence prior. Significant interactions are selected allowing the pharmacokinetic parameters to depend on gene sets of variable dimension. We discuss posterior inference and prediction based on RJ-MCMC simulation.

email: donatello.telesca@gmail.com

## NETWORK BASED GENE SET ANALYSIS UNDER TEMPORAL CORRELATION

Ali Shojaie\*, University of Michigan  
George Michailidis, University of Michigan

Cellular functions of living organisms are carried out through complex systems of interacting components. Including such interactions in the analysis and considering subsystems instead of individual components can unveil new facts about complex mechanisms of life. Networks are often used to demonstrate the interactions among components of biological systems and can be efficiently incorporated in the model to improve efficiency in estimation and inference. In this paper, we propose a model for incorporating external information about the underlying network in differential analysis of gene sets. The model provides a flexible framework for analysis of complex experimental designs and can efficiently incorporate temporal correlations among observations. The model is applied to real data on yeast environmental stress response (ESR) as well as simulated data sets.

email: shojaie@umich.edu

## NETWORK-BASED GENOMIC DISCOVERY: APPLICATION AND COMPARISON OF MARKOV RANDOM FIELD MODELS

Peng Wei\*, School of Public Health, University of Minnesota  
Wei Pan, School of Public Health, University of Minnesota

As biological knowledge accumulates rapidly, gene networks encoding genome-wide gene-gene interactions have been constructed. As extensions of the standard mixture model that tests all the genes

iid a priori, Wei and Li (2007) and Wei and Pan (2008) proposed two methods to incorporate gene networks into statistical analysis of genomic data via Discrete- and Gaussian-Markov random field (DMRF and GMRF) based mixture models, respectively. However, it may not be clear how the two methods compare with each other in practical applications. This paper is aimed at this comparison. We also propose two novel constraints on prior specifications for the GMRF model and a fully Bayesian approach to the DMRF model. In addition, we assessed the accuracy of direct posterior probability approach to estimating the False Discovery Rate (FDR) in the context of MRF models. Applications to a ChIP-chip data set and simulated data showed that the modified GMRF models had superior performance as compared with other models, while both MRF-based mixture models, with reasonable robustness to misspecified gene networks, outperformed the standard mixture model that assumes independent genes.

email: weixx035@umn.edu

## 68. ROC ANALYSIS

### THE CASE FOR FROC ANALYSIS

Dev P. Chakraborty\*, University of Pittsburgh

This is a rebuttal to a recent paper (Acad. Radiol. 15:1312-1315, 2008) that raises issues regarding the applicability of FROC methodology to imaging systems evaluations. Unlike many tests, diagnostic imaging provides information about the location(s) of disease, in addition to its presence or absence. Unlike FROC, the ROC paradigm only considers the disease present or absent information and disregards location. I am responsible for JAFROC, a method for analyzing FROC data. JAFROC usage has been gaining acceptance but resistance to it is also increasing. The authors have done a service by expressing their concerns publicly and I am grateful to ENAR for giving me the opportunity to respond in open forum. While the authors note several issues with FROC, with some of which I concur, I strongly disagree with their position that ROC is clinically more relevant than FROC. Much of my response will focus on the clinical relevance issue. I will describe a specific scenario where ROC is less clinically relevant than FROC. Specific rebuttals to issues regarding applicability to screening, data clustering, acceptance target, search model, figure of merit, validation, etc, will be presented. It is my hope that the debate will influence those vested in the ROC method to embrace and contribute to FROC research, rather than feel threatened by it.

email: dpc10@pitt.edu

## BAYESIAN NONPARAMETRIC COMBINATION OF MULTIPLE DIAGNOSTIC MEASUREMENTS

Lorenzo Trippa\*, MD Anderson Cancer Center

Receiver operating characteristic (ROC) curves are widely applied for evaluating the discriminatory ability of diagnostic tests. In many cases a multitude of tests for disease diagnosis are available. In presence of multiple diagnostic measurements it is often of interest to construct a composite score to improve the classification accuracy. We propose

a Bayesian nonparametric model in order to synthesize multiple diagnostic measurements. The model allows to construct composite scores and to evaluate their discriminatory ability. We will illustrate the use of the model through simulated data and an oncology study.

email: ltrippa@mdanderson.org

## OPTIMAL CUTPOINT ESTIMATION WITH CENSORED DATA

Mithat Gonen, Memorial Sloan-Kettering Cancer Center  
Cami Sima\*, Memorial Sloan-Kettering Cancer Center

We consider the problem of selecting an optimal cutpoint for a continuous marker when the outcome of interest is subject to right censoring. Maximal chi square methods and receiver operating characteristic (ROC) curves-based methods are commonly used when the outcome is binary. In this article we show that selecting the cutpoint that maximizes the concordance, a metric similar to the area under an ROC curve, is equivalent to maximizing the Youden index, a popular criterion when the ROC curve is used to choose a threshold. We use this as a basis for proposing maximal concordance as a metric to use with censored endpoints. Through simulations we evaluate the performance of two concordance estimates and three chi-square statistics under various assumptions. Maximizing the partial likelihood ratio test statistic has the best performance in our simulations.

email: simac@mskcc.org

## SEMIPARAMETRIC ROC MODELS WITH MULTIPLE BIOMARKERS

Eunhee Kim\*, University of North Carolina at Chapel Hill  
Donglin Zeng, University of North Carolina at Chapel Hill

In medical diagnostic research, biomarkers are used as the basis for detecting or predicting disease. There has been an increased interest in using the receiver operating characteristic (ROC) curve to assess the accuracy of biomarkers. Even though numerous methods have been developed for a single biomarker, few statistical methods exist to accommodate multiple biomarkers simultaneously. In this paper, we propose a multivariate binormal ROC model to assess multiple biomarkers. Our model assumes that biomarkers follow multivariate normal distribution after unknown and marker-specific transformations. Random effects are introduced to account for within-subject correlation among biomarkers. Nonparametric maximum likelihood estimation is used for inference and parameter estimators are shown to be asymptotically normal and efficient. Both simulation study and real data application are used to illustrate the proposed method.

email: ekim@bios.unc.edu

# ABSTRACTS

## COMPARISON OF TWO BINARY DIAGNOSTIC TESTS: CIRCUMVENTING AN ROC STUDY

Andriy I. Bandos\*, University of Pittsburgh  
Howard E. Rockette, University of Pittsburgh  
David Gur, University of Pittsburgh

Accuracy of a binary diagnostic test is conventionally characterized by Sensitivity and Specificity. A binary test that is better in both operating characteristics is naturally superior. However, when one of the two binary systems is better only in Sensitivity the comparison is not as straightforward. One approach to compare such binary tests is based on expected utilities. This method often requires knowledge of a difficult-to-deduce utility function. Another approach is a conventional ROC analysis. This technique is associated with an additional burden of conducting an ROC study which in some cases may cast doubts on validity and usefulness of the resulting ROC curves. Exploiting an often-satisfied convexity property of ROC curves we describe the conditions under which two diagnostic tests can be compared circumventing utility functions or an ROC study. Under the first set of conditions one binary system augmented by a suitably biased random guess is superior regardless of the utility structure. These conditions also imply that the two latent ROC curves are different. The satisfaction of the second set of conditions further strengthens the conclusions by comparing the area under the latent ROC curves. We describe statistical tests for each set of conditions and present simulation results.

email: anb61@pitt.edu

## TIME-DEPENDENT PREDICTIVE ACCURACY IN THE PRESENCE OF COMPETING RISKS

Paramita Saha\*, University of Washington  
Patrick J. Heagerty, University of Washington

Competing risks arise naturally in time-to-event studies. In this article we propose time-dependent accuracy measures for a marker when we have censored survival times and competing risks. Time-dependent versions of sensitivity or True Positive (TP) fraction naturally correspond to consideration of either cumulative (or prevalent) cases that accrue over a fixed time period, or alternatively to incident cases that are observed among event-free subjects at any select time. Time-dependent (dynamic) specificity ( $1 - \text{False Positive (FP)}$ ) can be based on the marker distribution among event-free subjects. We extend these definitions to incorporate cause of failure for competing risks outcomes. The proposed estimation for cause-specific cumulative TP/dynamic FP is based on the nearest neighbor estimation of bivariate distribution function of the marker and the event time. On the other hand, incident TP/dynamic FP can be estimated using a possibly non-proportional hazards Cox model for the cause-specific hazards and riskset reweighting of the marker distribution. The proposed methods extend the time-dependent predictive accuracy measures of Heagerty, Lumley, and Pepe (2000) and Heagerty and Zheng (2005).

email: psaha@u.washington.edu

## 69. MODEL SELECTION/ASSESSMENT

### THE GRADIENT FUNCTION FOR CHECKING GOODNESS-OF-FIT OF THE RANDOM-EFFECTS DISTRIBUTION IN MIXED MODELS

Geert Verbeke\*, Katholieke Universiteit Leuven and Universiteit Hasselt, Belgium  
Geert Molenberghs, Katholieke Universiteit Leuven and Universiteit Hasselt, Belgium

Inference in mixed models is often based on the marginal distribution obtained from integrating out random effects over a pre-specified, often parametric, distribution. In this paper, we present the so-called gradient function as a simple graphical diagnostic tool to assess whether the assumed random-effects distribution produces an adequate fit to the data, in terms of marginal likelihood. The method does not require any additional calculations in addition to the computations needed to fit the model, and can be applied to every type of mixed model (linear, generalized linear, non-linear), with univariate as well as multivariate random effects. The diagnostic value of the gradient function is extensively illustrated using some simulated examples, as well as in the analysis of a real longitudinal study with binary outcome values.

email: geert.verbeke@med.kuleuven.be

## THE MODEL SELECTION OF ZERO-INFLATED MIXTURE POISSON REGRESSION

Huaiye Zhang\*, Virginia Tech  
Inyoung Kim, Virginia Tech

Poisson regression provides a standard framework for the analysis of the counting data. However in many application, count data has many zeros and also has the mixture distributions. The zero-inflated mixture Poisson regression can be handled for the data. However, it is not obvious to choose a zero-inflated Poisson model without any statistical evidence and also difficult to select the number of mixing components in mixture distributions. Hence, in this paper, we propose a score test for zero-inflated mixture Poisson regression and give a procedure of component selections based on several criterions. And then the mixture model can be estimated using Expectation-Maximization algorithm and be made inference using bootstrapping approach. We demonstrate the advantage of our approaches using the example which motivated this work.

email: zhanghy@vt.edu

## CLAN: A NOVEL, PRACTICAL METHOD OF CURVATURE ASSESSMENT IN NONLINEAR REGRESSION MODELS

Jieru Xie\*, University of Louisville  
Linda Jane Goldsmith, University of Louisville

A novel, practical method of assessing nonlinearity behavior is developed to assess the extent of the nonlinearity in a nonlinear

regression model with design data points. We consider the geometric aspects of nonlinear regression modeling and use the familiar concept of confidence level as the criterion for nonlinearity assessment. The computation is based on the difference between the linear approximation inference ellipsoid region and the likelihood region, the often “banana-shaped” confidence region computed without the linear assumption. The method is applied to 23 published nonlinear datasets. It is found that the new method, CLAN (Confidence Level Assessment of Nonlinearity), is in good agreement with the root mean squared estimates of parameter effects and intrinsic nonlinearity introduced by Bates & Watts in their 1980s paper and book.

email: jieru.xie@louisville.edu

## BAYESIAN CASE INFLUENCE MEASURES AND APPLICATIONS

Hyunsoon Cho\*, 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

Three types of Bayesian case influence measures based on case deletion, namely phi-divergence, Cook's posterior mode distance and Cook's posterior mean distance are introduced to evaluate the effects of deleting a single observation in Bayesian regression models. The aim of this paper is to examine the statistical properties of these three diagnostic measures and their applications to model assessment. We derive their asymptotic approximations and establish their asymptotic equivalence. Moreover, we show that the sums of the proposed Bayesian case-deletion diagnostics measure model complexity, which is related to the effective number of parameters in the Deviance Information Criterion (DIC). We illustrate the proposed methodologies on generalized linear models and survival models. In addition, we present two real data examples to demonstrate the proposed methodology.

e-mail: hscho@bios.unc.edu

## A BAYESIAN CHI-SQUARED GOODNESS-OF-FIT TEST FOR CENSORED DATA MODELS

Jing Cao\*, Southern Methodist University  
 Ann Moosman, Patrick Air Force Base  
 Valen Johnson, University of Texas M.D. Anderson Cancer Center

We propose a Bayesian chi-squared model diagnostic for analysis of data subject to censoring. The test statistic has the form of Pearson's chi-squared test statistic and is easy to calculate from standard output of Markov chain Monte Carlo algorithms. The key innovation of this diagnostic is that it is based only on observed failure times. Because it does not rely on the imputation of failure times for observations that have been censored, we show that it can have higher power for detecting model departures than a comparable test based on the complete data. In a simulation study, we show that tests based on this diagnostic exhibit comparable power and better nominal Type I error rates than a commonly used alternative test proposed by Akritas (1988). An important advantage of the proposed diagnostic is that it applies to a broad class of censored data models, including generalized linear models and other models with non-identically distributed

and non-additive error structures. We illustrate the proposed model diagnostic for testing the adequacy of two parametric survival models for Space Shuttle main engine failures.

e-mail: jcao@smu.edu

## BAYES FACTOR CONSISTENCY IN LINEAR MODELS

Ruixin Guu\*, University of Missouri  
 Paul L. Speckman, University of Missouri

Liang et al. (2008) considered Bayesian model selection problem in the linear regression and showed consistency of the Bayes factor by using mixtures of g-priors, where they adapted the idea of Zellner and Siow (1980) to put flat priors on the common parameters and Zellner's g prior on the parameters only in the more complex model. In this case, the prior on g must be proper. Marin and Robert (2007) suggested to put Zellner's g-prior in each model so that an improper prior on g can be used. Later they suggested to use Jeffery's prior for g. In this paper, we adapt their idea to put the g-prior in each model, and we show consistency of the Bayes factor associated with the reference prior for g, which is improper, when model dimensions are fixed. We also discuss the consistency and inconsistency problems for the Bayes factor associated with this improper prior when comparing any model with fixed dimension and a model whose p can grow with n. We obtain consistency and inconsistency depending on the limiting behavior of p/n.

e-mail: rgd27@mizzou.edu

## SURROGATE DECISION RULE IN Q-LEARNING

Peng Zhang\*, University of Michigan  
 Bin Nan, University of Michigan  
 Susan A. Murphy, University of Michigan

Q-learning is a technique in reinforcement learning to learn an action function which maximizes the value function. We consider the problem where the action is binary and the decision rule is linear. By controlling the difference in the value function between the optimal decision and the surrogate decision, we developed the algorithm and theory to find the best surrogate decision rule in a subspace. This gives us another decision rule with the following properties: 1) It depends on fewer variables, and therefore is more feasible in practice; 2) The surrogate decision rule helps to eliminate covariates with no qualitative interactions; 3) The new decision usually has lower variance; 4) One can obtain a sequence of surrogate decisions at different levels of bounds on the difference in value, and choose one by the expert option.

e-mail: pczhang@umich.edu

# ABSTRACTS

## 70. JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

### JOINT MODELING OF MULTIVARIATE LONGITUDINAL DATA FOR MIXED RESPONSES AND SURVIVAL WITH APPLICATION TO MULTIPLE SCLEROSIS DATA

Pulak Ghosh\*, Emory University

Anneke Neuhaus, Sylvia Lawry Centre for Multiple Sclerosis Research  
Martin Daumer, Sylvia Lawry Centre for Multiple Sclerosis Research  
Sanjib Basu, Northern Illinois University

Multiple sclerosis (MS) is one of the most chronic neurological diseases in young adults with around 2.5 million affected individuals worldwide (Compston 2006). The most common presenting symptoms are inflammation of the optic nerve, weakness, sensory disturbances, gait disturbances and bladder dysfunction. So far only standard analysis methodology to estimate risks for relapse occurrence has been used. This includes mostly single endpoint survival analysis in which MRI information is shrunken to baseline values or aggregated measures such as means. In the present analysis we aim to establish a model that allows the description and prediction of occurrence of relapses by considering processes in the brain (visualized on T1 and T2 weighted MRI) simultaneously. These complex processes, together with clinical baseline information, have never been considered in one model so far. We will use our model to evaluate strength of dependencies of multivariate longitudinal MRI measures with the occurrence of MS relapses.

email: pulakghosh@gmail.com

### EVALUATING PREDICTIONS OF EVENT PROBABILITIES FROM A JOINT MODEL FOR LONGITUDINAL AND EVENT TIME DATA

Nicholas J. Salkowski\*, University of Minnesota  
Melanie M. Wall, University of Minnesota

Many clinical trials periodically measure participants until an event, generating longitudinal and event time data. Joint models simultaneously model the longitudinal data and the event times. One class of joint models uses latent variables to model associations between longitudinal trajectories and event risks. Both event times and longitudinal data influence the fitted values of the latent variables, which determine a subject's true trajectory and survival. We may wish to predict the probability of an event conditional on observed longitudinal measurements. Clinical heart failure trial data are used to evaluate approaches to predicting survival using longitudinal measurements of disease impact on daily life. The approach of Schoop et al. (2008) is used to compare the prediction errors from a joint model, a two-step model, a nonparametric Kaplan-Meier model, and a parametric Weibull model without covariates for several scenarios. The two-step model had the lowest prediction error, and the joint model had the highest prediction error. Simulations show that mean prediction error for the two-step model and the joint model are comparable under a scenario like the motivating data set.

email: salk0008@umn.edu

## SEMIPARAMETRIC ESTIMATION OF TREATMENT EFFECT WITH TIME-LAGGED RESPONSE IN THE PRESENCE OF INFORMATIVE CENSORING

Xiaomin Lu\*, University of Florida

Anastasios Tsiatis, North Carolina State University

In many randomized clinical trials, the primary response variable, for example, the survival time, is not observed directly after the patients enroll in the study but rather observed after some period of time (lag time). It is often the case that such a response variable is missing for some patients due to censoring that occurs when the study ends before the patient's response is observed or when the patients drop out of the study. It is often assumed that censoring occurs at random which is referred to as noninformative censoring; however, in many cases such an assumption may not be reasonable. If the missing data are not analyzed properly, the estimator or test for the treatment effect may be biased. In this paper, we use semiparametric theory to derive a class of consistent and asymptotically normal estimators for the treatment effect parameter which are applicable when the response variable is right censored. The baseline auxiliary covariates and post-treatment auxiliary covariates, which may be time-dependent, are also considered in our semiparametric model. These auxiliary covariates are used to derive estimators that both account for informative censoring and are more efficient than the estimators which do not consider the auxiliary covariates.

email: xlu2@phhp.ufl.edu

## SOME RESULTS ON LENGTH-BIASED AND CURRENT DURATION SAMPLING

Broderick O. Oluyede\* Georgia Southern University

In the analysis of longitudinal data, two semiparametric models that are often used are the Cox proportional hazards model and the accelerated failure time model. In Cox proportional hazards model for failure time, one assumes that the covariate effect is captured via a proportional constant between hazard functions, with unspecified underlying hazard functions. In accelerated hazards model, the hazards functions are related via the scale-time change, which is a function of covariates and the parameters. In a medical setting, current duration sampling require knowledge of the duration of the disease of a group of patients up to the present, but length-biased sampling requires time needed to observe the full duration of the disease of the sampled patients. In this talk, some results on current duration and length-biased sampling for the accelerated failure time model and Cox proportional hazards model are presented.

email: boluyede@georgiasouthern.edu

## DIAGNOSTIC TOOLS FOR JOINT MODELS FOR LONGITUDINAL AND TIME-TO-EVENT DATA

Dimitris Rizopoulos\*, Erasmus MC, Netherlands

The majority of the statistical literature for the joint modeling of longitudinal and time-to-event data has focused on the development of models that aim at capturing specific aspects of the motivating case

studies. However, little attention has been given to the development of diagnostic and model-assessment tools. The main difficulty in using standard model diagnostics in joint models is the non-random dropout in the longitudinal outcome caused by the occurrence of events. In particular, the reference distribution of statistics, such as the residuals, in missing data settings is not directly available and complex calculations are required to derive it. In this paper we propose a multiple-imputation-based approach for creating multiple versions of the completed data set under the assumed joint model. Residuals and diagnostic plots for the complete data model can then be calculated based on these imputed data sets. Our proposals are exemplified using two real data sets.

email: d.rizopoulos@erasmusmc.nl

## ON ESTIMATING THE RELATIONSHIP BETWEEN LONGITUDINAL MEASUREMENTS AND TIME-TO-EVENT DATA USING A SIMPLE TWO-STAGE PROCEDURE

Paul S. Albert\*, Biometric Research Branch National Cancer Institute  
Joanna H. Shih, Biometric Research Branch National Cancer Institute

Joint modeling of longitudinal measurements and time-to-event data is an active area of biostatistical research. Many of these methods require complex methodology which cannot easily be implemented with standard software packages. Recently, various authors have proposed a two-stage regression calibration approach which is simpler to implement than a joint modeling approach. In the first stage, the posterior expectation of an individual's random effects from a mixed model is fit without regard to the time-to-event data. In the second stage, the posterior expectation of an individual's random effects from the mixed-model are included as covariates in a Cox model. Although this approach is conceptually appealing, we demonstrate that this regression calibration approach may be biased due to informative dropout. We propose an alternative regression calibration approach which alleviates much of this bias. This new approach is developed for both discrete-time and continuous-time time-to-event data. Using simulation studies, we demonstrate that, in both cases, this new regression calibration procedure is nearly unbiased. Thus, this new regression calibration approach provides a simpler alternative to more complex joint modeling.

email: albertp@mail.nih.gov

## ADJUSTING FOR MEASUREMENT ERROR WHEN SUBJECT-SPECIFIC VARIANCE ESTIMATES ARE USED AS COVARIATES IN A PRIMARY OUTCOME MODEL

Laine E. Thomas\*, North Carolina State University  
Marie Davidian, North Carolina State University  
Stefanski A. Leonard, North Carolina State University

In biological studies of health effects a primary endpoint may be related to the longitudinal profiles of a continuous response. Outcome models where covariates are subject-specific random effects have been well studied (Li, Zhang and Davidian 2004). Instead, we may want to understand the relationship between disease outcome and

subject-specific variability over time (Yang et al 2007, Lyles et al 1999). Substantial bias can occur when variance estimates are imputed as covariates in the outcome model. To account for this, Lyles et al 1999 developed a maximum likelihood method which assumes that the subject-specific variances come from a lognormal distribution. We compare two approaches that require no assumptions on the distribution of the subject-specific variances. We adapt the conditional score strategy of Stefanski and Carroll (1987) to account for error in the variance estimators which has a chi-square distribution. This method provides consistent parameter estimation but is limited to a particular specification of the outcome model. An alternative approach is motivated by the unpublished thesis of Bay (1997) which adjusts the estimates to have unbiased sample moments for the distribution of subject-specific variances in the population. This has the benefit of being easily implemented in a variety of outcome models.

email: leellio2@ncsu.edu

**TUESDAY, MARCH 17, 2009  
10:15 AM -12:15 PM**

## 71. PRESIDENTIAL INVITED ADDRESS

### STATISTICAL MODELLING FOR REAL-TIME EPIDEMIOLOGY

Peter Diggle, Ph.D.  
Lancaster University School of Health and Medicine and Johns Hopkins University School of Public Health

Large volumes of data on a range of health outcomes are now collected routinely by many health care organisations but, at least in the UK, are often not analysed other than for retrospective audit purposes. Each data-record will typically be referenced both in time and in space; for example, in the UK the temporal reference will be a date, and in some cases a time of day, whilst the spatial reference will usually be the individual's post-code which, in urban settings, corresponds to a spatial resolution of the order of 100 metres. By real-time epidemiology, I mean the analysis of data-sets of this kind as they accrue, to inform clinical or public health decision-making. Such analyses would be triggered and the results posted automatically, for example on a web-site, by the arrival of new data. In this talk I will review work in spatial, temporal and spatio-temporal modelling that seems especially relevant to this general task, and will describe a number of applications, including some or all of: real-time syndromic surveillance (Diggle, Rowlingson and Su, 2005); tropical disease prevalence mapping (Crainiceanu, Diggle and Rowlingson, 2008); early warning of incipient renal failure in primary care patients (Diggle and Sousa, 2009).

email: p.diggle@lancaster.ac.uk

# ABSTRACTS

TUESDAY, MARCH 17, 2009

1:45 - 3:30 PM

## 72. PREDICTION AND CURE MODELING IN MODERN MEDICAL DATA ANALYSIS

### TRANSFORMATION MODELS WITH GAMMA-FRAILTY FOR MULTIVARIATE FAILURE TIMES

Joseph G. Ibrahim\*, University of North Carolina  
Donglin Zeng, University of North Carolina  
Qingxia Chen, Vanderbilt University

We propose a class of transformation models for multivariate failure times. The class of transformation models generalize the usual gamma-frailty model and yields a marginally linear transformation model for each failure time. Nonparametric maximum likelihood estimation is used for inference. The maximum likelihood estimators for the regression coefficients are shown to be consistent and asymptotically normal, and their asymptotic variances attain the semiparametric efficiency bound. Simulation studies show that the proposed estimation procedure provides asymptotically efficient estimates and yields good inferential properties for small sample sizes. The method is illustrated using real data from a cardiovascular study.

email: ibrahim@bios.unc.edu

### PREDICTION OF U.S. MORTALITY COUNTS USING SEMIPARAMETRIC BAYESIAN TECHNIQUES

Ram Tiwari\*, U.S. Food and Drug Administration

Accurate prediction of cancer mortality figures for the current and upcoming year are extremely essential for public health planning and evaluation. Due to delay in reporting cause-specific mortality for the US, there is a 3-year lag between the latest year for which such figures are available and the current year. Prior to 2004, the American cancer Society (ACS) used to predict cancer mortality counts by first fitting a time series model with quadratic trend and autoregressive error to the past data and then projecting this model into the future. Beginning 2004 the ACS has begun to implement a new methodology in its annual publication Cancer facts & Figures, 2004. This method, known as the state-space method (SSM), uses a quadratic trend with random time-varying coefficients to model the mortality counts. In this talk, Bayesian versions of the SSM are presented. In particular, we present two models for short-term prediction of the number of deaths that arise from common cancers in the United States. The first is a local linear model, in which the slope of the segment joining the number of deaths for two consecutive time periods is assumed to be random with a nonparametric distribution, which has a Dirichlet process prior. For slightly, longer prediction periods, we present a local quadratic model. The hierarchical and nested Dirichlet models are also considered. Bayesian models are compared with both the SSM and the previous ACS method.

email: lekyla.whitaker@fda.hhs.gov

## A FAMILY OF CURE MODELS

Jeremy MG Taylor\*, University of Michigan  
Ning Smith, University of Michigan

Cure models are a useful approach for failure time data when it is known that a fraction of the subjects would never experience the event of interest even if they could be followed for a long time. Two general types of cure models have been developed. A mixture cure model, developed by Boag, Farewell and others, and bounded cumulative hazard cure model, developed by Yakovlev, Tsodikov, Chen and others. In this paper we present a family of cure models indexed by an extra parameter in which both of these types are special cases. The family involves a Box-Cox transformation of a distribution function, and the extra parameter is the power parameter of the Box-Cox family. We demonstrate that large sample sizes will be needed to distinguish between the two type of cure models.

email: jmgt@umich.edu

## ANALYSIS OF CURE RATE SURVIVAL DATA UNDER PROPORTIONAL ODDS MODEL

Debajyoti Sinha\*, Florida State University,  
Sudipto Banerjee, University of Minnesota  
Yu Gu, Florida State University

With rapid improvements in medical treatment and health care, many survival data sets now reveal a substantial portion of patients who are cured (that is, who never experience the endpoint). Extended survival models called cure rate models account for the probability of a subject being cured. Our present work proposes a new class of cure rate models that has a proportional odds structure as a function of the covariates. This class also has some unique properties which are different from those of classical proportional odds survival models. In this article we address issues such as regression effects on the cure fraction, associated Bayesian analysis and propriety of the associated posterior distributions under different modelling assumptions. Finally, we illustrate with reanalysis of two data sets (one on melanoma and the other on breast cancer) our model's distinguishing features of our models and implementation of Bayesian data analytic tools for this model. We also develop a new set of methods for Bayesian model selection and model assessment for the cure-rate as well as classical survival models.

email: sinhad@stat.fsu.edu



## **73. NETWORK ANALYSIS MODELS, METHODS AND APPLICATIONS**

### **NEURAL FUNCTIONAL CONNECTIVITY NETWORKS**

Crystal Linkletter\*, Brown University  
Hernando Ombao, Brown University  
Mark Fiecas, Brown University

Network models are increasingly being used to represent structure in complex networks. One application of such models is to explore the impact of networks as substrates for dynamic processes (e.g. the spread of an infectious disease). That is, given a network which constrains the flow of the process, the resulting behavior is explored. In this talk, we consider the inverse problem: Can observing a process happening on a network inform us about the underlying network? In neuroscience, functional connectivity between different regions of the brain can be represented as a network. The presence of correlation between fMRI scans taken at two regions is indicative of functional connectivity between those regions. We propose a model-based approach to estimating functional connectivity in the brain based on the premise of uncovering underlying latent network structure.

email: cdlinkle@stat.brown.edu

### **NETWORK FILTERING WITH APPLICATION TO DETECTION OF GENE DRUG TARGETS**

Eric D. Kolaczyk\*, Boston University

A canonical problem in statistical signal and image processing is the detection of faint targets against complex backgrounds, which has been likened to the proverbial task of ‘finding a needle in a haystack’. We consider the task of target detection when the ‘background’ is neither one- or two-dimensional but rather in the form of an association network. We model the acquisition of network data, including the potential presence of targets, using a system of sparse simultaneous equation models (SSEM). In this context, detection is approached as a two-step procedure, involving (i) statistical inference and removal of ‘background’ network structure, using tools of sparse inference, and (ii) outlier detection in the network-filtered residuals. Theoretical performance of the methodology can be characterized using a combination of tools and concepts from sparse inference, compressive sampling, random matrix theory, and spectral graph theory. We illustrate the practical capabilities of this approach using simulations and the problem of drug target detection in the context of a network of gene interactions.

email: kolaczyk@math.bu.edu

### **EXPONENTIAL-FAMILY RANDOM GRAPH MODELS FOR BIOLOGICAL NETWORKS**

David Hunter\*, Penn State University

In the field of social networks, the use of Exponential-Family Random Graph Models, or ERGMs, is becoming increasingly popular as the theoretical understanding and software implementations of these

models improve. Recently, researchers have begun to use the same tools to explore biological networks as well. After giving a brief introduction to ERGMs, this talk critically examines recent work applying ERGMs to biological networks, offers some suggestions for future work in this field, and provides a brief illustration of analysis of a biological network using ERGMs.

email: dhunter@stat.psu.edu

## **74. STATISTICAL METHODS IN GENOME-WIDE GENE REGULATION STUDIES**

### **INTEGRATIVE MODELING OF TRANSCRIPTION AND EPIGENETIC REGULATION**

Xiaole Shirley Liu\*, Harvard University

High throughput chromatin immunoprecipitation experiments (ChIP-chip and ChIP-seq) have accelerated transcription regulation studies, yet created challenges for data analyses. I will discuss a few algorithms we have developed for the analysis of ChIP-seq, and report some observations about epigenetic regulation. I will also discuss an integrative algorithm we used to assign regulated gene targets and predict gene expression changes from binding.

email: xsliu@jimmy.harvard.edu

### **LEARNING GENE REGULATORY NETWORK PROFILE ACROSS MULTIPLE EXPERIMENTAL CONDITIONS**

Qing Zhou\*, UCLA  
Michael J. Mason, UCLA

Gene regulatory network changes dynamically across different cellular conditions. When transcription factor binding data (ChIP-chip/seq) and gene expression data are generated in multiple related experimental conditions or developmental stages, inferring a set of related regulatory networks is possible. We propose a flexible statistical model for this inference problem. Under each experimental condition, we assume that the binding of a transcription factor (TF) to a gene is determined by the expression of the TF under the same condition and the enrichment of its binding sites in the regulatory region of the gene. Multiple measures can be incorporated into a unified framework. Advanced learning methods based on ensemble learning or L1 norm penalty are employed to identify regulators in a set of networks. This approach is applied to the reprogramming of mouse embryonic stem cells from fibroblasts.

email: zhou@stat.ucla.edu

# ABSTRACTS

## A HIERARCHICAL SEMI-MARKOV MODEL FOR DETECTING ENRICHMENT WITH APPLICATION TO CHIP-SEQ EXPERIMENTS

Sunduz Keles\*, University of Wisconsin, Madison  
Pei Fen Kuan, University of Wisconsin, Madison

Protein-DNA interactions play a fundamental role in gene regulation. Significant progress has been made in profiling transcription factor binding sites and histone modifications with ChIP-chip, and more recently ChIP-Seq experiments. Despite the numerous model based approaches developed for the analysis of ChIP-chip data, limited statistical tools are available for ChIP-Seq data. We develop a comprehensive model based approach for detecting enrichment from ChIP-Seq experiments with a hierarchical semi-Markov model. The proposed model is applicable for the analysis of experiments with (1) single ChIP-Seq sample with and without input control and (2) multiple ChIP-Seq samples. We introduce a new meta analysis approach for controlling the FDR at peak level and allow for the boundaries of the binding sites to be declared probabilistically. This bypasses the common heuristic postprocessing methods to merge contiguous bins as peaks. We also propose and investigate various models for observed tag counts from ChIP-Seq experiments, that account for sequencing depths and biases due to amplification and sequence-specific affinities. Although our discussion is dedicated to ChIP-Seq experiments, the proposed hierarchical semi-Markov model can be applied to other types of data which exhibit spatial structure (e.g., ChIP-chip), by modifying the emission distributions.

email: keles@stat.wisc.edu

## A CORRELATION MOTIF BASED HIDDEN MARKOV MODEL FOR POOLING INFORMATION FROM MULTIPLE CHIP-CHIP EXPERIMENTS

Hongkai Ji\*, Johns Hopkins Bloomberg School of Public Health  
Hao Wu, Johns Hopkins Bloomberg School of Public Health

Chromatin immunoprecipitation coupled with genome tiling arrays (ChIP-chip) is a widely used technology to identify transcription factor binding sites in the genome. It is common to obtain noisy data from such experiments due to various reasons such as unoptimized protocols, varying qualities of antibodies produced at different time, etc. Due to the relatively high cost, most ChIP-chip experiments contain only a few replicates. Reliably detecting transcription factor binding sites is challenging when data are noisy and the number of replicates within the experiment is small. However, when the same transcription factor has been studied by multiple labs, it is possible to pool information from multiple data sets to improve the inference for each individual data set. Effective information pooling depends on adequate but parsimonious modeling of the unknown correlation structures among data sets. We propose a correlation motif based hidden Markov model to jointly identify such correlations and transcription factor binding sites. Both simulations and real data analyses show that the proposed method can substantially increase the sensitivity and specificity of transcription factor binding site detection compared to the approach that analyzes each data set individually.

email: hji@jhsph.edu

## 75. EXPERIMENTAL DESIGNS IN DRUG DISCOVERY & CLINICAL TRIALS

### PENALIZED DESIGNS OF MULTI-RESPONSE EXPERIMENTS

Valerii V. Fedorov\*, GlaxoSmithKline  
Rongmei Zhang, GlaxoSmithKline

The major theme of the presentation is optimal design of dose-response experiments when several (continuous or categorical) potentially correlated responses are observed simultaneously for every experimental unit. While the narration is built around two endpoints the generalization for the higher dimension is also discussed. I also address some ethical aspects of dose response experiments. In the traditional optimal design setting one tries to gain as much information as possible without explicit concern about patients in the trial, i.e. doing what is best for the targeted population (collective ethics). The currently popular procedures gravitate to individual ethics: doing what is best (accordingly to the current knowledge) for a newly arriving patient. To compromise between these two extremes we maximize information per a unit of penalty, which depends on efficacy and toxicity. Necessary and sufficient conditions, algorithms and software are developed and discussed for locally optimal, composite and adaptive designs.

email: valeri.v.fedorov@gsk.com

### ASPECTS OF OPTIMAL DOSE RESPONSE DESIGN

Randall D. Tobias\*, SAS Institute Inc.  
Alexander N. Donev, University of Manchester

Dose response experiments are involved early and late in the development of new drugs. At the early stages of drug development, a single design is used to screen many different compounds for biologically interesting activity; at the late stages, confirmatory bioassay experiments are performed on compounds whose properties are relatively well-known *a priori*. In this talk we will discuss applications of optimal design theory and techniques to both of these problems. In the confirmatory case, we will show how useful properties of the minimum support designs revealed by optimal design theory can lead to cost-efficient experiments. At the other end, lack of prior knowledge makes screening experiments more problematic from a theoretical point of view, but we can show that the serial dilution designs that are universally employed in practice have good Bayesian optimality properties.

email: Randy.Tobias@sas.com

### AN ADAPTIVE OPTIMAL DESIGN FOR THE EMAX MODEL AND ITS APPLICATION IN CLINICAL TRIALS

Sergei Leonov\*, GlaxoSmithKline  
Sam Miller, GlaxoSmithKline

We discuss an adaptive design for a first-time-in-humans dose-escalation study. The project team wished to maximize the efficiency of

the study by using doses targeted at maximizing information about the dose-response relationship within certain safety constraints. We have developed a GUI-based adaptive optimal design tool to recommend doses when the response follows an Emax model, with functionality for pre-trial simulation and in-stream analysis. We present the results of a simulation to investigate the operating characteristics of the applied algorithm and discuss potential extensions for other models.

email: Sergei.2.Leonov@gsk.com

## 76. STATISTICAL METHODS FOR FLOW CYTOMETRY DATA

### AUTOMATED FEATURE EXTRACTION FOR FLOW CYTOMETRY

Errol Strain, BD Technologies  
Perry Haaland, BD Technologies

Flow Cytometry High Content Screening (FC-HCS) experiments generally focus on testing a small number of markers on a large number of samples. FC-HCS analyses are performed on 96 or 384 well microtiter plates, and cells are stained with the same panel of antibodies in each well. Feature extraction is used to find cell populations in the multidimensional fluorescence signal space. Interesting populations are ones that change in response to treatment conditions, or show variability among cell sources or donors. BD FACS™ CAP is a different type of FC-HCS that looks at a large number of markers (212) on a small number of samples. The goal of a BD FACS™ CAP experiment is to develop profiles of surface markers that are expressed on a particular cell types. Determining whether or not a particular marker is actually present is difficult since each of the 212 antibodies can have a different level of background staining, and we do not have a matched cell type that is negative in expression for all markers. Instead of non expressing cells, we use non-specific antibodies, or isotypes, that are conjugated to the same fluorophore as the test antibodies for our negative controls. Features extracted for each marker include the fraction of cells with staining greater than the isotype threshold, and the ratio of the median fluorescence intensity (MFIs) for the test antibody to its associated isotype. The fraction of positive cells indicates whether or not the marker is expressed (off or on) while the MFI ratio gives an idea of the expression level (low or high). The basic framework for automating a large portion of plate-based FC-HCS analysis has been implemented in the Bioconductor package plateCore.

email: errol\_strain@bd.com

### BIOCONDUCTOR TOOLS FOR HIGH-THROUGHPUT FLOW-CYTOMETRY DATA ANALYSIS

Florian M. Hahne\*, Fred Hutchinson Cancer Research Center

Automation technologies developed during the last several years have enabled the use of flow cytometry high content screening (FC-HCS) to generate large, complex datasets in both basic and clinical research applications. A serious bottleneck in the interpretation of existing studies and the application of FC-HCS to even larger, more complex

problems is that data management and data analysis methods have not advanced sufficiently far from the methods developed for applications of flow cytometry (FCM) to small-scale, tube-based studies. Some of the consequences of this lag are difficulties in maintaining the integrity and documentation of extremely large datasets, assessing measurement quality, developing validated assays, controlling the accuracy of gating techniques, automating complex gating strategies, and aggregating statistical results across large study sets for further analysis. In this presentation, we introduce a range of computational tools developed in Bioconductor that enable the automated analysis of large flow cytometry data sets, from the initial quality assessment to the statistical comparison of the individual samples.

email: fhahne@fhcrc.org

### CHARACTERIZING IMMUNE RESPONSES VIA FLOW CYTOMETRY

Martha Nason\*, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Intracellular Cytokine Staining, a specific application of flow cytometry, can be used to measure the secretions of individual blood cells that have been exposed to an antigen. When applied to hundreds of thousands of cells in a particular blood sample, this technique allows a characterization of the immune response an individual makes to a virus or a component of a virus. For viruses such as HIV, the features that define a beneficial immune response are not entirely clear, and therefore the kind of response should be targeted by vaccine developers is likewise still nebulous. This talk will describe some of the current aspects of individual immune responses to HIV under investigation, and explore some measures and methods for characterizing the features of these immune responses.

email: mnason@niaid.nih.gov

### AUTOMATED GATING OF FLOW CYTOMETRY DATA VIA ROBUST MODEL-BASED CLUSTERING

Kenneth Lo, University of British Columbia  
Ryan Brinkman, BC Cancer Agency  
Raphael Gottardo\*, Clinical Research Institute of Montreal

The capability of flow cytometry to offer rapid quantification of multidimensional characteristics for millions of cells has made this technology indispensable for health research and medical diagnosis. However, the lack of statistical and bioinformatics tools to parallel recent high-throughput technological advancements has hindered this technology from reaching its full potential. We propose a flexible statistical model-based clustering approach for identifying cell populations in flow cytometry data based on t mixture models with a Box-Cox transformation. This approach generalizes the popular Gaussian mixture models to account for outliers and allow for non-elliptical clusters. We describe an Expectation-Maximization (EM) algorithm to simultaneously handle parameter estimation and transformation selection. Using two publicly available datasets, we demonstrate that our proposed methodology provides enough flexibility and robustness to mimic manual gating results performed by an expert researcher. The proposed clustering methodology is well-

# ABSTRACTS

adapted to automated analysis of flow cytometry data. It tends to give more reproducible results, and helps reduce the significant subjectivity and human time cost encountered in manual gating analysis.

email: raphael.gottardo@ircm.qc.ca

## 77. MULTIPLE TESTING IN GENOME-WIDE ASSOCIATION STUDIES

### WINNER'S CURSE AND BIAS CORRECTION IN GENOME-WIDE ASSOCIATION AND CANDIDATE GENE STUDIES

Lei Shen\*, Eli Lilly and Company

Recently there has been increasing attention to the problem induced by multiplicity in large-scale analyses known as the Winner's Curse, referring to the over-estimation of effect sizes of the top candidate biomarkers inferred by the analysis, which can be regarded as a form of selection bias. In this talk, we focus on genome-wide association and candidate gene studies, in which a large number of potential genetic biomarkers are routinely screened for their associations with clinical endpoints. Typically additional investment, often in the form of follow-up studies, is made for the top candidate biomarkers. We study the severity of the bias and its implications to a clinical program of drug development, such as resource allocation, business decisions and planning of confirmatory studies, in some typical settings. We then look closely at potential questions of interest and determine their relevance to pharmaceutical research. Finally we investigate some methods to correct for the selection bias and compare their performance.

email: shen\_lei@lilly.com

### EXTENDED HOMOZYGOSITY SCORE TESTS TO DETECT POSITIVE SELECTION IN GENOME-WIDE SCANS

Ming Zhong\*, Texas A&M University  
Kenneth Lange, UCLA  
Jeanette C. Papp, UCLA  
Ruzong Fan, Texas A&M University

In this article, we develop test statistics to detect excess homozygosity: (a) an extended genotype-based homozygosity test (EGHT), (b) a hidden Markov model test (HMMT), and (c) an extended haplotype-based homozygosity test (EHHT). The null hypothesis of all three tests assume random mating and Hardy-Weinberg equilibrium (HWE). They differ in how they treat linkage disequilibrium. The null hypothesis of EGHT assumes linkage equilibrium in addition to HWE. The EHHT conditions on existing linkage disequilibrium and it tests haplotype version of HWE. The HMMT stands between these two extremes and assumes pairwise but no higher-order disequilibrium interactions. We evaluate by simulation the false positive rates for the EGHT and HMMT under the null hypothesis of EGHT and find that HMMT is more conservative. All three methods are then applied to the HapMap Phase II data. We are able to replicate previous findings of strong positive selection in 19 autosome genomic regions

out of 20 candidates. Over the entire genome, our EGHT and HMMT statistics score lowest in African sample (YRI) and highest in east Asian sample (CHB+JPT), with European sample (CEU) intermediate. Across the genome, EHHT scores are generally low with sharp spikes in only a few regions. Based on the high EHHT scores and population differentiations, we identify new candidate regions which may undergo recent selection.

email: rfan@stat.tamu.edu

## BAYESIAN ASSOCIATION TESTING OF SNP MARKERS AND WOOD CHEMISTRY TRAITS IN CLONAL TRIALS OF LOBLOLLY PINE

Xiaobo Li\*, University of Florida  
Dudley A. Huber, University of Florida  
George Casella, University of Florida  
David B. Neale, University of California-Davis  
Gary F. Peter, University of Florida

With the advance of the sequencing technology, single nucleotide polymorphism (SNP) markers are now available for genome wide association testing in our loblolly pine population. We focus on the association testing using Hierarchical Bayesian Models for genome wide association for wood chemistry traits. A total number of 7,600 SNPs, which represents approximately 6,500 genes within the loblolly pine genome sequenced for 999 genotypes, were used to test association. A small set of 46 SNPs from the previous study was used to test the model and simulation to test the model was done in this project.

email: xbli@ufl.edu

## WITHIN-CLUSTER RESAMPLING (MULTIPLE OUTPUTATION) FOR ANALYSIS OF FAMILY DATA: READY FOR PRIME-TIME?

Hemant K. Tiwari\*, University of Alabama at Birmingham  
Amit Patki, University of Alabama at Birmingham  
David B. Allison, University of Alabama at Birmingham

Hoffman et al. (2001) proposed an elegant resampling method for analyzing clustered binary data. The focus of their paper was to perform association tests on clustered binary data using within-cluster-resampling (WCR) method. Follman et al., (2003) extended Hoffman et al.'s (2001). Follmann et al. (2003) termed their procedure multiple outputation because all excess data within each cluster is thrown out multiple times. Herein, we refer to this procedure as WCR-MO. For any statistical test to be useful for a particular design, it must be robust, have adequate power, and be easy to implement and flexible. WCR-MO can be easily extended to continuous data and is a computationally intensive but simple and highly flexible method. Considering family as a cluster, one can apply WCR to familial data in genetic studies. Using simulations, we evaluated WCR-MO's robustness for analysis of a continuous trait in terms of type I error rates in genetic research. WCR-MO performed well at the 5%  $\pm$ -level. However, it provided inflated type I error rates for  $\pm$ -levels less than 5% implying the procedure is liberal and may not be ready for

application to genetic studies where  $\alpha$  levels used are typically much less than 0.05.

email: htiwari@uab.edu

## ESTIMATION OF THE CONTRIBUTION OF RARE DISEASE-PREDISPOSING VARIANTS TO COMPLEX DISEASES

Weihua Guan\*, University of Michigan  
Laura J. Scott, University of Michigan  
Michael Boehnke, University of Michigan

For many complex diseases, common risk variants identified through genome-wide association studies explain only a small proportion of disease risk. Rapid advances in next-generation sequencing technologies will allow a more complete survey of genetic variants in the genomic regions of interest, and an opportunity to identify rare disease-predisposing variants. It is often assumed that a candidate gene contains a number of such variants to cause noticeable effect on the disease. One test for the association of the rare variants with disease status is to compare the total number of variants within a gene or a region between cases and controls. When affected family members are available, an additional approach is to look for co-transmission of disease with the possible linked rare alleles. Through analytically calculations and computer simulations, we have estimated the plausible numbers of linked disease-predisposing variants given the observed signal using a family-based test, assuming all these variants have low allele frequencies but moderate to high disease penetrance. We compared the power of the family-based test to a case-control approach. Our results will provide help to understand the importance of the common diseases, rare variants assumption in practice and its impact on the analyses of re-sequencing data.

email: wguan@umich.edu

## GENOMICS OF COMPLEX DISEASES

Li Luo\*, The University of Texas School of Public Health  
Gang Peng, School of Life Science, Fudan University  
Eric Boerwinkle, University of Texas School of Public Health  
Momiao Xiong, The University of Texas School of Public Health

Recent deep-resequencing reveals that there are a large number of rare variants that play an important role in causing complex diseases. Most traditional statistical methods in current genome-wide association studies (GWAS) have mainly focused on investigation of common variants individually. Due to their rarity, to individually test association of rare variants with diseases has little power and may not be robust. In this report, we propose to collectively analyze multiple rare variants as a general framework for association studies of rare variants. We develop three novel statistics that employ information about Poisson process characterization of occurrence of rare variants along a segment of genome. The number of rare variants each individual carries can be approximated by a Poisson random variable. The intensity of Poisson process can be interpreted as the mutation rate. The developed statistics compare the mutation rates between cases and controls. To study validity and evaluate performance of the proposed statistics, we estimate type I error rates by assuming infinite

allele models and large simulation; we also compare power with the standard chi-square test and apply them to real examples. Preliminary results show that the proposed statistics have higher power and smaller P-value than the standard chi-square test.

email: li.luo@uth.tmc.edu

## REDUCING COSTS OF TWO STAGE GENOME WIDE ASSOCIATION STUDIES

Michael D. Swartz\*, University of Texas M.D. Anderson Cancer Center  
Sanjay Shete, University of Texas M. D. Anderson Cancer Center

Genome Wide Association (GWA) Studies continue to increase in popularity. Although the cost of genotyping continues to go down, GWA studies continue to carry a high price tag. This is due to the large sample of individuals required to detect the small effect contributed to disease risk from individual SNPs. As a result, methodologies for GWA studies continue to strike a balance between cost and power. This presentation proposes a two stage design that reduces the cost of a GWA study by reducing the genotyped individuals in stage 2 without sacrificing power. We introduce an ascertainment scheme for individuals in stage 1 such that only a subset of the stage 1 individuals are brought forward to stage 2. We used the simulated data from the Genetic Analysis Workshop 15 to evaluate our method and compare its performance to the typical two stage design. Our simulation studies show that by ascertaining individuals from stage 1 from cases and controls neither increases false positives nor decreases power, and still can substantially reduce the cost of the study.

email: mdswartz@mdanderson.org

## 78. ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

### MEDIAN POLISH ALGORITHMS FOR AUTOMATED ANOMALY DETECTION IN ENVIRONMENTAL SENSOR NETWORKS

Ernst Linder\*, University of New Hampshire  
Zoe Cardon, Marine Biology Laboratory  
Woods Hole, Marine Biology Laboratory  
Jared Murray, University of New Hampshire

Environmental sensors malfunction often due to a variety of reasons, such as power failure, hardware deterioration, or interference by animals. Sensor networks are typically deployed for the purpose of new discovery, thus, an a-priori statistical model of the underlying scientific reality is not available. Therefore it is a challenge to decide which data represent typical behavior and which data is anomalous or extreme, indicating sensor malfunctioning. For contemporaneous sensor responses we propose to fit a sequential median polish. The two factors of the median polish are sensor effect, and time effect. We apply two outlier rules for automated anomalous data detection: One for extreme traces, and another for extreme residuals. We train the algorithm for various outlier thresholds relative to a manually cleaned data from 147 psychrometers that were utilized in an experiment on water uptake, transpiration and redistribution of 21 sage plant in a

# ABSTRACTS

semi-arid area in Utah during the summer of 2007. We also examine the use of median polish algorithms for discovery of anomalous and interesting behavior of individual plants.

email: elinder@unh.edu

## DATA ASSIMILATION FOR PREDICTION OF FINE PARTICULATE MATTER

Ana G. Rappold\*, U.S. Environmental Protection Agency  
Marco A. Ferreira, University of Missouri - Columbia

A substantial portion of air quality research and management is concerned with the nature of adverse health effects associated with exposure to fine particulate matter (PM<sub>2.5</sub>). Current assessment of population based daily exposure is limited by the spatial and temporal availability of the monitoring networks. There is a strong interest in developing methodology for assimilation of the observed data and mathematical model predictions such as EPA's Community Multi-scale Air Quality model (CMAQ). CMAQ is a deterministic model of atmospheric pollutant transport which provides volume averaged predictions on the grid. Although such model predictions are biased they contribute important spatial and temporal features by information such as weather patterns, land use, emissions, etc. The process of data assimilation involves several important statistical problems such as integration of areal and point data, and treatment of spatial and temporal similarities in data as well as bias. To address these problems, we develop a Bayesian hierarchical model for space-time assimilation of two types of data. The joint distribution of the data is modeled as a spatial process dynamically evolving through time. We provide insight in spacially varying bias associated with CMAQ data by introducing additive multi-scale spatio-temporal model for bias.

email: rappold.ana@epa.gov

## BAYESIAN MODEL AVERAGING APPROACH IN HEALTH EFFECTS STUDIES

Ya-Hsiu Chuang\*, University of Pittsburgh  
Sati Mazumdar, University of Pittsburgh

Determining the lagged effects of ambient air levels of a pollutant on cardiac distress is important in health effect studies. Standard model selection procedures where a set of predictor variables is selected ignore the associated uncertainties and may lead to overestimation of effects. Bayesian model averaging approach takes account of model uncertainty by combining information from all possible models. Zellner's g-prior containing a hyperparameter g can account for model uncertainty and has potential usefulness in this endeavor. We present results from a sensitivity analysis for Bayesian model averaging with different calibrated hyperparameter g, viz., Akaike Information Criterion prior, Bayes Information Criterion prior, and Local Empirical Bayes estimate. Data from Allegheny County Air Pollution Study and the simulated data sets are used.

email: yac14@pitt.edu

## EFFECT MODIFICATION OF PRENATAL MERCURY EXPOSURE ASSOCIATION WITH DEVELOPMENTAL OUTCOMES BY SOCIAL AND ENVIRONMENTAL FACTORS

Tanzy M. Love\*, University of Rochester Medical Center  
Sally Thurston, University of Rochester Medical Center

The Seychelles Child Development Study (SCDS) is testing the hypothesis that prenatal exposure to low doses of MeHg from maternal consumption of fish is associated with the child's developmental outcomes. No deleterious relationships between exposure to MeHg and cognitive functions have been identified in the primary analysis of the main cohort. However, secondary regression analysis of this cohort found a small effect modification by both caregiver IQ and household socio-economic status(SES) score (Davidson et al. 1999). They showed that children with higher IQ caregivers and higher SES had a significant positive relationship between Mercury levels and intelligence. We use latent classification techniques and a new cohort of children with measured nutrient information (particularly long-chain fatty acids like Omega 3) to further examine the relationship.

email: tanzy@cmu.edu

## LATENT SPATIAL MODELLING FOR SPECIES ABUNDANCE

Avishek Chakraborty\*, Duke University  
Alan E. Gelfand, Duke University

Modeling abundance pattern for one or more species using environmental features is of increasing importance in current ecological studies. The Cape Floristic Region (CFR) in South Africa provides a rich class of species data for such modeling. Here we propose a two stage Bayesian hierarchical model for explaining the species abundance over the region. Ordinally categorized abundance figures are given for about 10000 grid cells along with cell-wise environmental and soil-type factors. We formulate the empirical abundance pattern as a degraded version of the potential pattern, with the degradation effect coming from land transformation and classification error. Since most of the CFR region was sparsely sampled, the potential abundance is of interest from a predictive as well as conservation perspective. An areal level spatial regression model was used for modeling the dependence of species abundance on the environmental factors. Categorical abundance statistics was induced by a continuous latent surface and a conditionally autoregressive prior (CAR) was specified for the distribution of spatial random effects. Parallelized computation techniques were employed to improve the runtime. Various types of inference such as comparing different parts of region in terms of species richness, mutual comparison of two or more species etc. naturally follow from our model.

email: ac103@stat.duke.edu

## A FLEXIBLE REGRESSION MODELING FRAMEWORK FOR ANALYZING SEAGRASS AREAL COVERAGE AS CHARACTERIZED BY BRAUN-BLANQUET (BB) VEGETATION COVER SCORES

Paul S. Kubilis\*, University of Florida  
Mary C. Christman, University of Florida  
Penny Hall, Florida Fish and Wildlife Research Institute

The Florida Fish and Wildlife Conservation Commission (FWC) has been monitoring several species of seagrass in Florida Bay since 1995. Of interest is whether the areal coverage and extent of individual species have been changing over time, particularly in response to recent efforts at restoring natural fresh water flow through the Everglades. Seagrass bottom cover sampling is done using a spatially stratified design with restricted randomization of sampling locations within strata. At each location, 4-12 subsamples (quadrats) are observed and species presence/absence is recorded. If a particular species is present, the Braun-Blanquet (BB) vegetation ground cover score, an unequally spaced ordinal categorical variable, is used to describe the proportion of bottom area covered by the seagrass species within each quadrat. By viewing these BB-scores as interval-censored observations of a continuous random variable representing percent bottom cover, we have developed a flexible regression modeling framework for characterizing the status and trend of BB-score seagrass abundance. This model-based approach assumes an underlying Bernoulli distribution for the presence/absence of a seagrass species within a sampling quadrat, and an underlying Beta distribution for the true proportion of cover when the species is present.

email: pkubilis@ufl.edu

## ON SHORTEST PREDICTION INTERVALS IN LOG-GAUSSIAN RANDOM FIELDS

Victor De Oliveira\*, University of Texas at San Antonio and Changxiang Rui, University of Arkansas

This work considers the problem of constructing prediction intervals in log-Gaussian random fields. New prediction intervals are derived that are shorter than the standard prediction intervals of common use, where the reductions in length can be substantial in some situations. We consider both the case when the covariance parameters are known and unknown. For the latter case we propose a bootstrap calibration method to obtain prediction intervals with better coverage properties than the plug-in (estimative) prediction intervals. The methodology is illustrated using a spatial dataset consisting of cadmium concentrations from a contaminated region in Switzerland.

email: victor.deoliveira@utsa.edu

## 79. CATEGORICAL DATA ANALYSIS

### ON TESTS OF HOMOGENEITY FOR PARTIALLY MATCHED-PAIR DATA

Hani Samawi, College of Public Health, Georgia Southern University  
Robert L. Vogel\*, College of Public Health, Georgia Southern University  
Wilson A. Koech, College of Public Health, Georgia Southern University  
Jiann-Ping Hsu College of Public Health, Georgia Southern University

In this project, we are investigating several tests of homogeneity between two groups when the data is partially matched-pair by comparing their power. Also we propose a weighted test of homogeneity based on Pearson and McNemar chi-squared statistics. Numerical and simulation studies will be conducted to compare the power of the proposed test against tests that are currently used in the literature. Real data from the National Survey of Child's Health (NSCH) 2003 provided by Center for Disease Control and Prevention, Hyattsville Maryland is used to illustrate the method we developed.

email: rvogel@georgiasouthern.edu

### A NEW METHOD FOR ESTIMATING THE ODDS RATIO FROM INCOMPLETE MATCHED DATA

Kelly Miller, Medical College of Georgia  
Stephen Looney\*, Medical College of Georgia

Matched case-control studies are commonly used to evaluate the association between the exposure to a risk factor and a disease. The odds ratio is typically used to quantify this association. Difficulties in estimating the true odds ratio arise, however, when the exposure status is unknown for one individual in a pair. In the case where the exposure status is known for both individuals in all pairs, the true odds ratio is estimated as the ratio of the counts in the discordant cells of the observed two-by-two table. In the case where all data are independent, the odds ratio is estimated using the cross-product ratio from the observed table. In this paper we suggest a method for estimating the odds ratio when the sample consists of a combination of paired and unpaired observations. This method uses a weighted average of the two odds ratio calculations described above. We compare our method to existing methods via simulation.

email: slooney@mccg.edu

### LATENT VARIABLE MODEL FOR THE ANALYSIS OF BINARY DATA COLLECTED ON NUCLEAR FAMILIES

Yihao Deng\*, Indiana University Purdue University - Fort Wayne  
Roy Sabo, Virginia Commonwealth University  
N. Rao Chaganty, Old Dominion University

Many genetic and phenotypic studies are focused on binary variables within nuclear families. The within family genetic and phenotypic

# ABSTRACTS

binary data are naturally dependent and hence correlated. In this paper we present a latent variable model based on multivariate probit to analyze the familial correlations. We will discuss some theoretical properties of the model including feasible ranges of the correlations. Using a stochastic representation we simplify maximum likelihood estimation of the parameters and the Fisher information. We illustrate our methods with a real life example concerning the effects of erythrocyte adenosine triphosphate (ATP) levels among Caucasian family members.

email: dengy@ipfw.edu

## A SELF-CONSISTENT APPROACH TO MULTINOMIAL LOGIT MODEL WITH REPEATED MEASURES

Shufang Wang\*, University of Michigan  
Alex Tsodikov, University of Michigan

Modeling repeatedly observed categorical response is a challenge in terms of computation, due to the complex form of likelihood function and the presence of random effects. The computation is costly especially when the categorical response has a large number of categories, since it involves a high-dimensional integration. In this paper, we develop a stable MLE approach to the problem, based on generalized self-consistency and quasi-EM algorithm. The method transforms the complex multinomial likelihood to Poisson likelihood and hence allows for the estimates to be obtained iteratively and through a set of smaller problems. Simulation study indicates that the parameter estimates are consistent and stable.

email: sfwang@umich.edu

## A FULL EXCHANGEABLE NEGATIVE BINOMIAL LIKELIHOOD PROCEDURE FOR MODELING CORRELATED OVERDISPERSED COUNT DATA

Xiaodong Wang\*, Sanofi-aventis  
Hanxiang Peng, Indiana University Purdue University - Indianapolis

We introduce an exchangeable negative binomial by relaxing independence to exchangeability in the negative binomial. Based on this model, we propose a full likelihood procedure for investigating overdispersed and correlated binary data common in biomedical sciences and teratology. The proposed model can be characterized by a completely monotone sequence of infinitely many parameters and used to model all distributional information. We give two methods about converting from the distribution of infinitely many parameters to a parsimonious distribution of finitely many parameters, i.e., truncation and completely monotone links. We calculate the moments and perform simulation to illustrate the distribution. We also provide an estimating procedure based on maximum likelihood and the empirical estimates.

email: sheldon.wang@sanofi-aventis.com

## ANALYSIS OF MULTIVARIATE LONGITUDINAL BINARY DATA USING MARGINALIZED RANDOM EFFECTS MODELS

Keunbaik Lee\*, Louisiana State University Health Science Center  
Yongsung Joo, Dongguk University, South Korea  
Jae Keun Yoo, University of Louisville  
JungBok Lee, Korea University

Generalized linear models with random effects are often used to explain the serial dependence of longitudinal categorical data. Marginalized random effects models (MREMs) permit likelihood-based estimations of marginal mean parameters and also explain the serial dependence of longitudinal data. In this paper, we extend the MREM to accommodate multivariate longitudinal binary data using a new covariance matrix with a Kronecker decomposition which easily explains both the serial dependence and time specific response correlation. A maximum marginal likelihood estimation is proposed utilizing a Quasi-Newton algorithm with Quasi-Monte Carlo integration of the random effects. Our approach is applied to analyze metabolic syndrome data from the Korean Genomic Epidemiology Study (KGES) for Korean adults.

email: klee4@lsuhsc.edu

## ROBUST INFERENCE FOR SPARSE CLUSTERED COUNT DATA

John J. Hanfelt, Rollins School of Public Health, Emory University  
Ruosha Li\*, Rollins School of Public Health, Emory University  
Yi Pan, Rollins School of Public Health, Emory University  
Pierre Parent, Institute Armand-Frappier, Canada

Standard methods for the analysis of cluster-correlated count data fail to yield valid inferences when the study is finely stratified and the interest is in assessing the intra-cluster correlation structure. We present an approach, based upon exactly adjusting an estimating function for the bias induced by the fitting of stratum-specific effects, that requires modeling only the first two joint moments of the observations and that yields consistent and asymptotically normal estimators of the correlation parameters. The approach is motivated by a study of the health effects of the consumption of drinking water, where the outcomes were the counts of gastrointestinal illness episodes for each person within households from a population stratified by small geographic area.

email: rli2@emory.edu

## 80. INFECTIOUS DISEASES

### STATISTICAL ANALYSIS OF HIV-1 ENV SEQUENCES AND THEIR ROLE IN SELECTION PROCESS OF VIRAL VARIANTS IN MTCT

Rongheng Lin\*, University of Massachusetts Amherst  
Mohan Somasundaran, University of Massachusetts Medical School  
Michael Kishko, University of Massachusetts Medical School

Mother-to-child transmission (MTCT) accounts for 16% of new

infections of HIV-1 virus every year. Recent studies show that in MTCT some strains of virus preferentially get transmitted. Although a variety of viral, host, and obstetric factors can affect the selective transmission process, transmission and pathogenesis is potentially influenced by HIV-1 genetic variation. HIV-1 env gene encodes the gp160 protein and is believed to be associated with the selection process. A clear understanding of the diversity of the early viral quasispecies, whether selective viral variants are transmitted, and whether they change over time within individuals or populations may be obtained by the genetic characterization of viruses from mother-infant pairs. Characterization of changes in viral sequences in individuals and populations could provide important data about the sensitivity of transmitted viruses to antibody neutralization or other selective pressures. In this talk, we'll present our recent study of a data set consisting of 159 virus sequences from 5 mother-child pairs. We'll explore the association between specific sequence mutations and transmissibility from mother to child and discuss the statistical challenges presented by the data set, including partially observed transmissibility, non-perfect alignment of the sequences and high dimensionality, etc.

email: rlin@schoolph.umass.edu

## AN EPIDEMIOLOGICAL MODEL FOR GENETIC MAPPING OF VIRAL PATHOGENESIS

Yao Li\*, University of Florida  
Arthur Berg, University of Florida  
Maryon M. Chang, University of Florida  
Rongling Wu, Penn State University

Several serious human infectious diseases, such as AIDS, hepatitis B, influenza, and rabies, are the consequence of the interactions between the causative agent (virus) and host through the regulation of environmental and epidemiological factors. The identification of genes from the virus and host genomes will facilitate the understanding of the etiology of the diseases and the application of this information to develop antiviral drugs. Here, we will present a statistical model for characterizing genes and their interactions responsible for an infectious disease in a human population. The model incorporates the epidemiological mechanisms of the disease into a statistical framework for genetic haplotyping with multi-locus sequence data. In particular, the effects of genes from the transmitters (close contacts of patients) will be embedded in the model, allowing the characterization and test of genes from three different genomes (virus, host and transmitter). A testing procedure is constructed for study main genetic effects of different genomes and their epistatic interactions with different orders. We detect that high-order epistasis can be estimated, provided that an adequately large sample size (e.g., 2000 or more) is given. The new epidemiological model was investigated and validated through simulation studies.

email: yaoli@ufl.edu

## STATISTICAL MODEL FOR A DUAL-COLOR TAG SYSTEM FOR INVESTIGATING VIRUS-VIRUS INTERACTIONS

Jing Zhang, Miami University  
Douglas A. Noe\*, Miami University  
Stephen E. Wright, Miami University  
John Bailer, Miami University

A test system has been developed for staining cells to detect whether the cells have been infected by one or more viruses. A distinction between the states of viral infectivity (virus 1 infection, virus 2 infection, both virus 1 and 2 infection, no infection) leads to a multinomial framework in which cell probabilities can be defined in terms of attack rates and a possible window of time in which a cell might be infected by both viruses. This framework is developed and used to develop tests and interval estimates of model parameters in a study of a single cell line and to develop a test of virus infectivity equivalence between two cell lines. The performance of this test and estimation procedure is explored in a small simulation study.

email: noeda@muohio.edu

## POOLED NUCLEIC ACID TESTING TO IDENTIFY ANTIRETROVIRAL TREATMENT FAILURE DURING HIV INFECTION

Susanne May\*, University of Washington  
Anthony Gamst, University of California-San Diego  
Richard Haubrich, University of California-San Diego Medical Center  
Constance Benson, University of California-San Diego Medical Center  
Davey M. Smith, University of California-San Diego, School of Medicine

**Background:** Pooling strategies have been used to reduce the costs of polymerase chain reaction based screening for acute HIV infection in populations where the prevalence of acute infection is low (<1%). Only limited research has been done for conditions where the prevalence of screening positivity is higher (>1%). **Methods and Results:** We present data on a variety of pooling strategies that incorporate the use of PCR-based quantitative measures to monitor for virologic failure among HIV-infected patients receiving antiretroviral therapy. For a prevalence of virologic failure between 1% and 25%, we demonstrate relative efficiency and accuracy of various strategies. These results could be used to choose the best strategy based on the requirements of individual laboratory and clinical settings, such as required turnaround time of results, and availability of resources. **Conclusions:** Virologic monitoring during antiretroviral therapy is not currently being performed in many resource constrained settings largely because of costs. The presented pooling strategies may be used to make such monitoring feasible and to optimally limit the development and transmission of HIV drug resistance in resource constrained settings. They may also be used to design efficient pooling strategies for other settings where screening involves quantitative measures.

email: sjmay@u.washington.edu

# ABSTRACTS

## REGRESSION ANALYSIS OF CLUSTERED INTERVAL CENSORED DATA WITH INFORMATIVE CLUSTER SIZE

Yang-Jin Kim\*, Ewha Womans University, Korea

Interval censored data are often observed in the study of infection data where time to infection is not exactly observed and is only known to occur within a certain interval. Several techniques to analyze interval censored data have been suggested with independent assumption. Sometimes, such interval censored data comprised of clusters and consequent failure times do not hold independent assumption any more. Furthermore, when cluster size includes some information about the failure time, informative cluster size should have been considered in the analysis. In this study, we discuss a joint model for the analysis for clustered interval censored data with informative cluster size. To investigate possible association between failure time and cluster size, PH model and ordinal regression model are applied for failure time and cluster size, respectively. Then a bivariate random effect is adopted to connect two models. Simulation studies are performed to evaluate a finite sample properties and a lymphatic filariasis study is analyzed as an example.

email: yjinkim@ewha.ac.kr

## ESTIMATING INCUBATION PERIOD DISTRIBUTIONS WITH COARSE DATA

Nicholas Reich\*, Johns Hopkins School of Public Health

The incubation period, the time between infection and disease onset, is critical in the surveillance and control of infectious diseases but it is often coarsely observed. Coarse data arises either because the time of infection, the time of disease onset or both may not be known precisely. Methods that can efficiently estimate parameters of an incubation period distribution provide results that are useful in real-time outbreak investigations, in identifying possible sources of exposure and in modeling the efficacy of public health interventions. We compare two methods of estimating the incubation period distribution. The first method uses all available information about both exposure and disease onset, representing the data as doubly interval censored. The second introduces a data reduction technique which makes the computation considerably more tractable. In a simulation study, the methods performed similarly when estimating the median of the distribution: empirical 95% confidence interval coverages were within two percentage points of 95%. However, when estimating the tails of the distribution, the first method yielded more reliable estimates. We conducted an analysis of the sensitivity of the two methods to violations of model assumptions, in particular the distribution of infection times. We used the methods to estimate the incubation period distribution for influenza A and respiratory syncytial virus. The estimates of the medians are similar and the doubly interval censored method estimated slightly larger variability within the distributions. The analysis of reduced data is less computationally intensive than the doubly interval censored analysis and performs well for estimating the median incubation period under a wide range of conditions. However for estimation of the extreme tails of the incubation period distribution, the doubly interval censored analysis is the recommended procedure.

email: nick.reich@gmail.com

## RECURSIVE PARTITIONING FOR LONGITUDINAL MARKERS BASED ON A U-STATISTIC

Shannon Stock\*, Harvard University  
Victor DeGruttola, Harvard University  
Chengcheng Hu, University of Arizona

The development of HIV resistance mutations can reduce the efficacy of specific antiretroviral drugs used to treat HIV infection, and cross-resistance within classes of drugs is common. Recursive partitioning has been extensively used to identify resistance mutations associated with a reduced virological response measured at a single time point; here we describe a statistical method that accommodates a large set of genetic or other covariates and a longitudinal response. Our recursive partitioning approach for continuous longitudinal data uses the kernel of a U-statistic as the splitting criterion. The method is flexible and avoids the need for parametric assumptions regarding the relationship between the observed response trajectories and covariates. Under the assumption that some longitudinal measurements are missing at random, we extend our estimators to accommodate such missingness by incorporating inverse probability weights using either recurrent event processes or direct modeling of the probability of observing a response. The performance of our method is explored using simulation studies and by investigating its asymptotic properties. We illustrate this method using data combined from a variety of clinical research studies that investigated the drug Abacavir among patients for whom baseline HIV genotype was available.

email: sstock@hsph.harvard.edu

## 81. RATER AGREEMENT AND SCREENING TESTS

### EVALUATION OF INDIVIDUAL OBSERVER AGREEMENT FROM DATA WITH REPEATED MEASUREMENTS

Jingjing Gao\*, Emory University  
Michael Haber, Emory University  
Huiman Barnhart, Duke University

Coefficients of individual agreement between observers or methods of measurement are based on the comparison of the between and within-observer mean squared deviation (MSD). Methods for estimation of these coefficients from data where each observer makes replicated measurements on each subject have been developed. In this presentation we introduce a simple method for estimation of coefficients of individual agreement when data consists of matched sets of repeated measurements made under different conditions. The conditions may represent different time points, raters, laboratories, treatments, etc. Our approach allows the values of the measured variable and the magnitude of disagreement to vary across the conditions. The new approach is illustrated via two examples from studies designed to compare (a) methods of evaluating carotid stenosis and (b) methods of measuring percent body fat.

email: jgao@emory.edu



## A MISSING DATA APPROACH FOR ADJUSTING DIAGNOSES OF POST-TRAUMATIC STRESS DISORDER THAT ARE SUBJECT TO RATER BIAS

Juned Siddique\*, Northwestern University  
Bonnie L. Green, Georgetown University  
Robert D. Gibbons, University of Illinois at Chicago

Compared to other rating systems that are based on counts or frequencies of readily observable behaviors, assessments of Post-Traumatic Stress Disorder (PTSD) allow more scope for disagreement among raters due to the fact that PTSD assessments require raters to make subjective judgments about the meaning and representations of target behaviors. We describe a multiple imputation approach using a Bayesian censored ordinal probit model to adjust diagnoses of PTSD that are subject to rater bias. Adjusted diagnoses can be used as the basis for revised estimates of PTSD prevalence or as a covariate in subsequent analyses. We apply our methods to data from a depression study where nurse practitioners were twice as likely to diagnose participants with PTSD as compared to diagnoses made by clinical psychologists.

email: siddique@northwestern.edu

## MULTIVARIATE CONCORDANCE CORRELATION COEFFICIENT

Sasiprappa Hiriote\*, Eberly College of Science, Penn State University  
Vernon M. Chinchilli, Penn State College of Medicine

In many clinical studies, the concordance correlation coefficient introduced by Lin (1989) is a common tool to assess the agreement of a continuous response measured by two raters or methods. However, the need of measures of agreement may arise for more complex situations, such as when the responses are measured on more than one occasion by each rater or method. In this work, we propose a new version of the concordance correlation coefficient, called the multivariate concordance correlation coefficient, which possesses the properties needed to characterize the level of agreement between two  $p \times 1$  vectors of random variables. It reduces to Lin's concordance correlation coefficient when  $p = 1$ . The proposed estimators are proven to be asymptotically normal and their performances are evaluated via simulation studies. Real data from asthma clinical trials are used for demonstration.

email: sxh350@psu.edu

## BAYESIAN PERFORMANCE ASSESSMENT FOR RADIOLOGISTS INTERPRETING MAMMOGRAPHY

Dawn Woodard\*, Cornell University

We use Bayesian hierarchical modeling techniques to infer about interpretive performance of individual radiologists in screening mammography. Our approach accounts for differences due to patient mix and radiologist attributes (for instance, years of experience or interpretive volume). We model at the mammogram level, and then use these models to assess radiologist performance. Our approach is demonstrated with data from mammography registries and radiologist

surveys. Using these Bayesian hierarchical models we estimate several radiologist performance metrics, including the difference between the sensitivity or specificity of a particular radiologist and that of a hypothetical "standard" radiologist with the same attributes and the same patient mix.

email: woodard@orie.cornell.edu

## MODELING THE CUMULATIVE RISK OF A FALSE POSITIVE SCREENING MAMMOGRAM

Rebecca A. Hubbard\*, Group Health Center for Health Studies  
Diana L. Miglioretti, Group Health Center for Health Studies

Mammography is the only screening modality shown to reduce breast cancer mortality among women 50 and older in clinical trials. However, screening mammography is associated with adverse outcomes including false-positive recalls and benign biopsies. False-positive results lead to increased cost as well as patient anxiety. There are several existing statistical methods for estimating the cumulative risk of a false positive screening exam. Specifically, regression models have been developed for estimating the effect of fixed and random effects on the cumulative risk of a false positive screening exam. However, these models make use of possibly unrealistic assumptions about independence of the history of exam results and the number of screens obtained. Additional methods have been developed that do not rely on this assumption, but these methods do not allow for the inclusion of random effects. We will review existing methods and discuss their limitations and assumptions. We then apply these methods to a population based study of screening mammography, comparing inference on the false positive recall rate and investigating the appropriateness of modeling assumptions. Based on the performance of existing statistical methods, we will propose extensions to address limitations or unrealistic assumptions.

email: hubbard.r@ghc.org

## REEXAMINATION AND FURTHER DEVELOPMENT OF THE ROE AND METZ SIMULATION MODEL FOR MULTIPLE READER ROC DECISION DATA

Stephen L. Hillis\*, VA Iowa City Medical Center

The simulation model proposed by Roe and Metz (RM) is the most commonly used model for evaluating the performance of methods designed to analyze multireader ROC data that take into account both reader and case variability. In this talk I examine the RM model in more detail. I reformulate the model using more standard statistical notation to show that the model is actually a four-factor model with certain effects set to zero, discuss the conceptual motivation for the model, discuss extensions of the model, derive the population parameters of interest, and show the relationship between the RM variance components and the variance components for the two estimation methods proposed by Dorfman, Berbaum & Metz, and Obuchowski & Rockette.

email: steve-hillis@uiowa.edu

# ABSTRACTS

## BAYESIAN INFERENCE FOR TRUE-BENEFIT AND OVER-DIAGNOSIS IN PERIODIC CANCER SCREENING

Dongfeng Wu\*, School of Public Health, University of Louisville  
Gary L. Rosner, University of Texas, MD Anderson Cancer Center

We developed a probability model for evaluating true benefit and over-diagnosis in periodic cancer screening. Peoples who take part in the screening program are categorized into 4 mutually exclusive groups: Pure-Waste, No-Benefit, True-Benefit, and Over-Diagnosis. For each case, the probability was derived. Simulation studies using the HIP (Health Insurance Plan for Greater New Yorker) study's data provide estimates for these probabilities. Our model can provide policy makers with important information regarding the percentage of true early detection and the percentage of over diagnosis. Though the study focuses on breast cancer screening, it is also applicable to other kinds of chronic disease.

email: dongfeng.wu@louisville.edu

## 82. APPLIED DATA ANALYSIS, GRAPHICAL DISPLAYS, AND BIOSTATISTICAL LITERACY

### ANALYSIS OF VARIANCE ON A CATEGORIZED CONTINUOUS VARIABLE

Wenyaw Chan\*, School of Public Health, University of Texas-Health Science Center at Houston  
Lin-An Chen, Institute of Statistics, National Chiao Tung University-Hsin Chu, Taiwan  
Younghun Han, University of Texas- M. D. Anderson Cancer Center

It is not uncommon in epidemiological studies that observations from an independent continuous variable are categorized and the analysis of variance (ANOVA) method is applied to compare the means of the response variable on these categories. For example, among articles published in American Journal of Epidemiology from 2005 to October of 2007, 8 of them used this method. For this method, the statistical appropriateness has never been rigorously examined. We develop the model for this problem under the assumption that response variable and categorized variable follow a bivariate normal distribution. We found that the classical assumptions of normality and constant variance are violated. However, under the null hypothesis of equal means, the classical ANOVA technique is still valid. When the null hypothesis is rejected, the conditional means are monotone. The analytical results were verified by the simulation. These results may help the researchers in making inferences on relationship between the response variable and the categorized independent variable.

email: wenyaw.chan@uth.tmc.edu

### BAYESIAN CANCER TREND ANALYSIS

Pulak Ghosh, Emory University  
Kaushik Ghosh\*, University of Nevada Las Vegas  
Ram C. Tiwari, U.S. Food and Drug Administration

Annual Percentage Change (APC) summarizes trends in age-adjusted cancer rates over short time-intervals. This measure assumes linearity

of the logarithms of rates over the intervals in question, which may not be valid, especially for relatively longer time-intervals. An alternative is the Average Annual Percentage Change (AAPC), which computes a weighted average of APC values over intervals where log-rates are piecewise linear. In this article, we propose a Bayesian approach to calculating APC and AAPC values from age-adjusted cancer rate data. The procedure involves modeling the corresponding counts using age-specific Poisson regression models with a log-link function that contains unknown joinpoints. The regression slope parameters, including slopes at the joinpoints are assumed to have a normal-inverse-gamma setup and the joinpoints are assumed to be uniformly distributed subject to order-restrictions. Finally, age-specific intercept parameters are modeled nonparametrically using a Dirichlet process prior. The proposed method can be used to construct Bayesian credible intervals for AAPC using age-adjusted mortality rates. Simulation studies are used to demonstrate the success of the method in capturing trend-changes. Finally, the proposed method is illustrated using data on prostate cancer incidence.

email: kaushik.ghosh@unlv.edu

## MORTALITY MODEL FOR PROSTATE CANCER

Shih-Yuan Lee\*, University of Michigan  
Alex Tsodikov, University of Michigan

Since the introduction of prostate cancer screening using the Prostate Specific Antigen (PSA), more than thirty percent prostate cancer mortality decline was observed. We propose a statistical model to assess and predict the effect of PSA screening on prostate cancer mortality in United State. The model contains four major components. Marginal incidence model predicts age at diagnosis of prostate cancer. Stage and grade specific model predicts the probability of being diagnosed at a specific stage and grade at cancer incidence. Treatment model describes the probability of receiving a certain treatment combination at the time of cancer diagnosis. Survival model calculates the survival time from the diagnosis to death after adjusting for lead time. Treatment effect parameters were obtained from clinical trials and the model was fitted using Surveillance, Epidemiology and End Results (SEER) data. Age adjusted observed and predicted prostate cancer mortalities were compared.

email: shihylee@umich.edu

## A CLASS OF DISTRIBUTIONS WITH NORMAL SHAPE DENSITIES ON FINITE INTERVALS

Ahmad Reza Soltani\*, Kuwait University

In a series of papers van Dorp and Kotz (2002-2006) introduced classes of distributions, called two sided power distributions and their generalizations, on finite intervals to model real data with finite range, mostly economical and medical data. They were apparently interested in normal densities on finite intervals. Symmetric two sided power distributions lack this property, due to certain irregularities. Peak of densities is too high, and densities are too narrow near their peaks. In this work by using certain transforms of certain Dirichlet distributions we present new classes of symmetric distributions on finite intervals that their densities are very similar to the densities of normal distributions.

The density is formulated using certain Guass Hypergeometric functions. Potentials of these distributions in modelling real data is brought into light.

email: soltani@kuc01.kuniv.edu.kw

## SIMULATION BASED VISUALIZATION OF INFERENCE FUNCTIONS

Daeyoung Kim\*, University of Massachusetts-Amherst  
Bruce G. Lindsay, Penn State University

This paper presents a new simulation based methodology designed to facilitate visual analysis of the confidence sets generated by an inference function such as the likelihood. This methodology generates a sample from the parameter space using a fiducial-like distribution. This distribution is designed so that its probabilities on the parameter space are equal to the coverage probabilities of the targeted confidence sets. Plotting these samples provides picture of the inference function surface around the point estimator optimizing the inference function. Once the sample is created, one can picture the profile inference function confidence sets for multiple parameters of interest without further complicated optimization. We illustrate the methodology with four different inference functions. Although this methodology is related to Fisher's concept of fiducial inference, the fiducial-like distribution we create here is chosen for its ability to recover the confidence sets generated by the inference function and for its ease in computation. Unlike resampling methods such as parametric bootstrap, our method uses the original data set, just as is done in Bayesian inference. We use illustrative examples to compare simulation-based confidence sets with those based on numerical optimization, and to compare the confidence regions generated by different inference functions.

email: daeyoung@math.umass.edu

## ANIMATED GRAPHICS AND VISUAL METAPHORS HELP EXPLAIN COMPLEX MATHEMATICAL RELATIONSHIPS

John T. Brinton\*, University of Colorado Health Science Center  
Deborah H. Glueck, University of Colorado at Denver and Health Sciences Center

Statisticians often need to explain complicated mathematical relationships to scientists and medical professionals. While symbols and equations help people trained in mathematics to understand issues, they can make results unclear to people with training in other scientific disciplines. A successful presentation accurately conveys the mathematics behind the science to all listeners. We advocate abandoning equations and symbols when addressing general audiences. We suggest summarizing statistical results using simple, consistent visual metaphors, in both animated graphics, and still pictures. We demonstrate a number of graphic principles that can produce clarity in communication. Graphics from recent invited talks on bias serve as a demonstration of our ideas. Pie charts represent percent disease verification, ellipses express the magnitude of correlation between screening test scores, and vertical bar charts illustrate the rate of disease in a population. The effect of

each factor on bias is demonstrated with a movie and stills showing small graphical elements changing next to moving receiver operating characteristic curves. We provide step by step instructions, SAS and Mathematica code, and presentation quality examples.

email: john.brinton@uchsc.edu

## ASSESSING BIOSTATISTICAL LITERACY AND STATISTICAL THINKING

Felicity B. Enders\*, Mayo Clinic

One of the tenets of statistical education is that outcome assessment should be based on an independent instrument rather than course-specific outcomes. DelMas, Garfield, Ooms, and Chance (2006) have developed the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) and a smaller version called the Statistics Thinking and Reasoning Test (START; Garfield, DelMas, Chance, and Ooms; 2007), but no such assessment has been available for biostatistics courses for nonstatisticians. Windish et al (2007) developed an instrument to assess physicians' comprehension of biostatistical results typically represented in the medical literature (the Windish Quiz). In this study, the START and the Windish Quiz were used to jointly assess biostatistical literacy and statistical thinking among physicians completing an introductory course in biostatistics. The two assessments were independently associated with final exam scores (START  $p=0.004$ ; Windish  $p<0.001$ ); together, they predicted 46% of exam variability. This project suggests that biostatistical literacy varies somewhat from statistical literacy and that a different outcome instrument is needed for biostatistics courses.

email: enders.felicity@mayo.edu

**TUESDAY, MARCH 17, 2009  
3:45 - 5:30 PM**

## 83. IMS MEDALLION LECTURE

### STATISTICAL CHALLENGES IN NANOSCALE BIOPHYSICS

Samuel Kou, Ph.D.\*, Harvard University

Recent advances in nanotechnology allow scientists to follow a biological process on a single molecule basis. These advances also raise many challenging stochastic modeling and statistical inference problems. First, by zooming in on single molecules, recent nanoscale experiments reveal that many classical stochastic models derived from oversimplified assumptions are no longer valid. Second, the stochastic nature of the experimental data and the presence of latent processes significantly complicate the statistical inference. In this talk we will use the modeling of enzymatic reaction and the inference of biochemical kinetics to illustrate the statistical and probabilistic challenges in single-molecule biophysics.

email: kou@stat.harvard.edu

# ABSTRACTS

## 84. MEDIATION AND CAUSAL INFERENCE

### BAYESIAN INFERENCE FOR MEDIATION EFFECTS USING PRINCIPAL STRATIFICATION

Michael R. Elliott\*, University of Michigan  
Trivellore E. Raghunathan, University of Michigan

Most investigations in the social and health sciences aim to understand the causal relationship between a treatment or risk factor and outcome. Furthermore, given the multitude of pathways through which the treatment or risk factor may affect the outcome, there is also interest on decomposing the effect of a risk factor into "direct" and "mediated" effects. Building on the potential outcome framework for causal inference, we develop a Bayesian approach to estimate direct and mediating effects. This approach recognizes that direct and mediating effects can only be expressed as a range of plausible values of the population parameters constructed from potential populations. This range can be reduced by making further assumptions. For example, monotonicity or exclusionary restrictions used in the causal inference from randomized experiments reduces this range to a single estimable parameter. Such assumptions may not be reasonable in observational studies, or might be reasonable only in a stochastic, rather than deterministic, fashion. Here we use principal stratification (Rubin and Frangakis 2002) to draw inferences on the range of plausible values conditional on the observed data and perform sensitivity analysis using different prior distributions. The methodology is illustrated using both real and simulated examples.

email: mrelliot@umich.edu

### CONTROLLED DIRECT AND MEDIATED EFFECTS: DEFINITION, IDENTIFICATION AND BOUNDS

Tyler J. VanderWeele\*, University of Chicago

A new identification result for controlled direct effects is given for settings in which data is available for a set of variables that intercept all paths between a treatment and an outcome. Furthermore, in this setting it is possible to provide a definition not just for controlled direct effects but also for controlled indirect effects (or controlled mediated effects). Further results are given which provide bounds for controlled direct effects when the no-unmeasured-confounding assumptions required for the identification of these effects do not hold. Previous results concerning bounds for controlled direct effects rely on monotonicity relationships between the treatment and the outcome; the results presented in this paper instead assume that monotonicity relationships hold between the unmeasured confounding variable or variables and the treatment, mediator and outcome. The results on bounds for controlled direct effects are motivated by and applied to a problem concerning the effects of prenatal care on birth outcomes.

email: vanderweele@uchicago.edu

## ESTIMATING CONTROLLED DIRECT EFFECTS IN RANDOM AND OUTCOME-DEPENDENT SAMPLING DESIGNS

Stijn Vansteelandt\*, Ghent University, Belgium

Estimating what is the effect of an exposure on an outcome other than through some given mediator, requires adjustment for all risk factors of the mediator that are also associated with the outcome. When these risk factors are themselves affected by the exposure, then standard regression methods do not apply. In this presentation, I will briefly review methods for accommodating this and discuss their limitations for estimating the controlled direct effect, in particular focussing on studies with a continuous mediator. In addition, I will propose a powerful and easy-to-apply alternative which uses G-estimation in structural nested models to address these limitations both for cohort and outcome-dependent sampling designs.

email: stijn.vansteelandt@ugent.be

## 85. CHALLENGES IN THE BAYESIAN SPATIO-TEMPORAL ANALYSIS OF LARGE AND HETEROGENEOUS DATASETS

### SPATIAL MISALIGNMENT IN TIME SERIES STUDIES OF AIR POLLUTION AND HEALTH DATA

Roger D. Peng\*, Johns Hopkins Bloomberg School of Public Health  
Michelle L. Bell, Yale University

Time series studies of environmental exposures often involve comparing daily changes in a toxicant measured at a point in space with daily changes in an aggregate measure of health. Spatial misalignment of the exposure and response variables can bias the estimation of health risk and the magnitude of this bias depends on the spatial variation of the exposure of interest. In air pollution epidemiology, there is an increasing focus on estimating the health effects of the chemical components of particulate matter. One issue that is raised by this new focus is the spatial misalignment error introduced by the lack of spatial homogeneity in many of the particulate matter components. Current approaches to estimating short-term health risks via time series modeling do not take into account the spatial properties of the chemical components and therefore could result in biased estimation of those risks. We present a spatial-temporal statistical model for quantifying spatial misalignment error and show how adjusted health risk estimates can be obtained using a regression calibration approach and a two-stage Bayesian model. We apply our methods to a database containing information on hospital admissions, air pollution, and weather for 20 large urban counties in the United States.

email: rpeng@jhsph.edu

## **BAYESIAN VARIABLE SELECTION FOR MULTIVARIATE SPATIALLY-VARYING COEFFICIENT REGRESSION: APPLICATION TO PHYSICAL ACTIVITY DURING PREGNANCY**

Montserrat Fuentes\*, North Carolina State University  
Brian Reich, North Carolina State University  
Amy Herring, University of North Carolina-Chapel Hill

For pregnant women, the American College of Obstetricians and Gynecologists currently recommends 30 minutes of moderate exercise on most days. Epidemiologists, policy makers, and city planners are interested in whether characteristics of the physical environment in which women live and work have influence on physical activity levels during pregnancy. We study the associations between physical activity and several factors including personal characteristics, meteorological/air quality variables, and neighborhood characteristics in pregnant women in four counties of North Carolina. We simultaneously analyze six types of physical activity and investigate cross-dependencies between these activity types. Exploratory analysis suggests that the associations are different in different regions. Therefore we use a multivariate regression model with spatially-varying regression coefficients. This model includes a regression parameter for each covariate at each spatial location. For our data with many predictors, some form of dimension reduction is clearly needed. We introduce a spatial Bayesian variable selection procedure to identify subsets of important variables. Our stochastic search algorithm determines the probabilities that each covariate's effect is null, non-null but constant across space, and spatially-varying.

email: fuentes@stat.ncsu.edu

## **HIERARCHICAL SPATIAL MODELING OF GENETIC VARIANCE FOR LARGE SPATIAL TRIAL DATASETS**

Sudipto Banerjee\*, University of Minnesota  
Andrew O. Finley, Michigan State University  
Patrik Waldmann, Swedish University of Agricultural Sciences-Sweden  
Tore Ericsson, Swedish University of Agricultural Sciences-Sweden

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. Recently much attention has been devoted to this problem. In this talk we primarily focus upon the use of a predictive process derived from the original spatial process that projects process realizations to a lower-dimensional subspace thereby reducing the computational burden. This approach can be looked upon as a process-based approach to reduced-rank methods for 'kriging' but offers additional complexities. We discuss attractive theoretical properties of this predictive process as well as its greater modeling flexibility compared to existing methods. We also discuss some pitfalls of this and other reduced-rank methods and offer remedies. A computationally feasible template that encompasses these diverse settings will be presented and illustrated.

email: sudiptob@biostat.umn.edu

## **86. ADVANCES IN META-ANALYSIS**

### **GENERALIZING DATA FROM RANDOMIZED TRIALS**

Eloise Kaizar\*, The Ohio State University

Randomized controlled trials are often thought of as the gold standard of evidence in medicine because they offer strong internal validity. However, subject recruitment may introduce selection bias that limits trials' external validity. We describe and illustrate an approach that relies on the often smaller selection bias seen in observational data to make what we call generalizability judgments for trials. Further, we examine a simple framework for combining randomized and observational data to model and adjust for the selection bias of a randomized study. We consider the approximations this model requires for the structure of the data under a simple linear model, and explore its feasibility for real applications.

email: ekaizar@stat.osu.edu

### **MULTIVARIATE META-ANALYSIS: MODELLING CORRELATION STRUCTURES**

Robert W. Platt\*, McGill University  
Khajak Ishak, United BioSource Corporation

Meta-analyses of randomized trials with multiple outcomes typically treat the outcomes as independent and analyze them separately. Several authors have proposed frequentist and Bayesian multivariate models for multiple outcomes, designed to perform joint analysis to account for correlation between outcomes. We review assumptions implicit in these models, in particular related to the type and source of apparent correlation between outcomes. We examine whether the correlations measured from aggregated meta-analytic data reflect the relationships of interest, or whether these are distorted by other factors like correlated errors within studies that might induce similarity between observed outcomes. The focus of our analysis was not the estimation of correlations, but rather the dependencies reflected in the true correlations underlying each study. We simulated studies of the effect of a treatment on two dichotomous endpoints, incorporating various sources of correlation. We found considerable overlap in correlations observed across different scenarios in which we varied the strengths of the associations of interest; thus, the true relationship between treatment effects could not always be inferred accurately.

email: robert.platt@mcgill.ca

### **NON-PARAMETRIC ROC CURVE META-ANALYSIS WITH VARYING NUMBER OF THRESHOLDS**

Vanja M. Dukic\*, University of Chicago

Standard meta-analytic methods combine information on only one parameter, such as a simple treatment effect. For meta-analysis of diagnostic test accuracy, measures of both sensitivity and specificity from different trials are of meta-analytic interest, summarized as a bivariate measure of accuracy, or possibly as a receiver operating characteristic (ROC) curve. Motivated by an analysis of serum

# ABSTRACTS

progesterone tests for diagnosing non-viable pregnancy, we develop simple fixed-effects and random-effects summary ROC estimators, based on a flexible density estimation technique. We contrast the performance and risk of the new estimator to the simpler bivariate normal summary ROC estimator in a series of simulations.

email: vanja@uchicago.edu

## 87. STATISTICAL METHODS FOR GENOME-WIDE ASSOCIATION STUDIES

### ESTIMATING GENETIC EFFECTS AND GENE-ENVIRONMENT INTERACTIONS WITH MISSING DATA

Danyu Lin\*, University of North Carolina

Missing data arise in genetic association studies when genotypes are unknown or when haplotypes are of direct interest. We provide a general likelihood-based framework for estimating genetic effects and gene-environment interactions with such missing data. We allow genetic and environmental variables to be correlated while leaving the distribution of environmental variables completely unspecified. We consider three major study designs --- cross-sectional, case-control, and cohort designs --- and construct appropriate likelihood functions for all common phenotypes (e.g., case-control status, quantitative traits, and potentially censored ages at onset of disease). The likelihood functions involve both finite- and infinite-dimensional parameters. The maximum likelihood estimators are shown to be consistent, asymptotically normal, and asymptotically efficient. Fast and stable numerical algorithms are developed to implement the corresponding inference procedures. Extensive simulation studies demonstrate that the proposed inferential and numerical methods perform well in practical settings. Applications to two genomewide association studies are provided.

email: lin@bios.unc.edu

### DETECTING GENE-GENE INTERACTIONS USING GENOME-WIDE ASSOCIATION STUDIES (GWAS) IN THE PRESENCE OF POPULATION STRATIFICATION

Nilanjan Chatterjee\*, National Cancer Institute  
Samsiddhi Bhattacharjee, National Cancer Institute

Some existing approaches to measure gene-gene interactions, such as the case-only approach, make a crucial assumption of gene-gene independence in the underlying population for physically distant genes. While this strategy is known to increase power considerably, we demonstrate empirically in GWAS setting that the presence of population stratification causes large scale violation of the independence assumption and hence severe inflation of type-I error in the case-only and related methods. To solve this problem, we use the idea of “genetic matching” of cases and controls based on principal component analyses of null genetic markers. We proposed to analyze matched sets using extensions of conditional logistic regression that can derive additional power from a “weak” gene-gene independence assumption that is required to hold only within genetically

homogeneous matched sets. We compare our approach to some of the existing methods in terms of bias and efficiency, both using real GWAS data and simulations under varying degrees of stratification.

email: bhattacharjees@mail.nih.gov

## SCORE STATISTICS FOR FAMILY-BASED GENETIC ASSOCIATION STUDIES OF QUANTITATIVE TRAITS

Samsiddhi Bhattacharjee\*, National Cancer Institute  
Eleanor Feingold, University of Pittsburgh

Family-based tests of genetic association protect from spurious associations by ignoring and/or conditioning on certain data for the pedigree founders. Hence, these tests are often considerably less powerful than population-based tests that use all of the available information but are sensitive to stratification. We present a unified likelihood for quantitative traits in families and derive several score statistics for testing association. Our statistics make varying assumptions on the nature of substructure that may be present in the dataset, and provide a range of options between purely population-based and traditional family-based tests. Under certain assumptions about the stratification, we are able to incorporate founder phenotypes and derive significant additional power from the founder genotype-phenotype correlation and the environmental correlation between founders and non-founders. We extend these score tests to handle known linkage, and also derive formulas for conditional moments required to compute them in general pedigrees. Finally, we use simulations to compare the performance of these statistics to the standard approaches under varying extents of stratification.

email: bhattacharjees@mail.nih.gov

## PREDICTIVE MODELS FOR GENOME-WIDE ASSOCIATION STUDIES

Charles Kooperberg\*, Fred Hutchinson Cancer Research Center  
Michael LeBlanc, Fred Hutchinson Cancer Research Center  
Valerie Obenchain, Fred Hutchinson Cancer Research Center

One of the objectives of genome-wide association studies (GWAS) is often to attempt to refine ‘risk prediction models’ based on traditional epidemiological data, using SNPs that are associated with disease. Methods such as the lasso, lars, and the elastic net have been proposed in recent years to construct risk prediction models. While these methods can deal with many predictors, dealing with the 100,000s of predictors of a GWAS is still often beyond the capacity of these methods. Some multi-stage strategies, which may influence the selection of smoothing parameters, are needed. An additional complication is that in GWAS often some interest is in gene x gene and gene x environment interactions. The components of these interactions often are multi-dimensional, and require additional tools to be combined with individual SNPs in risk prediction models.

email: clk@fhcrc.org

## **Population Stratification Evaluation and Adjustment in Genome Wide Association Studies**

Kai Yu\*, Qizhai Li  
National Cancer Institute

Population stratification (PS) can lead to an inflated rate of false positive findings in genome-wide association studies (GWAS). The commonly used approach of adjustment for a fixed number of principal components (PCs) could have a deleterious impact on power when selected PCs are equally distributed in cases and controls, or the adjustment of certain covariates, such as self-identified ethnicity or recruitment center, already included in the association analyses, correctly maps to major axes of genetic heterogeneity. We propose a computationally efficient procedure, PC-Finder, to identify a minimal set of PCs while permitting an effective correction for PS. A general pseudo F statistic, derived from a non-parametric multivariate regression model, can be used to assess whether PS exists or has been adequately corrected by a set of selected PCs. Empirical data from two GWAS conducted as part of the Cancer Genetic Markers of Susceptibility (CGEMS) project demonstrates the application of the procedure. Furthermore, simulation studies show the power advantage of the proposed procedure in GWAS over currently used PS correction strategies, particularly when the PCs with substantial genetic variation are distributed similarly in cases and controls and therefore do not induce PS.

email: yukf@mail.nih.gov

## **88. INFERENCE FOR CLINICAL TRIALS**

### **A COMPARATIVE SIMULATION STUDY BETWEEN ETRANK® METHODOLOGY AND CONSTRAINT LONGITUDINAL DATA ANALYSIS (cLDA)**

Lian Liu\*, Merck & Co., Inc.  
Richard Entsuah, Merck & Co., Inc.

ETRANK® method uses a unique nonparametric (randomization) technique to address treatment related withdrawals (informative censoring) in longitudinal clinical trials. ETRANK® uses data-dependent scoring schemes and a unified test statistic which incorporates all available data, and adjusts for withdrawal patterns, proportion of withdrawals, and level of response prior to withdrawal. An empirical significance level for each scoring system under both the Fulldata and Endpoint methods for a specified parameter configuration is computed using this methodology. A simulation study under clinical trial missing data mechanisms is performed to compare the power and type I error of the ETRANK® methodology with the constraint longitudinal data analysis (cLDA) method which uses restriction of same baseline mean across treatment groups.

e-mail: lian\_liu@merck.com

## **RANDOMIZATION TESTS OF MULTI-ARM RANDOMIZED CLINICAL TRIALS**

Youngsook Jeon\*, University of Virginia  
Feifang Hu, University of Virginia

Randomization test is a standard inference method for two-arm randomized clinical trials. However, how to conduct randomization tests for multi-arm clinical trials is unclear. It is because treatments other than those of interest are also involved in a random rearrangement process in the standard randomization procedure for pairwise comparisons. In this talk, we deal with this multiple-treatment issue of randomization tests. We propose new randomization testing method by which true difference in the pair of treatments can be assessed without other treatments interference. The proposed method is theoretically well-founded and can be easily implemented in the practice due to its computational feasibility. Some numerical studies are also presented.

e-mail: yj6k@virginia.edu

## **COMPARISON OF VARIATIONS OF THE DUFFY-SANTNER CONFIDENCE INTERVALS FOR THE ONE-SAMPLE PROPORTION BASED ON MULTISTAGE DESIGNS**

Haihong Li\*, Vertex Pharmaceuticals

There exist several methods to calculate the exact confidence intervals for the one-sample proportion from multistage designs. One of them, which extends Crow's method from single stage to multistage design, was proposed by Duffy and Santner. We compared several variations of this method. It turned out that the 'keeping left' method has favorable properties compared with the original DS method. Four Fleming 3-stage plans and a Simon 2-stage plan were used for comparison. In most cases, the 'keeping left' method resulted in CIs with smaller total lengths and expected lengths, and coverage probabilities closer to the nominal level.

e-mail: haihong\_li@vrtx.com

## **WEIGHTED KAPLAN-MEIER ESTIMATOR FOR TWO-STAGE TREATMENT REGIMES**

Sachiko Miyahara\*, University of Pittsburgh  
Abdus S. Wahed, University of Pittsburgh

In two stage randomization designs, patients are randomized to one of the available initial treatments, and at the end of the first stage, they are randomized to one of the second stage treatments depending on the outcome of the initial treatment. Statistical inference for survival data from these designs uses methods such as marginal mean models and weighted risk set estimates. In this article, we propose a weighted Kaplan-Meier (WKM) estimator based on the method of inverse-probability weighting and compare its properties to that of the standard Kaplan-Meier (SKM) estimator, marginal mean model based (MM) estimator and weighted risk set (WRS) estimator. Simulation study reveals that the WKM estimator is asymptotically unbiased, and provides coverage rates similar to the MM and WRS estimators. The

# ABSTRACTS

SKM estimator, however, is biased when the second randomization rates are not equal between the responders and non-responders to initial treatment. The methods described are demonstrated by applying to a leukemia dataset.

e-mail: sam976@pitt.edu

## BORROWING STRENGTH WITH NON-EXCHANGEABLE PRIORS OVER SUBPOPULATIONS

Peter Mueller, M.D. Anderson Cancer Center  
Benjamin N. Bekeley, M.D. Anderson Cancer Center  
Luis G. Leon Novelo\*, Rice University  
Kyle Wathen, M.D. Anderson Cancer Center  
Fernando A. Quintana, Pontificia Universidad Católica de Chile

We introduce a non-parametric Bayesian model for success rates in a phase II clinical trial with patients presenting different subtypes of the disease under study. The subtypes are not a priori exchangeable. The lack of a priori exchangeability hinders straightforward use of traditional hierarchical models to implement borrowing of strength across disease subtypes. We propose instead a random partition model for the set of disease subtypes. All subtypes within the same cluster share a common success probability. Our model is a variation of the product partition model with a non-exchangeable prior structure. In particular the data arises from a clinical trial of patients with sarcoma, a rare cancer affecting connective and soft tissues (e.g., cartilage and fat). Each patient presents one subtype of the disease and subtypes are grouped by good, intermediate and poor prognosis. The prior model respects the varying prognosis across disease subtypes. Two subtypes with equal prognosis are more likely a priori to co-cluster than two subtypes with different prognosis. The practical motivation for this approach is that the number of accrued patients within each subtype is too small to assess the success rates with the desired precision if analyzing the data for each subtype separately.

e-mail: lgl1@rice.edu

## INFERENCE FOR NONREGULAR PARAMETERS IN OPTIMAL DYNAMIC TREATMENT REGIMES

Bibhas Chakraborty\*, University of Michigan  
Susan Murphy, University of Michigan

A dynamic treatment regime is a set of decision rules, one per stage, that takes a patient's treatment and covariate history as input, and outputs a recommended treatment. In the estimation of the optimal dynamic treatment regime from longitudinal data, the treatment effect parameters at any stage prior to the last can be nonregular under certain distributions of the data. This results in biased estimates and invalid confidence intervals for the treatment effect parameters. In this paper, we discuss the problem of nonregularity, and present an estimation method that addresses the problem. We also provide a simulation study to compare our proposed estimator with the original estimator under a variety of nonregular scenarios.

e-mail: bibhas@umich.edu

## REVERSE REGRESSION IN RANDOMIZED CLINICAL TRIALS

Zhiwei Zhang\*, U.S. Food and Drug Administration

In clinical trials, treatment comparisons are often cast in a regression framework that evaluates the dependence of the relevant clinical outcome(s) on treatment assignment and possibly other baseline characteristics. This presentation introduces a reverse regression approach to randomized clinical trials, with focus on the dependence of treatment assignment on the clinical outcome(s) of interest. A reverse regression model is essentially a semiparametric density ratio model for the outcome distributions in the two treatment groups. The resulting inferences can be expected to be more robust than those based on fully parametric models for the outcome distributions and more efficient than nonparametric inferences. In the presence of multiple endpoints, the reverse regression approach leads to a novel procedure for multiplicity adjustment that is readily available in standard logistic regression routines.

e-mail: zhiwei.zhang@fda.hhs.gov

## 89. STATISTICAL GENETICS

### HAPLOTYPE-BASED REGRESSION ANALYSIS AND INFERENCE OF CASE-CONTROL STUDIES WITH UNPHASED GENOTYPES AND MEASUREMENT ERRORS IN ENVIRONMENTAL EXPOSURES

Iryna V. Lobach\*, New York University School of Medicine  
Raymond J. Carroll, Texas A&M University  
Christine Spinka, University of Missouri  
Mitchell Gail, Division of Cancer Epidemiology and Genetics,  
National Cancer Institute  
Nilanjan Chatterjee, Division of Cancer Epidemiology and Genetics,  
National Cancer Institute

It is widely believed that risks of many complex diseases are determined by genetic susceptibilities, environmental exposures, and their interaction. Chatterjee and Carroll (2005, Biometrika92, 399-418) developed an efficient retrospective maximum-likelihood method for analysis of case-control studies that exploits an assumption of gene-environment independence and leaves the distribution of the environmental covariates to be completely nonparametric. Spinka, Carroll, and Chatterjee (2005, Genetic Epidemiology29, 108-127) extended this approach to studies where certain types of genetic information, such as haplotype phases, may be missing on some subjects. We further extend this approach to situations when some of the environmental exposures are measured with error. Using a polychotomous logistic regression model, we allow disease status to have  $K+1$  levels. We propose use of a pseudolikelihood and a related EM algorithm for parameter estimation. We prove consistency and derive the resulting asymptotic covariance matrix of parameter estimates when the variance of the measurement error is known and when it is estimated using replications. Inferences with measurement error corrections are complicated by the fact that the Wald test often behaves poorly in the presence of large amounts of measurement error. The likelihood-ratio (LR) techniques are known to be a good alternative. However, the LR tests are not technically correct in this setting because the likelihood function is based on an incorrect model,

i.e., a prospective model in a retrospective sampling scheme. We corrected standard asymptotic results to account for the fact that the LR test is based on a likelihood-type function. The performance of the proposed method is illustrated using simulation studies emphasizing the case when genetic information is in the form of haplotypes and missing data arises from haplotype-phase ambiguity. An application of our method is illustrated using a population-based case-control study of the association between calcium intake and the risk of colorectal adenoma.

e-mail: iryna.lobach@nyumc.org

## MODELING HAPLOTYPE-HAPLOTYPE INTERACTIONS IN CASE-CONTROL GENETIC ASSOCIATION STUDIES

Li Zhang\*, Cleveland Clinic Foundation  
Rongling Wu, Penn State University

Haplotype analysis has been increasingly used to study the genetic basis of human diseases, but models for characterizing genetic interactions between haplotypes from different chromosomal regions have not well been developed in the current literature. In this talk, we will present a statistical model for testing haplotype-haplotype interactions for human diseases with a case-control genetic association design. The model is formulated on a contingency table in which cases and controls are typed for the same set of molecular markers. We derive the EM algorithm to estimate and test differences in the pattern of genetic variation between cases and controls. Traditional quantitative genetic principles are integrated into the model to characterize epistatic interactions of haplotypes from different chromosomal regions. The model allows the partition of epistasis into different components due to additive x additive, additive x dominant, dominant x additive, and dominant x dominant interactions. A testing procedure is framed to test the roles of each of these components in the pathogenesis of human diseases. Simulation studies and real examples are used to validate the usefulness and utilization of the model.

e-mail: zhangl3@ccf.org

## NUCLEOTIDE MAPPING COMPLEX DISEASE AND THE LIMITING DISTRIBUTION OF THE LIKELIHOOD RATIO TEST

Yuehua Cui\*, Michigan State University  
Dong-Yun Kim, Virginia Tech

Detecting the pattern and distribution of disease variants across the genome is essential in understanding the etiology of complex disease in human. Statistical methods based on single nucleotide polymorphism (SNP) or haplotype analysis have been developed. Different from these methods, we propose a nucleotide-based mapping approach which can estimate and test the effect of a risk haplotype on a disease risk. The model is developed under the mixture model framework which presents challenges in assessing statistical significance when using the traditional likelihood ratio test (LRT). The widely used permutation tests can be applied to assess the statistical significance, but is computationally intensive. Here we study the

limiting distribution of the LRT under the proposed framework and show that the distribution of the LRT asymptotically follows a chi-square distribution under the null of no association. Simulation studies show that the asymptotic chi-square distribution performs well in finite samples with disease trait following an exponential family distribution.

e-mail: cuiy@msu.edu

## MODELING SNP GENOTYPE DATA WITH INFORMATIVE MISSINGNESS IN SAMPLES OF UNRELATED INDIVIDUALS

Nianjun Liu\*, University of Alabama at Birmingham

Even with the advancement of modern technology, data with missing genotypes are still common in genetic studies. Although some statistical methods can handle missing data, they usually assume that genotypes are missing at random either explicitly or implicitly, that is, at a given marker, different genotypes and different alleles are missing with the same probability. This assumption is over-simplified and may not hold in practice. In this study, we demonstrate that the violation of this assumption may lead to serious bias in allele frequency estimates, and association analysis based on this assumption can be biased. To address this limitation in the current methods, we propose a novel missing data model which can estimate allele frequency and missing rate without assumption about the missing data distribution. Analytically, we prove that the allele frequency and missing probability are identifiable under our model. Empirically, simulation studies illustrate that our proposed model can reduce the bias for allele frequency estimates and association analysis due to incorrect assumption on the missing data mechanism. In addition, we evaluate the impact of departure from Hardy-Weinberg equilibrium on the model. Lastly, we illustrate the utilities of our method through its application to HapMap data and another real data.

e-mail: nliu@uab.edu

## CONTRIBUTION OF GENETIC EFFECTS TO GENETIC VARIANCE COMPONENTS WITH EPISTASIS AND LINKAGE DISEQUILIBRIUM

Tao Wang\*, Department of Population Health, Medical College of Wisconsin

Zhao-Bang Zeng, Bioinformatics Research Center, North Carolina State University

Genetic models provide a basis to analyze genetic properties in a study population. For quantitative traits, a popular model that has been widely used in genetic association studies is referred to as the F\_infinity model whose parameters are often defined as the additive, dominance and epistasis effects. Another type of models, which has long been used in experimental designed populations for analysis of quantitative trait loci (QTL), are the so-called Fisherian or Cockerham models. The Cockerham model focuses on partition of genotypic variance into additive, dominance and epistatic variances, and its parameters are called the average additive, dominance or epistasis effects. Over years, there has been some confusion about the definition and interpretation of additive, dominance and epistatic effects of

# ABSTRACTS

QTL, and their relationship to the additive, dominance and epistatic variances. In this study, we explore differences and links between the F\_infinity and Cockerham models. We discuss ways of establishing the relationship between the average effects and genetic effects parameters. Some practical issues related to using of reduced models instead of full-parameterized models are also addressed.

e-mail: taowang@mcw.edu

## ASSOCIATION STUDY OF G PROTEIN-COUPLED RECEPTOR KINASE 4 GENE VARIANTS WITH ESSENTIAL HYPERTENSION IN NORTHERN HAN CHINESE

Yaping Wang\*, Emory University

Biao Li, Institute of Biophysics, Chinese Academy of Sciences-Beijing, China

Weiyan Zhao, Cardiovascular Institute, Fu Wai Hospital  
Pei Liu, The School of Public Health Southeast University-Nanjing, China

Qi Zhao, Tulane University

Shufeng Chen, Cardiovascular Institute and Fu Wai Hospital

Hongfan Li, Cardiovascular Institute and Fu Wai Hospital

Dongfeng Gu, Cardiovascular Institute and Fu Wai Hospital

To investigate the association between polymorphisms in the G protein-coupled receptor kinase 4 gene (GRK4) (R65L, A142V and A486V) and essential hypertension in northern Han Chinese, we conducted a case-control study consisting of 503 individuals with essential hypertension (HT) and 490 age-, gender-, and area-matched normotensive (NT) controls. The three GRK4 variants were genotyped by PCR-RFLP analysis. Both haplotype and single locus analysis were used to process the genotyping data. The A486 allele showed a significant association with HT ( $P < 0.001$ ). A total of 6 haplotypes were observed in the entire population, with the haplotypes L-V-A and R-A-A being found to be significantly related to hypertension ( $P = 0.001$ ).

e-mail: ywang42@emory.edu

## COMPOSITE LIKELIHOOD: ISSUES IN EFFICIENCY

Jianping Sun\*, Penn State University

Maximum likelihood is a popular statistical method largely because it provides estimators with optimal statistical efficiency. However, many realistic statistical models are so complex in structure that it becomes computationally infeasible to find the MLE, especially in large data sets. One approach to solving this problem is the method of composite likelihood, which can reduce the complexity of computation at the price of some loss of efficiency. Because of its promising features, composite likelihood has recently become more and more popular in many fields such as longitudinal data, survival analysis, time series, spatial data and genetic data. A composite likelihood is constructed by taking a product of likelihood terms, each one of which is a likelihood, conditional or marginal, for some subset of the data. The statistical efficiency of such a composite likelihood then depends on the how it was constructed. In this talk, we will introduce the composite likelihood approach and then compare several methods

for constructing them from an optimal efficiency point of view. To illustrate the method I will consider its use in a recombination model for DNA sequence data.

e-mail: jxs1021@psu.edu

## 90. HEALTH SERVICES RESEARCH

### STATISTICAL METHODS IN HEALTHCARE QUALITY IMPROVEMENT

Claudia Pedroza\*, University of Texas School of Public Health

In this talk, we review the methods that are currently being used to evaluate quality improvement in health care. In particular, we discuss the use of SPC control charts and other methods originally derived for industrial quality control. We discuss the advantages and limitations of these methods and propose that a Bayesian approach is ideally suited for QI.

e-mail: claudia.pedroza@uth.tmc.edu

### STRUCTURAL EQUATION MODELING FOR QUALITY OF LIFE DATA IN CARDIOVASCULAR DISEASE

Zugui Zhang\*, Christiana Care Health System

Paul Kolm, Christiana Care Health System

William S. Weintraub, Christiana Care Health System

Quality of life data frequently arise in medical studies when the comprehensive measurements of medical cares of subjects are investigated. For instance, in cardiovascular study, patient-reported health status data were collected. However, the widely used traditional statistical analysis involves mainly descriptive or basic trends, ignoring the characteristics of unobserved factors, which may result in misleading conclusions. The purpose of this study is to apply Structural Equation Modeling (SEM), a useful and effective tool for evaluating theories involving health related quality of life and other related patient outcomes, to evaluate the construct validity of the outcomes through the specification and testing of hypotheses linking different quality of life measures. Patients and data were from COURAGE trial, comparing a strategy of percutaneous coronary intervention plus optimal medical therapy to optimal medical therapy alone. General health status, measured using the RAND-36 with 7 domains, and Health status related to angina, evaluated via the Seattle Angina Questionnaire with 5 domains, were assessed directly from patients at baseline and at 1, 3, 6 and 12 months followed by annual evaluations for two more years. Results highlight the evaluation of relationships among latent variables derived from quality of life data via SEM.

e-mail: zzhang@ChristianaCare.org



## BAYESIAN HIERARCHICAL MODELS FOR EXTRACTING USEFUL INFORMATION FROM MEDICATION ERROR REPORTS

Jessica A. Myers\*, Bloomberg School of Public Health, Johns Hopkins University

Francesca Dominici, Bloomberg School of Public Health, Johns Hopkins University

Laura Morlock, Bloomberg School of Public Health, Johns Hopkins University

Medical errors originating in healthcare facilities are a significant source of unnecessary morbidity, mortality, and healthcare costs. Voluntary error report systems that collect information on the causes and contributing factors of medical errors and the resulting harm may be useful for developing effective harm prevention strategies. Some patient safety experts question the utility of data from errors that did not result in harm, also called near-misses. We use data from a large voluntary reporting system of 836,174 medication errors from 1999 to 2005 to provide evidence for the causal continuum hypothesis, which states that the causes and contributing factors of errors that result in patient harm are similar to the causes and contributing factors of errors that do not result in patient harm. Bayesian hierarchical models are developed for estimating the log odds of each cause (or contributing factor) given harm and the log odds of each cause given that harm did not occur. The posterior distribution of the correlation between these two vectors of log odds is used as a measure of the evidence for the hypothesis. In addition, we identify the causes and contributing factors that most likely deviate from the hypothesis, and therefore have the highest or lowest odds of harm.

e-mail: jamyers@jhsph.edu

## ANALYSIS OF DURATION TIMES WITH UNOBSERVED HETEROGENEITY THROUGH FINITE MIXTURES

Xiaoqin Tang\*, Michigan State University  
Hwan Chung, Michigan State University  
Joseph Gardiner, Michigan State University

In health services research, hospital length of stay (LOS) is often used as a proxy for health care utilization. Sometimes LOS distributions exhibit heterogeneity that cannot be adequately fitted through standard parametric models (eg, Log normal, Gamma) or explained by observed patient variables. Similar concerns are also present with time to event data. We use a latent Markov model to provide a set of principles for systematic identification of homogeneous stages of hospital duration of stay or failure time and describe their stage-sequential process. Using methods from latent class analysis we make inference on the number of latent classes and their membership probabilities. A Bayesian estimation method via Markov chain Monte Carlo (MCMC) is developed as an alternative to traditional maximum-likelihood (ML) based methods. We illustrate the use of our strategy and compare it to ML in the context of a simulation study. We also demonstrate its application with survival times of patients who underwent heart transplantation. The strategy works well and can be implemented in SAS software.

e-mail: tang@stt.msu.edu

## A STUDY ON CONFIDENCE INTERVALS FOR INCREMENTAL COST-EFFECTIVENESS

Hongkun Wang\*, University of Virginia

Hongwei Zhao, Texas A&M Health Science Center

In health policy and economics studies, the incremental cost-effectiveness ratio (iCER) has long been used to compare the economic consequences relative to the health benefits of therapies. Due to the skewed distributions of the costs and ICERs, much research has been done on how to obtain confidence intervals of ICERs, using either parametric or nonparametric methods, with or without the presence of censoring. In this talk, we will examine and compare the finite sample performance of many approaches via simulation studies. For the special situation when the health effect of the treatment is not statistically significant, we will propose a new bootstrapping approach to improve upon the bootstrap percentile method that is currently available. The most efficient way of constructing confidence intervals will be identified and extended to the censored data case. Finally, a data example from a cardiovascular clinical trial is used to demonstrate the application of these methods.

e-mail: hw3r@virginia.edu

## JOINT MODELING OF ZERO-INFLATED DATA USING COPULAS

Joanne K. Dagg\*, Purdue University

Bruce A. Craig, University of Wisconsin-Madison

Zero-inflated data arise in many real-world problems. In the healthcare arena, zeroes typically arise for two reasons. If we consider annual in-patient hospital costs, the value could be a zero because the subject never entered the hospital that year or the value could be a zero because the subject did not incur any costs when in the hospital (i.e., covered by insurance). In this talk, we will discuss the joint modeling of correlated healthcare costs. Because of the large proportion of zeroes, one cannot use the multivariate normal even after an appropriate transformation. We propose using zero-inflated or two-part models to marginally describe each cost and using a copula to correlate the variables. We compare this approach to other alternatives such as a two-part Gamma with random effects. A simulation study and real data analysis are presented.

e-mail: joannefyffe@hotmail.com

## ESTIMATING A VOLUME-OUTCOME ASSOCIATION FROM AGGREGATE LONGITUDINAL DATA

Benjamin French\*, University of Pennsylvania  
Farhood Farjah, University of Washington

David R. Flum, University of Washington

Patrick J. Heagerty, University of Washington

Recently there has been much interest in using volume-outcome data to establish causal associations between measures of surgical experience and patient outcomes following a surgical procedure. However, there does not appear to be a standard approach to a volume-outcome analysis with respect to specifying a volume measure and selecting an

# ABSTRACTS

estimation method. We establish the recurrent marked point process as a general framework from which to approach a longitudinal volume-outcome analysis and examine the statistical issues associated with using longitudinal data analysis methods to model aggregate volume-outcome data. We conclude with the recommendation that analysis carefully specify a volume measure that most accurately reflects their scientific question of interest and select an estimation method that is appropriate for their scientific context.

e-mail: bcfrench@upenn.edu

## 91. EXPERIMENTAL DESIGN

### COUNTERINTUITIVE RESULTS WHEN CALCULATING SAMPLE SIZE IN ANOVA

Yolanda Munoz Maldonado\*, Michigan Technological University

When designing experiments for testing differences between the levels of two or more factors, it is usually advised to be parsimonious with the number of selected treatments. It is assumed that this convention will help reduce the sample size required for a given significance level and power. We will present a case encountered during a consulting project that contradicts this assumption. The choice of treatments and the magnitude of the size effect play a key role in this outcome. We will also comment on the consequences of this counterintuitive result in our daily statistical practice.

e-mail: ymunoz@mtu.edu

### BAYESIAN EXPERIMENTAL DESIGN FOR STABILITY STUDIES

Harry Yang\*, MedImmune  
Lanju Zhang, MedImmune

Recently methods for the selection of stability designs have been proposed. However there are few publications that directly deal with the optimal designs for shelf life estimation. This paper addresses such optimal designs from a Bayesian perspective, taking advantage of historical data concerning drug product stability. Criteria based on Bayesian decision theory are developed to optimize designs for shelf life estimation. The method is developed assuming the stability profile can be characterized by a linear model. However, it can be readily generalized to the cases of generalized linear and non-linear models. Simulations are conducted to evaluate the optimality criteria. An example based on the proposed method and real-life data is presented.

e-mail: yangh@medimmune.com

### CALCULATING SAMPLE SIZE FOR STUDIES WITH EXPECTED ALL-OR-NONE NONADHERENCE AND SELECTION BIAS

Michelle D. Shardell\*, University of Maryland School of Medicine  
Samer S. El-Kamary, University of Maryland School of Medicine

We develop sample size formulas for studies aiming to test mean differences between a treatment and control group when all-or-none

nonadherence (noncompliance) and selection bias are expected. Recently published work addressed the increased variances within groups defined by treatment assignment when nonadherence occurs, compared to the scenario of full adherence, under the assumption of no selection bias. In this paper, we extend this work to allow selection bias in the form of systematic differences in means and variances between latent adherence subgroups. We illustrate the approach by performing sample size calculations to plan clinical trials with and without pilot adherence data.

e-mail: mshardel@epi.umaryland.edu

### COMPARISON OF DIFFERENT SAMPLE SIZE DESIGNS - GROUP SEQUENTIAL VERSUS RE-ESTIMATION

Xiaoru Wu\*, Columbia University  
Lu Cui, Eisai Medical Research Inc.

Adaptive sample size designs include group sequential and sample size re-estimation methods. Although the group sequential likelihood ratio test has been proved to be optimal among all group sequential tests with a similar structure, it has some inherent inflexibility in the final sample size since its maximum sample size is fixed and it can only stop at several pre-specified information time. Taking the flexibility in the final sample size into consideration, the simulation results based on a new performance assessment criterion suggest that the re-estimation method generally performs better than the group sequential method in terms of delivering more on the targeted final sample size and power. This is also true even when we select the optimal group sequential likelihood ratio test among all group sequential tests.

e-mail: xw2144@columbia.edu

### ON THE ROLE OF BASELINE MEASUREMENTS FOR CROSSOVER DESIGNS UNDER THE SELF AND MIXED CARRYOVER EFFECTS MODEL

Yuanyuan Liang\*, University of Texas Health Science Center at San Antonio  
Keumhee Chough Carriere, University of Alberta

Generally, crossover designs are not recommended when carryover effects are present and when the primary goal is to obtain an unbiased estimate of the treatment effect. In some cases, baseline measurements are believed to improve design efficiency. This paper examines the impact of baselines on optimal designs using two different assumptions about carryover effects during baseline periods and employing a non-traditional crossover design model. As anticipated, baseline observations improve design efficiency considerably for two-period designs, which use the data in the first period only to obtain unbiased estimates of treatment effects, while the improvement is rather modest for three- or four-period designs. Further, we find little additional benefits for measuring baselines at each treatment period as compared to measuring baselines only in the first period. Although our study of baselines did not change the results on optimal designs that are reported in the literature, the problem of strong model dependency problem is generally recognized. The advantage of using multi-period designs is rather evident, as we found that extending

two-period designs to three- or four-period designs significantly reduced variability in estimating the direct treatment effect contrast.

e-mail: liangy@uthscsa.edu

## EFFICIENCY OF STUDY DESIGNS IN DIAGNOSTIC RANDOMIZED CLINICAL TRIALS

Bo Lu\*, The Ohio State University  
Constantine Gatsonis, Brown University

From the patients' management perspective, a good diagnostic test should contribute to both reflecting the true disease status and improving clinical outcomes. Two study designs for the randomized clinical trial—the two-arm design and the paired design are compared in the evaluation of diagnostic tests with patient outcomes as the primary endpoint. In the conventional two-arm design, patients are randomized to one of the diagnostic tests. In the paired design, patients undergo both tests and randomization occurs in the patients with discordant test results. Treatment will be applied based on test results. The follow-up clinical outcomes will be measured to determine the prognostic value of the tests. The paired design is shown to be more efficient than the two-arm design when the operating characteristics of the tests are given. The efficiency gain depends on the discordant rate of test results. Estimation of important quantities under the paired design is derived and simulation studies are also conducted to verify the theoretical results. The method is illustrated with an example of designing a randomized study on preoperative staging of bladder cancer.

e-mail: blu@cph.osu.edu

## STATISTICAL VALIDITY AND POWER FOR TESTING FOR HETEROGENEOUS EFFECTS WITH QUANTITATIVE TRAITS AND ITS APPLICATION TO PHARMACOGENETIC STUDIES

Todd G. Nick\*, Cincinnati Children's Hospital Medical Center  
Mi-ok Kim, Cincinnati Children's Hospital Medical Center  
Chunyan Liu, Cincinnati Children's Hospital Medical Center  
Yu Wang, Cincinnati Children's Hospital Medical Center

In pharmacogenetic studies, it is often of interest to look for evidence of a difference in treatment effect in complementary groups. For example, such groups are typically defined based on a traditional classification of metabolizing group (e.g. poor versus extensive metabolizes). When the trait is quantitative and the comparison is between two groups, the conventional t-test or rank sum test are often used. To detect differences when heterogeneity is present, O'Brien proposed extensions to the t-test and rank sum test and called these tests generalized t-test and generalized rank sum test. For the generalized tests, group membership is regressed against the trait using a quadratic model. We extend the generalized tests by using restrictive cubic splines. The statistical properties under different effect size and distributions were evaluated. Additionally, alternative hypothesis were generated assuming only a location shift, only a scale shift, and both location and scale shift. The generalized test using splines and O'Brien's generalized tests are compared with standard tests. The generalized t-test using splines performed well in regard to power

when the distributions are skewed or contaminated. The method provided no improvement over the generalized t-test when there was only a shift in location.

e-mail: todd.nick@cchmc.org

## 92. SURVIVAL ANALYSIS

### PREDICTION AND MISCLASSIFICATION IN RIGHT CENSORED TIME-TO-EVENT DATA

Keith A. Betts\*, Harvard School of Public Health  
David P. Harrington, Harvard School of Public Health and Dana-Farber Cancer Institute

A common goal in medical studies with survival data is to stratify patients according to predicted risk based on their covariate values. Typically, this consists of fitting a proportional hazards regression model and dividing patients by their subject specific estimated relative risk. Using this method, it is difficult to assess the model's predictive accuracy in an intuitive and interpretable manner. We reframe the problem in terms of prediction error. We propose methodology using working survival models to make predictions in terms of failure time intervals. We propose two measures of prediction error which are consistently estimated regardless of whether the model was correctly specified. We demonstrate a resampling technique that approximates the large sample distribution of the error statistics, and can be used to differentiate between two models on the basis of prediction error. We demonstrate our methodology through simulation study as well as through a data application.

e-mail: kbetts@fas.harvard.edu

### INCORPORATING RATE OF CHANGE INTO TREE STRUCTURED MODELS WITH TIME VARYING COVARIATES

Meredith J. Lotz\*, University of Pittsburgh  
Stewart J. Anderson, University of Pittsburgh  
Sati Mazumdar, University of Pittsburgh

Tree-structured survival analysis creates prognostic groups which can be used to predict risk and assist clinicians in making difficult treatment decisions. Typically, only baseline covariates are utilized in tree-structured survival analysis. However, covariate values are often measured multiple times after baseline, and often, their rate of change in addition to their baseline value may assist in predicting the event of interest. We propose a time-dependent tree-structured survival analysis model which can be used to update an individual's risk based on their changing covariate values. For each time dependent covariate, a linear model is used to regress the covariate versus time for each individual assuming that the slope is random. The slopes and the baseline values of the time dependent covariate are then included along with other baseline covariates in a tree-structured survival analysis model. The result is a model which provides prognostic groups based on not only baseline covariate values but also the rate of change over time of the time varying covariates. An illustrative example of our method is provided.

e-mail: mel20@pitt.edu

# ABSTRACTS

## EXPONENTIAL TILT MODELS IN THE PRESENCE OF CENSORING

Chi Wang\*, Johns Hopkins University  
Zhiqiang Tan, Rutgers University  
Thomas A. Louis, Johns Hopkins University

We study application of the Exponential Tilt Model (ETM) to compare survival distributions in two groups. As a semi-parametric model, the ETM assumes a parametric form for the density ratio of the two distributions. The ETM accommodates a broad array of parametric models such as the log-normal and gamma models and serves to complement the proportional hazards and accelerated failure models. We develop a non-parametric likelihood approach to estimating ETM parameters in the presence of censoring. In simulation studies, we compare the ETM to the Proportional Hazards Model (PHM). When the proportional hazards assumption is not satisfied but the ETM assumption is, we show that the ETM has better power for testing the hypothesis of no difference between the two groups. And, importantly, when the ETM relation is not satisfied but the PHM assumption is, the ETM can still have power close to that of the PHM.

e-mail: chwang@jhsph.edu

## A RISK-ADJUSTED O-E CUSUM WITH V-MASK IN A CONTINUOUS TIME SETTING

Jie (Rena) Sun, University of Michigan  
John D. Kalbfleisch, University of Michigan

Some recent research has examined the CUSUM control chart in a medical setting. Motivated by a risk-adjusted one-sided CUSUM procedure with a continuous time setting, we introduce a risk-adjusted O-E (Observed-Expected) CUSUM to simultaneously monitor 'worse than expected' and 'better than expected' by using a V-shaped mask (V-mask) as the decision criterion. This approach has several advantages over the one-sided CUSUM. Simulation studies were conducted to test the performance of the proposed method, which was compared to the one-sided CUSUM method. Control limits (or threshold value for signaling) were obtained for facilities with different sizes by controlling the false alarm rate over a period of given length. A case study was carried out for 67 liver transplantation programs, using the proposed O-E CUSUM with a V-mask, the one-sided CUSUM, and the current-used model for the Program-Specific Reports (PSR) in the Scientific Registry of Transplant Recipients (SRTR). The presentation concludes with a brief discussion on the unique features of the proposed procedure.

e-mail: renajsun@umich.edu

## BIAS-CORRECTED LOGRANK TEST WITH DEPENDENT CENSORING

Yabing Mai\*, Merck & Co., Inc.  
Eric V. Slud, University of Maryland

The asymptotic property of the logrank and stratified logrank tests varies across different types of assumptions regarding the dependence

of the censoring and the survival times. When the treatment group and the covariates are conditionally independent given that the subject is still at risk, the logrank statistic is asymptotically standard normally distributed under the null hypothesis. Given this assumption, the stratified logrank statistic has similar asymptotic properties to the logrank statistic. However, if the assumption of conditional independence fails, the logrank statistic is generally biased and the bias is considered non-negligible. We discuss and extend an available bias-correction method of DiRienzo and Lagakos (2001) with unknown and estimated censoring distribution function given the provided treatment group and covariates. We obtain the correct asymptotic distribution of the bias-corrected test statistic when stratum-based Kaplan-Meier estimators of the conditional censoring distribution are substituted into the statistic. Within this framework, we prove the asymptotic unbiasedness of the corrected test and find a consistent variance estimator. Major theoretical results and motivations of future studies are confirmed by a series of simulation studies.

e-mail: yabing\_mai@merck.com

## THE COMPARISON OF ALTERNATIVE SMOOTHING METHODS FOR FITTING NON-LINEAR EXPOSURE-RESPONSE RELATIONSHIPS WITH COX MODELS IN A SIMULATION STUDY

Usha S. Govindarajulu\*, Harvard Medical School  
Betty J. Malloy, American University  
Bhaswati Ganguli, University of Calcutta  
Donna Spiegelman, Harvard School of Public Health  
Ellen A. Eisen, University of California-Berkeley and Harvard School of Public Health

We examined the behavior of alternative smoothing methods for modeling environmental epidemiology data. Model fit can only be examined when the true exposure-response curve is known and so we used simulation studies to examine the performance of penalized splines (P-splines), restricted cubic splines (RCS), natural splines (NS), and fractional polynomials (FP). Survival data were generated under six plausible exposure-response scenarios with a right skewed exposure distribution, typical of environmental exposures. Cox models with each spline or FP were fit to simulated datasets. The best models, e.g. degrees of freedom, were selected using default criteria for each method. The root mean-square error (rMSE) and area difference were computed to assess model fit and bias (difference between the observed and true curves). The test for linearity was a measure of sensitivity and the test of the null was an assessment of statistical power. No one method performed best according to all four measures of performance, however, all methods performed reasonably well. The model fit was best for P-splines for almost all true positive scenarios, although fractional polynomials and RCS were least biased, on average.

e-mail: usha@alum.bu.edu

## **UTILIZING BIOSTATISTICAL METHODS IN THE ANALYSIS OF DATA IN DISCRIMINATION CASES**

Joseph L. Gastwirth\*, George Washington University  
Qing Pan, George Washington University

In the Diaz v. Eagle Produce case, both parties did not analyze the statistical information available to them. In reversing a lower court's summary judgment decision for the employer the appellate opinion described patterns of some simple measures, e.g. the average age of new hires decreased over time and that older workers were at a higher risk of being discharged after a new supervisor took charge. Using the Cochran-Armitage trend test to examine the hiring data and the proportional hazards model to analyze the discharge data, one finds statistical support for the appellate court's finding. Although there were only 44 employees, using Fisher's summary chi-square test to combine the p-values of both tests rejects the null hypothesis that both processes were fair with a p-value < .01. This area of application indicates that more research on combining the results of analyses of small datasets as well as further study of the power of these methods are needed.

e-mail: jlgast@gwu.edu

## **93. FUNCTIONAL DATA ANALYSIS**

### **GENERALIZED MULTILEVEL FUNCTIONAL REGRESSION**

Ciprian M. Crainiceanu, Johns Hopkins University  
Ana-Maria Staicu\*, University of Bristol-UK  
ChongZhi Di, Johns Hopkins University

We introduce Generalized Multilevel Functional Linear Models (GMFLM), a novel statistical framework motivated by and applied to the Sleep Heart Health Study (SHHS), the largest community cohort study of sleep. The primary goal of SHHS is to study the association between sleep disrupted breathing (SDB) and adverse health effects. An exposure of primary interest is the sleep electroencephalogram (EEG), which was observed for thousands of individuals at two visits, roughly 5 years apart. This unique study design led to the development of models where the outcome, e.g. hypertension, is in an exponential family and the exposure, e.g. sleep EEG, is multilevel functional data. We show that GMFLMs are, in fact, generalized multilevel mixed effect models. Two consequences of this result are that: 1) the mixed effects inferential machinery can be used for GMFLM and 2) functional regression models can be extended naturally to include, for example, additional covariates, random effects and nonparametric components. We propose and compare two inferential methods based on the parsimonious decomposition of the functional space.

e-mail: a.staicu@bristol.ac.uk

## **STOCHASTIC FUNCTIONAL DATA ANALYSIS: A DIFFUSION MODEL-BASED APPROACH**

Bin Zhu\*, University of Michigan  
Peter X.-K. Song, University of Michigan  
Jeremy M.G. Taylor, University of Michigan

We consider the problem of estimating an unknown smooth function given functional data. The unknown function is treated as the realization of a stochastic process. We propose a new type of continuous-discrete state space model, called a stochastic velocity model. The resulting model allows straightforward and meaningful interpretation. The method of smoothing splines is a special case of this approach. The likelihood of the model is derived with Euler approximation and data augmentation. Bayesian inference is carried out via a Markov Chain Monte Carlo algorithm. The proposed model and method are illustrated using a blood oxygenation-level dependent signal data, and prostate specific antigen data.

e-mail: BZHU@UMICH.EDU

## **WAVELET-BASED FUNCTIONAL MIXED MODELS VIA DPM**

Alejandro Villagran\*, Rice University  
Sang Han Lee, New York University  
Marina Vannucci, Rice University

Recently, various experimental designs in public health and bioinformatics require the use of functional mixed models (FMM), i.e., a functionalized extension of linear mixed models, conditional on modern technologies allowing researchers to record data sampled on a fine grid. Morris and Carroll (JRSS-B, 2006) developed wavelet-based functional mixed models that are flexible enough to accommodate a broad range of functional data. We attempt to extend these methods by fitting the functional mixed model in the wavelet domain and by relieving the assumption of normality on the random effect functions to any distributional form via a Dirichlet Process Mixture (DPM) prior. We use the Gibbs sampler algorithm to estimate the posterior distributions of the parameters. We illustrate this methodology with a simulated example.

e-mail: av8@rice.edu

## **DATA DRIVEN ADAPTIVE SPLINE SMOOTHING WITH APPLICATIONS TO EPILEPTIC EEG DATA**

Ziyue Liu\*, University of Pennsylvania School of Medicine  
Wensheng Guo, University of Pennsylvania School of Medicine

The band power of 26-50Hz of EEG has been shown to be a potential predictor of epileptic seizure and can help neurologists locating seizure's spatiotemporal initiation. It is flat before seizure, quick changes around seizure and returns to stationary after seizure. Traditional methods smooth it using a global smoothing method, leading to oversmooth the regions around seizure and undersmooth of other regions. In this paper we propose an adaptive smoothing spline model where the smoothing parameter changes across the time domain, allowing the model to adapt to the change of roughness in

# ABSTRACTS

the data. We model the penalty function by a step function with data driven segmentation. We impose a binary tree structure on the step function and propose a Best Basis like search algorithm. We propose an AIC like criterion based on the generalized likelihood to select the optimal segmentation. We derive the state space representation for efficient computation. The proposed method smaller true mean square errors comparing to non-adaptive smoothing spline, wavelet shrinkage and Bayesian adaptive P-spline for a wide range of signals. Application to epileptic EEG example reveals that after remaining flat for 7 minutes, the band power starts to rise 33 seconds before the onset of seizure. This finding will enable development of short term predictive model.

e-mail: zliu5@mail.med.upenn.edu

## MODELLING LABOR CURVES IN WOMEN ATTEMPTING A VAGINAL BIRTH AFTER A CESAREAN USING A B-SPLINES BASED SEMIPARAMETRIC NONLINEAR MIXED EFFECTS MODEL

Angelo Elmí\*, University of Pennsylvania  
Sarah Ratcliffe, University of Pennsylvania  
Sam Parry, University of Pennsylvania  
Wensheng Guo, University of Pennsylvania

New methodology is developed for the Semiparametric Nonlinear Mixed Effects Model and is applied for the analysis and comparison of labor curves from women attempting a vaginal birth after cesarean (VBAC) delivery. Existing approaches estimate the shape function with smoothing splines and use a backfitting approach that iterates between two mixed effects models and consequently two separate likelihoods. Such an algorithm will not be guaranteed to converge; and because it incorrectly assesses the variability in the parameter estimates, it will not give valid inferences. We develop a new model replacing the smoothing spline with B-splines. This approach requires only one mixed effects model, and since all parameters are estimated from the same likelihood, an iterative algorithm will be guaranteed to converge and likelihood-based inferences will be valid. We apply this model to make comparisons between the average labors of women who do and do not experience uterine rupture and to find the earliest time point where clinicians could distinguish between the average labor curves in different groups which enables them to make the potentially life-saving decision of whether or not to perform a cesarean delivery. Since the dimension of integration is reduced, we are able to obtain accurate representations of the log-likelihood based on Adaptive Gaussian Quadrature.

e-mail: afelmi@mail.med.upenn.edu

## A BAYESIAN REGRESSION MODEL FOR MULTIVARIATE FUNCTIONAL DATA

Ori Rosen\*, University of Texas at El Paso  
Wesley K. Thompson, University of California San Diego

In this paper we present a model for the analysis of multivariate functional data with unequally spaced observation times that may differ among subjects. Our method is formulated as a Bayesian

mixed-effects model in which the fixed part corresponds to the mean functions, and the random part corresponds to individual deviations from these mean functions. Covariates can be incorporated into both the fixed and the random effects. The random error term of the model is assumed to follow a multivariate Ornstein-Uhlenbeck process. For each of the response variables, both the mean and the subject-specific deviations are estimated via low-rank cubic splines using radial basis functions. Inference is performed via Markov chain Monte Carlo methods.

e-mail: ori@math.utep.edu

## ANALYSIS OF LONG PERIOD VARIABLE STARS WITH A NONPARAMETRIC SIGNIFICANCE TEST OF NO TREND

Woncheol Jang\*, University of Georgia  
Cheolwoo Park, University of Georgia  
Jeongyoun Ahn, University of Georgia  
Martin Hendry, University of Glasgow

The study of variable stars has a long and illustrious history in astronomy, making crucial contributions to our understanding of many fields, from stellar birth and evolution to the calibration of the extragalactic distance scale. Variable stars are characterized by showing significant variation in their brightness over time. We perform a time series analysis of the periods between maximum brightness of a group of 378 long period variable stars. The objective of this study is to identify the stars which display certain trends in their period, via multiple testing of a mean for non-stationary time series model. We test the null hypothesis of no trend in each time series based on high dimension normal mean inference, while controlling the false discovery rate to adjust multiplicity. Functional clustering and principal component analysis is used to account for the non-stationary error structure. We investigate the performance of the proposed methods using simulation studies.

e-mail: jang@uga.edu

## WEDNESDAY, MARCH 18, 2009 8:30-10:15 AM

### 94. EVALUATING MARKERS FOR RISK PREDICTION

## DECISION CURVE ANALYSIS: A SIMPLE, NOVEL METHOD FOR THE EVALUATION OF PREDICTION MODELS, DIAGNOSTIC TESTS AND MOLECULAR MARKERS

Andrew J. Vickers\*, Memorial Sloan-Kettering Cancer Center

Prediction models and molecular markers and models are currently evaluated in terms of accuracy, using metrics such as sensitivity and specificity or the area-under-the-curve (AUC). A model is thought to be a good one if it is accurate; a marker is claimed to be of value if it increases accuracy of clinical predictors. However, it is unclear exactly how high an AUC, or how great an increment in AUC, is sufficient

to justify clinical use of a model or marker. Reclassification metrics similarly fail to give a clear answer as to clinical value: how much reclassification is enough to warrant measuring a marker. Evaluating models and markers in terms of clinical consequences is the remit of a field known as 'decision analysis'. The drawback of traditional decision analysis is that it requires additional information, for example, on the benefits, harms and costs of treatment, or on patient preferences for different health states. Decision curve analysis is a simple, decision-analytic method that can be directly applied to the data set of a model or marker, without the need for external data. Because the method is decision analytic, it can be used to tell us whether or not to use a model in the clinic, or whether a marker is worth measuring. Further information on decision curve analysis can be found at [www.decisioncurveanalysis.org](http://www.decisioncurveanalysis.org)

e-mail: vickersa@mskcc.org

## **ON INCORPORATING BIOMARKERS INTO MODELS FOR ABSOLUTE RISK PREDICTION MODELS**

Ruth Pfeiffer\*, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health  
Mitchell Gail, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health

Absolute risk is the probability that an individual who is free of a given disease at an initial age,  $a$ , will develop that disease in the subsequent interval  $(a, t]$ . Absolute risk is reduced by mortality from competing risks. Models of absolute risk that depend on covariates have been used to design intervention studies, to counsel patients regarding their risks of disease and to inform clinical decisions. While models that predict the absolute risk of breast cancer, colorectal cancer or melanoma are well calibrated when validated in independent data, their discriminatory power is typically low, which makes them ill suited for screening applications. One hope is to improve their discriminatory ability by incorporating biomarker information. Several statistical issues need to be addressed for such an undertaking: 1) how to optimally incorporate these markers into a risk prediction model, and 2) how to assess the improvement of a model from adding markers. We present methods for comparing the performance of different absolute risk models. We illustrate our methods by comparing a model for breast cancer based on a woman's age, and personal characteristics including reproductive risk factors, with a model that also includes genetic markers.

e-mail: pfeiffer@mail.nih.gov

## **ESTIMATING THE CAPACITY FOR IMPROVEMENT IN RISK PREDICTION WITH A MARKER**

Wen Gu\*, University of Washington and Fred Hutchinson Cancer Research Center  
Margaret S. Pepe, University of Washington and Fred Hutchinson Cancer Research Center

Consider a set of baseline predictors  $X$  to predict a binary outcome  $D$  and let  $Y$  be a novel marker or predictor. This talk is concerned with evaluating the performance of the augmented risk model  $P(D =$

$1|Y,X)$  compared with the baseline model  $P(D = 1|X)$ . The diagnostic likelihood ratio,  $DLRX(y)$ , quantifies the change in risk obtained with knowledge of  $Y = y$  for a subject with baseline risk factors  $X$ . The notion is commonly used in clinical medicine to quantify the increment in risk prediction due to  $Y$ . It is contrasted here with the notion of covariate adjusted effect of  $Y$  in the augmented risk model. We also propose methods for making inference about  $DLRX(y)$ .

Case-control study designs are accommodated. The methods provide a mechanism to investigate if the predictive information in  $Y$  varies with baseline covariates. In addition, we show that when combined with a baseline risk model and information about the population distribution of  $Y$  given  $X$ , covariate specific predictiveness curves can be estimated. These curves are useful to an individual in deciding if ascertainment of  $Y$  is likely to be informative or not for him. Data from two studies, one is a study of the performance of hearing screening tests for infants; the other concerns the value of serum creatinine in diagnosing renal artery stenosis, will be used to illustrate the methodology.

e-mail: wengu@u.washington.edu

## **CALCULATING DISEASE RISK: EVALUATING RISK PREDICTION MODELS USING RISK STRATIFICATION TABLES**

Holly Janes\*, University of Washington and Fred Hutchinson Cancer Research Center  
Margaret S. Pepe, University of Washington and Fred Hutchinson Cancer Research Center  
Jessie Gu, University of Washington and Fred Hutchinson Cancer Research Center

A multitude of new markers are being evaluated in the hopes that they may aid patients and clinicians in predicting an individual's risk of disease. A new approach to evaluating these markers, called the risk stratification approach, was recently proposed by Nancy Cook and colleagues (Cook 2006). This involves cross-tabulating risk predictions on the basis of models with and without the new biomarker, and has been widely adopted in the literature. We argue that important information with regard to three important model validation criteria can be extracted from risk stratification tables: 1) model fit or calibration; 2) capacity for risk stratification; and 3) accuracy of classifications based on risk. However, we also caution against misuses of the method. For example, risk stratification tables are not useful for comparing non-nested risk prediction models, nor can they be directly applied to evaluate models fit to case-control data. We provide alternative suggestions for evaluating risk prediction models in these settings. The concepts are illustrated using models for predicting breast cancer risk.

e-mail: hjanest@scharp.org

## **95. ADVANCES IN FUNCTIONAL DATA ANALYSIS**

### **FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS WITH SURVEY DATA**

Herve, Cardot\*, Institut de Mathematiques, Universite de Bourgogne

This talk aims at presenting how functional principal components analysis can be performed when curve data are collected with survey

# ABSTRACTS

sampling strategies. We consider estimators based on the Horvitz-Thompson approach and derive asymptotic properties of the mean function and the eigenelements of the covariance function with the help of the influence function. We also prove that we can get consistent estimators of the asymptotic variance. A simulation study, which focuses on stratified sampling, allows to confirm the good properties of our estimators.

e-mail: herve.cardot@u-bourgogne.fr

## CONCEPT OF DENSITY FOR FUNCTIONAL DATA

Peter Hall\*, University of Melbourne and University of California at Davis  
Aurore Delaigle, University of Bristol

The notion of a probability density for a random function is not as straightforward as in finite-dimensional cases. While it is possible to rank points in the function space in terms of their density within a ball of given nonzero radius, the conventional concept of a probability density function, constructed with respect to a ball of infinitesimal radius, is not well defined. We suggest instead a transparent and meaningful surrogate for density, defined as the average value of the logarithms of the densities of the distributions of principal component scores, for a given dimension. This density approximation is readily estimable from data. Indeed, methodology for estimating densities of principal component scores is of independent interest; it reveals shape differences that have not previously been considered.

e-mail: halpstat@ms.unimelb.edu.au

## DECIDING THE DIMENSION OF EFFECTIVE DIMENSION REDUCTION SPACE FOR FUNCTIONAL DATA

Yehua Li, University of Georgia  
Tailen Hsing\*, University of Michigan

In this talk, we discuss regression models with a functional predictor and a scalar response, where the response depends on the predictor only through a finite number of projections. The linear subspace spanned by these projections is called the effective dimension reduction (EDR) space. To determine the dimensionality of the EDR space, we focus on the principal component scores of the functional predictor, and propose three sequential testing procedures under the assumption that the predictor has an elliptically contoured distribution. The proposed procedures are supported by theory and validated by a simulation study. Applications on two real data sets will also be presented.

e-mail: thsing@umich.edu

## 96. NEW STATISTICAL CHALLENGES AND ADVANCEMENTS IN GENOME-WIDE ASSOCIATION STUDIES

### ON THE ADJUSTMENT FOR COVARIATES IN GENETIC ASSOCIATION ANALYSIS: A NOVEL, SIMPLE PRINCIPLE TO INFER CAUSALITY

Stijn Vansteelandt, University of Ghent  
Christoph Lange\*, Harvard School of Public Health

In genetic association studies, different complex phenotypes are often associated with the same marker. Such associations can be indicative of pleiotropy (i.e. common genetic causes), of indirect genetic effects via one of these phenotypes, or can be solely attributable to non-genetic/environmental links between the traits. To identify the phenotypes with the inducing genetic association, statistical methodology is needed that is able to distinguish between the different causes of the genetic associations. Here, we propose a simple, general adjustment principle that can be incorporated into many standard genetic association tests which are then able to infer whether a SNP has a direct biological influence on a given trait other than through the SNP's influence on another correlated phenotype. Using simulation studies, we show that, in the presence of a non-marker related link between phenotypes, standard association tests without the proposed adjustment can be biased. In contrast to that, the proposed methodology remains unbiased. Its achieved power levels are identical to those of standard adjustment methods, making the adjustment principle universally applicable in genetic association studies. The principle is illustrated by an application to three genome-wide association analysis. Key-words: causal diagram; direct effect; genetic pathways; mediation; pleiotropy.

e-mail: clange@hsph.harvard.edu

## FAMILY-BASED ASSOCIATION TEST FOR MULTIPLE TRAITS SPEAKER

Heping Zhang\*, Yale University  
Ching-Ti Liu, Boston University  
Xueqin Wang, Sun-Yat Sen University  
Wensheng Zhu, Yale University

Early family studies of psychiatric disorders began about a century ago, but our understanding for the genetics of mental and behavioral disorders remains limited. One challenge arises from the fact that multiple phenotypes are needed to characterize psychiatric disorders that usually do not occur alone. In fact, comorbidity is a rule other than exception. To address this challenge, I will first demonstrate the usefulness of considering multiple traits in genetic studies of complex disorders. Then, I will present a non-parametric test to studying the association between multiple traits and a candidate marker. After a brief summary for the theoretical properties of the test, the nominal type I error and power of the proposed test will be compared with existing test through simulation studies. The advantage of the proposed test will also be demonstrated by a study of alcoholism.

e-mail: heping.zhang@yale.edu

## A PENALIZED LIKELIHOOD APPROACH TO HAPLOTYPE SPECIFIC ANALYSIS

Jung-Ying Tzeng, North Carolina State University  
Howard D. Bondell\*, North Carolina State University

Haplotypes can hold key information to understand the role of candidate genes in disease etiology. While many existing methods are available for studying haplotype effects at either the global or individual levels, few provide systematic evaluation on the pattern and structure of the haplotype effects. In most work, haplotype inference focuses on relative effects compared to a baseline haplotype. Ideally, all haplotype effects should be compared to determine a group structure among the haplotypes. This can be done as a secondary post-hoc analysis on pairwise differences as in ANOVA. However, this combined analysis often cannot identify the appropriate structure and tends to lack power. To resolve these issues, we propose a penalized likelihood approach using an L1 penalty on the pairwise differences of haplotype effects. The proposed method treats haplotypes as a factor of multiple levels without a pre-determined baseline level. It simultaneously carries out the effect estimation and comparison of all haplotypes, and outputs the haplotype group structure based on their effect sizes. We use simulation studies to demonstrate the informativeness and power of the proposed method, and to illustrate that the method can better identify the haplotype effect structure than the traditional haplotype association methods.

e-mail: bondell@stat.ncsu.edu

## STATISTICAL METHODS FOR GENE MAPPING USING HIGH DENSITY SNPS IN FAMILY SAMPLES

Josée Dupuis\*, Boston University School of Public Health

Genome wide association scans have been performed on multiple studies, including some family based cohorts. While association analysis in family based samples presents some statistical challenges because of the correlated nature of the observations, the advantages of family designs in genetic studies greatly outweigh the added analysis complexity. We present statistical approaches to exploit family features when looking for genetic variants influencing quantitative traits of interest. We illustrate our methods with a high density scan in the Framingham Heart Study cohorts.

e-mail: dupuis@bu.edu

## 97. RESPONSE-ADAPTIVE DESIGNS FOR CLINICAL TRIALS

### RESPONSE ADAPTIVE RANDOMIZATION IN NON-INFERIORITY TRIALS

Lanju Zhang\*, MedImmune LLC  
Harry Yang, MedImmune LLC

Response adaptive randomization has been studied extensively but only focused on superiority trials. The main drive for response adaptive randomization is ethical consideration, which is as compelling in

non-inferiority clinical trials as in superiority clinical trials. If the experimental drug is not worse than the active control, one should not randomize equal number of patients to the experimental drug arm. The talk focuses on how to derive optimal allocation proportions for non-inferiority trials. Power calculation method of Farrington and Manning for non-inferiority trials is used in the derivation. The optimal allocation proportion is a solution to an equation, and a closed form is not available. Simulation is used to study the properties of the proposed allocation proportion.

e-mail: zhangla@medimmune.com

### SEQUENTIAL MONITORING RESPONSE-ADAPTIVE RANDOMIZED CLINICAL TRIALS

Feifang Hu\*, University of Virginia  
Hongjian Zhu, University of Virginia

In clinical trials, sequential monitoring procedures are often implemented to stop the trials earlier, therefore reduce the cost and the sample size. Response-adaptive randomization has been developed and demonstrated to be desirable randomization for many clinical trials. Is it possible to combine these two sequential procedures together in a clinical trial? In this talk, we answer this question theoretically by showing that the joint distribution of some important sequential statistics is asymptotically Brownian process in response-adaptive randomized trials. Therefore, we can apply sequential monitoring procedures to response-adaptive randomized trials. Simulated studies also support our theoretical results.

e-mail: fh6e@virginia.edu

### USING SHORT-TERM RESPONSE INFORMATION TO FACILITATE ADAPTIVE RANDOMIZATION FOR SURVIVAL CLINICAL TRIALS

Xuelin Huang\*, The University of Texas MD Anderson Cancer Center  
Jing Ning, The University of Texas MD Anderson Cancer Center  
Yisheng Li, The University of Texas MD Anderson Cancer Center  
Donald A. Berry, The University of Texas MD Anderson Cancer Center

Increased survival is a common goal of cancer clinical trials. Due to the long periods of observation and follow-up to assess patient survival outcome, it is difficult to use outcome-adaptive randomization in these trials. In practice, often information about a short-term response is quickly available during or shortly after treatment, and this short-term response is a good predictor for long-term survival. For example, complete remission of leukemia can be achieved and measured after a few cycles of treatment. It is a short-term response that is desirable for prolonging survival. We propose a new design for survival trials when such short-term response information is available. We use the short-term information to ``speed up'' the adaptation of the randomization procedure. We establish a connection between a short-term response and long-term survival through a Bayesian model, first by using prior clinical information, and then by dynamically updating the model according to information accumulated in the ongoing trial. Interim monitoring and final decision making are based upon inference on

# ABSTRACTS

the primary outcome of survival. The new design can more effectively assign patients to the better treatment arms. We demonstrate these properties through simulation studies.

e-mail: xlhuang@mdanderson.org

## 98. DEVELOPMENT OF BAYESIAN SURVIVAL AND RISK ANALYSIS

### CLASSICAL AND BAYES ESTIMATION FOR ADDITIVE HAZARDS REGRESSION MODELS

Stuart R. Lipsitz\*, Brigham and Women's Hospital  
Debajyoti Sinha, Florida State University  
M. Brent McHenry, Bristol-Myers Squibb  
Malay Ghosh, University of Florida

For additive hazards regression models, when the hazard rates are close to 0 and/or censoring is high, there often exist convergence problems in the Newton-Raphson algorithm since the MLE is on or close to the boundary of the parameter space. As alternatives, we propose a weighted least squares (WLS) method-of-moments techniques, as well as a novel empirical Bayesian framework. The integrated likelihood in the empirical Bayesian framework is obtained via integrating the unknown prior of the nonparametric baseline cumulative hazard, and can be maximized using standard statistical software. Unlike the corresponding full Bayes method, our empirical Bayes estimates of regression parameters, survival curves and their corresponding standard errors have easy to compute closed form expressions and require no elicitation of hyperparameters of the prior. We illustrate the implementation and advantages of our methodology via a reanalysis of a survival dataset and a simulation study using existing statistical software such as SAS.

e-mail: SLIPSITZ@PARTNERS.ORG

### BAYESIAN DEVELOPMENT OF A GENERALIZED LINK FUNCTION FOR BINARY RESPONSE DATA

Xia Wang\*, University of Connecticut  
Dipak K. Dey, University of Connecticut

This paper introduces a flexible skewed link function for modeling binary data with covariates based on the generalized extreme value (GEV) distribution. Extreme value techniques have been widely used in many disciplines for risk analysis. However, its application in the binary data context is sparse and its strength as a link function has never been explored. The commonly used complementary log-log link is a special case in the GEV distribution family but it is prone to link misspecification because of its positive and fixed shape parameter. The GEV link is flexible in fitting the skewness in the data with an unknown shape parameter value. Using Bayesian methodology, it automatically detects the skewness in the data along with the model fitting. The propriety of posterior distributions under various proper and improper priors is explored in details. The flexibility of the proposed model is illustrated by a unique billing data set of the electronic payments system adoption from a Fortune 100 company. This link is especially attractive when there exists extreme difference

between 0 and 1 observations in the binary response such that the response curve could be extremely skewed.

e-mail: xia.wang@uconn.edu

## ANALYSIS OF EXTREME DRINKING IN PATIENTS WITH ALCOHOL DEPENDENCE USING PARETO REGRESSION

Sourish Das\*, Duke University  
Ofer Harel, University of Connecticut  
Dipak K. Dey, University of Connecticut  
Jonathan Covault, University of Connecticut Health Center, Psychiatry  
Henry R. Kranzler, University of Connecticut Health Center, Psychiatry

We developed a novel Pareto regression model with unknown shape parameter to analyze extreme drinking in patients with Alcohol Dependence (AD). We used a generalized linear models (GLM) framework and a log-link between the shape parameter of the random and systematic components and a Monte Carlo based Bayesian method to implement the analysis. We examined two issues of importance in the study of AD: First, we tested whether a single nucleotide polymorphism within GABRA2 gene, which encodes a subunit of the GABA\_A receptor and has been associated to AD, influenced extreme alcohol intake and second, the efficacy of three psychotherapies for alcoholism in treating extreme drinking behavior. Following 3-month treatment period, during which one of the three psychotherapy treatment, participants were followed up. We also found that women with the high-risk GABRA2 allele had a significantly higher probability of extreme drinking behavior than women with no high-risk allele. Among men, there was no significant effect of GABRA2 genotype on extreme drinking behavior. We found that women who received cognitive behavioral therapy had better outcomes than those receiving either of the other two therapies. Among men, motivational enhancement therapy was the best treatment for the extreme drinking behavior.

e-mail: sourish.das@gmail.com

## 99. PROTEOMICS / METABOLOMICS

### A BAYESIAN APPROACH TO THE ALIGNMENT OF MASS SPECTRA

Xiaoxiao Kong\*, School of Public Health, University of Minnesota  
Cavan Reilly, School of Public Health, University of Minnesota

The need to align spectra is a problem that arises in mass spectrometry. Here we present a novel Bayesian alignment approach. The approach is based on a parametric model which assumes the spectrum and alignment function are Gaussian processes, but the alignment function is monotone. We show how to use the EM algorithm to find the posterior mode of the alignment functions and the mean spectrum for a patient population. We apply the method to characterize the proteome of a set of patients receiving lung transplants.

e-mail: xiaoxiak@biostat.umn.edu

## **TWO-DIMENSIONAL CORRELATION OPTIMIZED WARPING ALGORITHM FOR ALIGNING GCXGC-MS DATA**

Dabao Zhang\*, Purdue University  
Xiaodong Huang, Purdue University  
Fred E. Regnier, Purdue University  
Min Zhang, Purdue University

A two-dimensional (2-D) correlation optimized warping (COW) algorithm has been developed to align 2-D gas chromatography coupled with time-of-flight mass spectrometry (GC $\times$ GC/TOF-MS) data. By partitioning raw chromatographic profiles and warping the grid points simultaneously along the first and second dimensions on the basis of applying a one-dimensional COW algorithm to characteristic vectors, nongrid points can be interpolatively warped. This 2-D algorithm was directly applied to total ion counts (TIC) chromatographic profiles of homogeneous chemical samples, i.e., samples including mostly identical compounds. For heterogeneous chemical samples, the 2-D algorithm is first applied to certain selected ion counts chromatographic profiles, and the resultant warping parameters are then used to warp the corresponding TIC chromatographic profiles. The developed 2-D COW algorithm can also be applied to align other 2-D separation images, e.g., LC $\times$ LC data, LC $\times$ GC data, GC $\times$ GC data, LC $\times$ CE data, and CE $\times$ CE data.

e-mail: zhangdb@purdue.edu

## **MONOISOTOPIC PEAK DETECTION FOR MASS SPECTROMETRY DATA**

Mourad Atlas\*, University of Louisville  
Susmita Datta, University of Louisville

Mass spectrometry has emerged as a core technology for high throughput proteomics profiling. It has enormous potential in biomedical research; however, the complexity of the data poses new statistical challenges for the analysis. Statistical methods and software developments for analyzing proteomics data will continue to be a major area of research in the coming years. In this paper, we develop novel statistical methods for analyzing high dimensional mass-spectrometry proteomics data. We propose to use the chemical knowledge of the isotopic distribution of peptide molecules along with quantitative modeling to detect chemically valuable peaks from each spectrum. A mixture of location-shifted Poisson distribution is fitted to the deamidated isotopic distribution of a peptide molecule. Maximum likelihood estimation by Expectation-Maximization (EM) technique is used to estimate the parameters of the distribution. We then determined the monoisotopic peak for each of the isotopic distribution. Our method is examined through simulations and real data. We compare our method with an existing method of peak detection. Keywords: Mass spectrometry, Proteomics, peaks, isotopic distribution, location-shifted Poisson, monoisotopic peaks.

e-mail: mourad.atlas@louisville.edu

## **A TWO-STAGE APPROACH FOR DETECTING CLUSTERS OF PEAKS WITH PERIODICITY IN NMR SPECTRA**

Anna K. Jolly\*, Emory University  
Amita Manatunga, Emory University  
Tianwei Yu, Emory University

Nuclear magnetic resonance (NMR) spectroscopy has been used increasingly in recent years as a means of obtaining metabolic information from individuals. In our study, we consider NMR spectra obtained from the blood plasma of subjects every hour over a 25-hour period. The identification of peaks with periodic behavior is of interest. A two-stage process is proposed, the first stage of which uses periodic regression to estimate the parameters corresponding to period for the various peaks. In the second stage, a mixture model is used to develop clusters of peaks, taking into account the variability of the estimates obtained in the first stage. Using simulation studies, we demonstrate the performance of the two-stage method and then apply the method to blood plasma spectra. Key words: Periodicity, NMR Spectra, Clustering

e-mail: ajolly@sph.emory.edu

## **SPARSITY PRIORS FOR PROTEIN-PROTEIN INTERACTION PREDICTIONS**

Inyoung Kim\*, Virginia Tech  
Yin Liu, University of Texas Medical School  
Hongyu Zhao, Yale University

Protein-protein interactions play important roles in most fundamental cellular processes including cell cycle, metabolism, and cell proliferation. Therefore, the development of effective statistical approaches to predicting protein interactions based on recently available large scale experimental data is a very important problem. However, due to the number of protein-protein interaction to be observed is very small, the number of parameters to be estimated is very large. Therefore the data is very sparse due to a few number of protein-protein interaction to be observed. In this paper, we incorporate a point-mass mixture prior in the analysis through a Bayesian method. The prediction results between with and without this prior are compared using the large-scale protein-protein interaction data obtained from high throughput yeast two-hybrid experiments. The result demonstrates the advantages of the Bayesian approach with a sparsity prior based on point-mass mixture prior.

e-mail: inyoungk@vt.edu

## **STATISTICALLY APPRAISING AFFINITY-ISOLATION EXPERIMENT PROCESS QUALITY**

Julia Sharp\*, Clemson University  
John Borkowski, Montana State University  
Denise Schmoyer, Oak Ridge National Laboratory  
Greg Hurst, Oak Ridge National Laboratory

Identifying valid protein-protein interactions relies on quality data from affinity-isolation experiments. The quality of an experiment

# ABSTRACTS

can be reduced from biological error, processing error, and random variability. If these errors are of any magnitude, the identification of interacting protein pairs will be hindered. A known mixture of proteins and peptides is processed through a mass spectrometer as a quality control mixture. Statistical quality control (SQC) procedures, including cumulative sum, the individual measurement, and moving range charts are used to assess the stability of the affinity isolation process using the quality control mixture. The SQC measures presented can assist in setting preliminary control limits for identifying out-of-control processes and investigate assignable causes for shifts.

e-mail: jsharp@clemson.edu

## SET ENRICHMENT ANALYSIS METHODS FOR INTEGROMIC STUDIES

Laila M. Poisson\*, University of Michigan  
Debashis Ghosh, Penn State University

The medical community is embracing omics technologies and many elements of systems biology are now being measured on a global scale for clinical samples. As with gene expression, the follow up to any list of differential elements is to search for evidence of enrichment of pathways or other groupings. In this talk we discuss the problem of integrating set enrichment analysis to incorporate data from multiple omics technologies. Issues of sampling and inference will be discussed. Examples are motivated by gene expression and metabolomic data on matched samples of prostate cancer progression.

e-mail: lpoisson@umich.edu

## 100. DETECTING GENE DEPENDENCIES AND CO-EXPRESSION

### DETECTING NON-LINEAR DEPENDENCIES IN GENE CO-EXPRESSION NETWORKS

Alina Andrei\*, University Of Wisconsin-Madison

The reconstruction of gene regulatory networks is a main goal of many biological endeavors. Gene coexpression networks (GCNs) have been widely used as a simple and efficient approach to summarize dependencies among genes, while being computationally inexpensive. In a GCN, the association between two genes is most often quantified by using the Pearson correlation coefficient (PCC) or the mutual information. However, the PCC is suitable for capturing linear dependencies, while the mutual information requires that expression data be discretized, hence incurring a certain information loss. In general, the problem concerning types of dependencies in microarray data is not well studied. We propose an efficient approach to automatically detect non-linear dependencies among genes. Simulation studies show that the method detects non-linear associations with high sensitivity for moderate sample sizes, while controlling the FDR. Practical advantages are demonstrated using a breast cancer study data set.

e-mail: aandrei@biostat.wisc.edu

## A NONPARAMETRIC APPROACH TO DETECT NONLINEAR CORRELATION IN GENE EXPRESSION

Yian Ann Chen\*, Moffitt Cancer Center, University of South Florida  
Jonas S. Almeida, The University of Texas, M.D. Anderson Cancer Center

Adam J. Richards, Medical University of South Carolina  
Peter Müller, The University of Texas, M.D. Anderson Cancer Center  
Raymond J. Carroll, Texas A&M University  
Baerbel Rohrer, Medical University of South Carolina

We propose a distribution-free approach to detect nonlinear relationships by reporting local correlation. The effect of our proposed method is analogous to piece-wise linear approximation although the method does not utilize any linear dependency. The proposed metric, maximum local correlation, was applied to both simulated cases and expression microarray data comparing the rd mouse, which exhibits photoreceptor degeneration with age-matched control animals. The rd mouse is an animal model (with a mutation for the gene Pde6b) for photoreceptor degeneration. Using simulated data, we show that maximum local correlation detects nonlinear association, which could not be detected using other correlation measures. In the microarray study, our proposed method detects nonlinear association between the expression levels of different genes, which could not be detected using the conventional linear methods.

e-mail: Ann.Chen@moffitt.org

## ANALYSIS FOR TEMPORAL GENE EXPRESSIONS UNDER MULTIPLE BIOLOGICAL CONDITIONS

Hong-Bin Fang\*, University of Maryland Greenebaum Cancer Center  
Dianliang Deng, University of Regina-Canada  
Jiuzhou Song, University of Maryland  
Ming Tan, University of Maryland Greenebaum Cancer Center

Temporal gene expression data are of particular interest to researchers as it contains rich information in characterization of gene function and have been widely used in biomedical studies and cancer early detection. However, the current temporal gene expressions usually have few measuring time series levels, extracting information and identifying efficient treatment effects without loss temporal information are still in problem. A dense temporal gene expression data in bacteria shows that the gene expression has various patterns under different biological conditions. Instead of analysis of gene expression levels, we consider the relative change-rates of gene in the observation period in this paper. We propose a semi-parametric model to characterize the relative change-rates of genes, in which individual expression trajectory is modeled as longitudinal data with changeable variance and covariance structure. Then, based on the parameter estimates, a chi-square test is proposed to test the equality of gene expressions. Furthermore, the Mahalanobis distance is used for the classification of genes. The proposed methods are applied to the dataset of 32 genes in *P. aeruginosa* expressed in 39 biological conditions. The simulation studies show that our methods are well performance for analysis of temporal gene expressions.

e-mail: hfang@som.umaryland.edu

## QUERY LARGE SCALE MICROARRAY COMPENDIUM DATASETS USING A MODEL-BASED BAYESIAN APPROACH WITH VARIABLE SELECTION

Ming Hu\*, School of Public Health, University of Michigan  
Zhaohui Qin, School of Public Health, University of Michigan

In microarray gene expression data analysis, it is often of interest to identify genes that share similar expression profiles with a particular gene such as a key regulatory protein. Multiple studies have been conducted using various correlation measures to identify co-expressed genes. While working well for small datasets, the heterogeneity introduced from increased sample size inevitably reduces the sensitivity and specificity of these approaches. We develop a model-based gene expression query algorithm built under the Bayesian model selection framework. It is capable of detecting co-expression profiles under a subset of samples/experimental conditions. In addition, it allows linearly transformed expression patterns to be recognized and is robust against sporadic outliers in the data. Both features are critically important for increasing the power of identifying co-expressed genes in large scale gene expression datasets. Our simulation studies suggest that this method outperforms existing correlation coefficients or mutual information-based query tools. When we apply this new method to the Escherichia coli compendium data, it identifies a majority of known regulons as well as novel potential target genes of numerous key transcription factors.

e-mail: hming@umich.edu

## MODELLING THREE DIMENSIONAL CHROMOSOME STRUCTURES USING GENE EXPRESSION DATA

Guanghua Xiao, University of Texas Southwestern Medical Center  
Xinlei Wang\*, Southern Methodist University  
Arkady Khodursky, University of Minnesota

Recently, many genomic studies have shown that significant chromosomal spatial correlation exists in gene expression of many organisms. Co-expression has been frequently observed among genes that are far apart along a chromosome chain, but brought into physical proximity by three-dimensional chromosome structures. Ignoring such correlation in statistical modeling can greatly reduce the efficiency of estimation and the power of statistical inference. Further, modeling the spatial correlation explicitly will be extremely useful for biologists to identify co-regulated genes and understand the underlying transcriptional regulation mechanism. In this paper, we construct a mathematical model for a spiral/helix-like folding structure suggested by several biological studies, and propose a statistical method to incorporate the induced correlation structure. The proposed method, can improve the estimation of gene expression. More importantly, it will be the first to model and infer the local 3D chromosome structure, and directly test its role in gene regulation.

e-mail: swang@smu.edu

## ORDER REVERSAL DETECTION (ORD) FOR ANALYSIS OF SPLICE-JUNCTION MICROARRAYS

Jonathan A. Gelfond\*, University of Texas Health Science Center, San Antonio

Luiz Penalva, University of Texas Health Science Center, San Antonio

Alternative splicing is a substantial source for expanding the functional capacity of the genome. Splice junction microarrays allow the interrogation of thousands of splice junctions, but the data analysis is difficult and error prone because of the increased complexity relative to conventional gene expressional analysis. We present Order Reversal Detection (ORD) as a method for identifying alternative splicing events based upon a straightforward probabilistic model comparing the over or underrepresentation of two or more competing isoforms. ORD has advantages over commonly used methods it has resistance to false positive errors due to nonlinear trends in microarray measurements. Further, ORD does not depend on prior knowledge of splicing isoforms. We also provide a means of visualizing the data, and a means of matching of alternative splicing patterns between different tissue types. The example data is from different cell lines of glioblastoma tumors assayed with JIVAN microarrays.

e-mail: gelfondjal@uthscsa.edu

## ANALYSIS OF CANCER-RELATED EPIGENETIC CHANGES IN DNA TANDEM REPEATS

Michelle R. Lacey\*, Tulane University  
Koji Tsumagari, Hayward Genetics Center, Tulane University School of Medicine  
Melanie Ehrlich, Hayward Genetics Center, Tulane University School of Medicine

DNA methylation is essential for the normal development and functioning of organisms, and frequent abnormal increases or decreases in DNA methylation tags are found in most human cancers and contribute to their development. Through hairpin bisulfite technology and Southern blots, the methylation characteristics of genomic regions can be studied on both a local and regional scale. We analyze the tandem repeats Sat2 and NBL2, revealing considerable differences in methylation patterns between somatic control tissues and ovarian carcinomas, and present a stochastic model for these epigenetic changes which reflects the site-to-site dependencies evident in the observed data.

e-mail: mlacey@math.tulane.edu

## 101. NONPARAMETRIC METHODS

### RANK INFERENCE FOR VARYING COEFFICIENT MODELS

Lan Wang, University of Minnesota  
Bo Kai\*, Penn State University  
Runze Li, Penn State University

Wilcoxon rank regression is a well developed technique in classical linear models, but less so in nonparametric regression models. Varying

# ABSTRACTS

coefficient models have been demonstrated to be very popular models to add the flexibility without incurring the curse of dimensionality. The goal of this work is to extend rank regression into varying coefficient models and develop nonparametric inferences for rank-based method. We have developed new estimation procedures and derived the asymptotic normality results. Our simulation shows that the newly proposed methods compare favorably with traditional techniques based on least squares.

e-mail: bokai@psu.edu

## COMPARISON OF TREATMENT EFFECTS-AN EMPIRICAL LIKELIHOOD BASED METHOD

Haiyan Su\*, University of Rochester  
Hua Liang, University of Rochester

To compare two treatment effects, which can be described as the difference of the parameters in two linear models, we propose an empirical likelihood based method to make inference for the difference. Our method is free of the assumptions of normally distributed and homogeneous errors, and equal sample sizes. The empirical likelihood ratio for the difference of the parameters of interest is shown to be asymptotically chi-squared. Simulation experiments illustrate that our method outperforms the published ones. Our method is used to analyze a data set from a drug study.

e-mail: Haiyan\_Su@urmc.rochester.edu

## A MULTIVARIATE LIKELIHOOD-TUNED DENSITY ESTIMATOR

Yejin Chung\*, Penn State University  
Bruce G. Lindsay, Penn State University

We consider an improved multivariate nonparametric density estimator which arises from treating the kernel density estimator as an element of the model that consists of all mixtures of the kernel, continuous or discrete. One can obtain the kernel density estimator with likelihood-tuning by using the uniform density as the starting value in an EM algorithm. The second tuning leads to a fitted density with higher likelihood than the kernel density estimator. In the univariate case, the two-step likelihood-tuned density estimator reduces asymptotic bias and performs robustly against a type of the true density. In addition, starting EM from a normal density centered at the sample mean, the esitmator is more robust against outliers than the estimator using the uniform initial value. We compare the performance of the new density estimator with other modified density estimators in higher dimensions.

e-mail: ychung@psu.edu

## SPLINE-BASED SIEVE SEMIPARAMETRIC GENERALIZED ESTIMATING EQUATION METHOD

Lei Hua\*, University of Iowa  
Ying Zhang, University of Iowa

We propose to analyze panel count data using a spline-based sieve semiparametric generalized estimating equation method with a semiparametric proportional mean model. The baseline mean function is approximated by monotone cubic B-spline functions. The estimates of regression parameters and spline coefficients are the roots of generalized estimating equations (sieve GEE) and computed by the generalized Rosen algorithm utilized in Zhang and Jamshidian (2004). The proposed method avoids assuming the parametric structure of the mean function and the underlying counting process. Selection of an appropriate covariance matrix that accounts for the over-dispersion and autocorrelation generally improves estimation efficiency. The asymptotic variances of the sieve semiparametric GEE estimates can be estimated using a sandwich formula and the semiparametric inference about the unknown parameters is robust to the misspecification of the covariance matrix. Simulation studies are conducted to investigate the finite sample performance of the sieve semiparametric GEE estimates with different sample sizes. Finally, the proposed method with different covariance matrices is applied to a real data from a bladder tumor clinical trial.

e-mail: lei-hua@uiowa.edu

## DIMENSION REDUCTION FOR NON-ELLIPTICALLY DISTRIBUTED PREDICTORS: SECOND-ORDER METHODS

Yuexiao Dong\*, Penn State University  
Bing Li, Penn State University

Many classical dimension reduction methods --- especially those based on inverse conditional moments --- require the predictors to have elliptical distributions, or at least to satisfy a linearity condition. Such conditions, however, are too strong for some applications. Li and Dong (2008) introduced the notion of the central solution space and used it to modify the first-order methods, such as Sliced Inverse Regression, so that they no longer rely on these conditions. In this paper we generalize this idea to the second-order methods, such as Sliced Average Variance Estimator and Directional Regression. In doing so we demonstrate that the central solution space is a versatile framework: we can use it to modify essentially all inverse conditional moment based methods to relax the distributional assumption on the predictors. Simulation studies and an application show a substantial improvement of the modified methods over their classical counterparts.

e-mail: yud101@psu.edu

## RANK-BASED SIMILARITY METRIC WITH TOLERANCE

Aixiang Jiang\*, Vanderbilt University  
Yu Shyr, Vanderbilt University

Many similarity metrics are available to measure the degree of similarity of two variables. Multivariate analyses such as dimension reduction, clustering, factor analysis, and classification all rely on similarity metrics or relative measurement distances. This implies that similarity metrics are crucial in high-dimensional data analysis. Among available similarity metrics, the rank-based Spearman correlation coefficient is the most robust, with the ability to overcome outlier or leverage effect. However, ranks themselves might artificially expand the real difference when two numbers are actually very close. To keep the robustness of the Spearman correlation coefficient, but make some correction to its drawback and also provide researchers with a method for calculating loose correlation coefficient with customized tolerance, we created a new similarity metric based on ranks with a tolerance. Both simulated and real data are used to illustrate our new metric.

e-mail: aixiang.jiang@vanderbilt.edu

## CLUSTERING TECHNIQUES FOR HISTOGRAM-VALUED DATA

Jaejik Kim\*, University of Georgia  
Lynne Billard, University of Georgia

Contemporary datasets are becoming increasingly larger and more complex, while techniques to analyse them are becoming more and more inadequate. Thus, new methods are needed to handle these new types of data. This paper introduces methods to cluster histogram-valued observations. First, three new dissimilarity measures for histogram data are proposed. Then, a polythetic clustering algorithm is developed (based on all p variables). Validity criteria to aid in the selection of the optimal number of clusters are described. The new methodology is illustrated on a large dataset collected from the US Forestry Service.

e-mail: jjkim@uga.edu

## 102. BAYESIAN SPATIAL/TEMPORAL MODELING

### BAYESIAN MODELLING OF WIND FIELDS USING SURFACE DATA COLLECTED OVER LAND

Margaret Short\*, University of Alaska Fairbanks  
Javier Fochesatto, University of Alaska Fairbanks

We propose an approach to modeling wind fields in which the error structure includes the type of instrumentation used to collect such data over land, namely anemometers (for wind speed) and vanes (for wind direction). Thus the model can handle both the periodicity of the wind direction and the non-negativity of the wind speed. Measurement error depends in part on the wind speed and is incorporated in the model using a circular distribution. We use a

Bayesian approach and fit the models using Markov chain Monte Carlo. Model performance is illustrated with a Swiss surface wind data set.

e-mail: ENAR@mshort.authorized.yon.net

### BAYESIAN NON-PARAMETRIC APPROACHES FOR DETECTING ABRUPT CHANGES IN DISEASE MAPS

Pei Li\*, University of Minnesota  
Sudipto Banerjee, University of Minnesota  
Timothy E. Hanson, University of Minnesota  
Alexander M. McBean, University of Minnesota

In many applications involving geographically referenced data in public health, investigators want to understand the underlying mechanisms causing disparities in the outcome variables. Statistical methods correctly accounting for uncertainty at various levels to elicit ‘boundaries’ or ‘zones’ of rapid change help epidemiologists and policy-makers better understand the factors driving these disparities. Such boundaries or zones often occur due to lurking spatial variables representing local disparities, such as those in income or access to health care. This talk will discuss Bayesian non-parametric approaches using areally dependent stick-breaking priors to achieve fully model-based inference on regions to reveal clusters or boundaries that suggest hidden risk factors. In particular, we concentrate on models that embed univariate and multivariate Markov random fields within a non-parametric framework. We illustrate with simulated data as well as Pneumonia and Influenza hospitalization data from the SEER-Medicare program in Minnesota.

e-mail: lixx0525@umn.edu

### A MARGINALIZED ZERO ALTERED MODEL FOR SPATIALLY CORRELATED COUNTS WITH EXCESSIVE ZEROS

Loni P. Philip\*, Harvard School of Public Health  
Brent Coull, Harvard School of Public Health

Excessive zero counts occur if the number of observed zeros exceeds that expected under the assumption of a traditional count distribution, such as a Poisson or Negative Binomial model. In health disparities research focusing on the spatial patterning of disease counts and its relationship with socioeconomic factors, spatial correlation among the counts is also typically present. Therefore, proper statistical analyses must account for the complex distribution of the disease counts as well as the correlation that exists among observations. We propose a marginalized zero-altered Poisson (ZAP) model that has the advantage of specifying an interpretable marginal relationship between the outcome and the covariates of interest in the presence of zero inflation or deflation. We describe a Bayesian approach to model fitting, and compare the approach to existing statistical approaches to analyzing spatially correlated zero-altered count data. We apply the model to motivating data on the association between premature mortality rates and socioeconomic status in Boston, MA during 1999 to 2001.

e-mail: lphilip@hsph.harvard.edu

# ABSTRACTS

## SPACE-TIME DIRICHLET PROCESS MIXTURE MODELS FOR SMALL AREA DISEASE RISK ESTIMATION

M.D. M. Hossain\*, Biostatistics/Epidemiology/Research Design (BERD) Core Center for Clinical and Translational Sciences, The University of Texas Health Science Center at Houston  
Andrew B. Lawson, Medical University of South Carolina

We propose a space-time Dirichlet process mixture model. The dependencies for spatial and temporal effects are introduced by using space-time dependent kernel stick-breaking processes. We compared this model with the space-time standard random effect model by checking each model's ability in terms of cluster detection of various shapes and sizes. This comparison was made for real and simulated data. For real data, we used twelve years of Georgia throat cancer mortality data. For simulated data, we used Ohio Geographies and twenty-one years of expected lung cancer cases.

e-mail: monir.hossain@uth.tmc.edu

## ASYMPTOTIC COMPARISON OF PREDICTIVE DENSITIES FOR DEPENDENT OBSERVATIONS

Xuanyao He\*, University of North Carolina at Chapel Hill  
Richard Smith, University of North Carolina at Chapel Hill  
Zhengyuan Zhu, University of North Carolina at Chapel Hill

This paper studies Bayesian predictive densities based on different priors and frequentist plug-in type predictive densities when the predicted variables are dependent on the observations. Average Kullback-Leibler divergence to the true predictive density is used to measure the performance of different inference procedures. The notion of second-order KL dominance is introduced, and an explicit condition for a prior to be second-order KL dominant is given using an asymptotic expansion. As an example, we show theoretically that for mixed effects models, the Bayesian predictive density with a prior from a particular improper prior family dominates the performance of REML plug-in density, while the Jeffreys prior is not always superior to the REML approach. Simulation studies are included which show good agreement with the asymptotic results for moderate sample sizes.

e-mail: xoyo@unc.edu

## ZERO-INFLATED BAYESIAN SPATIAL MODELS WITH REPEATED MEASUREMENTS

Jing Zhang\*, Miami University  
Chong Z. He, University of Missouri-Columbia

Zero-inflated data arises in many contexts. In this paper, we develop a Bayesian hierarchical model which deals with the spatial effects, correlation between repeated measurements as well as the excess zeros simultaneously. Inference, including the simulation from the posterior distributions, predictions on new locations, and hypothesis testing on the model parameters, is carried out by computationally efficient MCMC techniques. The posterior distributions are simulated using a

Gibbs sampler with embedded ratio-of-uniform method and the slice sampling algorithm. The approach is illustrated via the application to the herbaceous data collected in the Missouri Ozark Forest Ecosystem Project.

e-mail: zhangj8@muohio.edu

## BAYESIAN MODELING FOR NONSTATIONARY MULTIVARIATE SPATIAL PROCESSES

Anandamayee Majumdar\*, Arizona State University  
Debashis Paul, University of California-Davis  
Dianne Bautista, The Ohio State University

We propose a flexible class of nonstationary stochastic models for multivariate spatial data. The method is based on convolutions of spatially varying covariance kernels and produces mathematically valid covariance structures. This method generalizes the convolution approach suggested by Majumdar and Gelfand (2007) to extend multivariate spatial covariance functions to the nonstationary case. A Bayesian method for estimation of the parameters in the covariance model based on a Gibbs sampler is proposed, and applied to simulated data. Model comparison is performed with the coregionalization model of Wackernagel (2003) which uses a stationary bivariate model. Based on posterior prediction results, the performance of our model is seen to be considerably better.

e-mail: ananda@math.asu.edu

## 103. MISSING VALUES IN SURVIVAL AND/OR LONGITUDINAL DATA

### EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR THE RATIO AND DIFFERENCE OF TWO HAZARD FUNCTIONS

Yichuan Zhao\*, Georgia State University  
Meng Zhao, Georgia State University

In biomedical research and lifetime data analysis, the comparison of two hazard functions usually plays an important role. In this talk, we consider the standard two-sample framework under right censoring. We construct useful confidence intervals for the ratio and difference of two hazard functions using smoothed empirical likelihood (EL) method. The empirical log-likelihood ratio is derived and its asymptotic distribution is a chi-squared distribution. Simulation studies show that the proposed EL confidence intervals have better performance in terms of coverage accuracy and average length of confidence intervals than the traditional normal approximation method. Finally, our methods are illustrated with clinical trial data. It is concluded that the proposed EL methods provide better inferential results.

e-mail: dz2007@gmail.com

## MULTIPLE IMPUTATION BASED ON RESTRICTED MEAN MODELS FOR CENSORED SURVIVAL DATA

Lyrica Xiaohong Liu\*, University of Michigan  
Susan Murray, University of Michigan  
Alex Tsodikov, University of Michigan

Most multiple imputation methods for censored survival data either ignore patient characteristics when imputing a likely event time, or place quite restrictive modeling assumptions on the survival distributions used for imputation. In this research, we propose a multiple imputation approach that directly imputes restricted lifetimes over the study period based on a model of the mean restricted life as a linear function of covariates. This method retains patient characteristics through the model on the mean structure when making imputation choices, but does not make assumptions on the shapes of hazards or survival functions. Simulation results show that the resulting model of mean restricted life gives more precise parameter estimates than a pseudo-value approach for fitting a similar model for the restricted mean, without making additional parametric assumptions.

e-mail: lyrica@umich.edu

## NON-PARAMETRIC ESTIMATION OF A LIFETIME DISTRIBUTION WITH INCOMPLETE CENSORED DATA

Chung Chang\*, New Jersey Institute of Technology  
Wei-Yann Tsai, Mailman School of Public Health, Columbia University

In the analysis of lifetime data, under some circumstances, censoring times for unfailed units are missing or only known within an interval (e.g., warranty data). Motivated by such examples, we consider a statistical model in which censoring times are incomplete. We propose an iterative method to obtain a nonparametric estimator of the survival function and conduct a simulation study to discuss its property.

e-mail: cchang@njit.edu

## BAYESIAN MODEL AVERAGING FOR CLUSTERED DATA: IMPUTING MISSING DAILY AIR POLLUTION CONCENTRATIONS

Howard H. Chang\*, Johns Hopkins University  
Roger D. Peng, Johns Hopkins University  
Francesca Dominici, Johns Hopkins University

Missing observations are often present and imputation has become a popular approach to handle missing data. When multiple competing regression models can be used for missing data imputation, Bayesian model averaging (BMA) provides a powerful tool for missing data imputation and prediction. We develop a BMA-based missing data imputation strategy for clustered data. Our approach has the feature of allowing the weights assigned to competing models to vary between clusters while borrowing information across clusters in estimating model parameters. We first demonstrate the benefits of carrying out

our proposed cluster-specific BMA through simulation studies. We then apply the newly proposed method to a national dataset of daily ambient coarse particulate matter (PM10-2.5) concentration between 2003 and 2005. Using cross-validation, we demonstrate that cluster-specific BMA for imputing missing air pollution time series data outperforms standard approaches. Finally by using the national dataset with imputed PM10-2.5 data, we estimate the posterior probability of PM10-2.5 nonattainment status for 95 US counties based on the Environmental Protection Agency's proposed 24-hour standard.

e-mail: hhchang@jhsph.edu

## NON-IGNORABLE MODELS FOR INTERMITTENTLY MISSING CATEGORICAL LONGITUDINAL RESPONSES

Roula Tsonaka\*, Katholieke Universiteit-Leuven, Belgium  
Dimitris Rizopoulos, Erasmus Medical Center-The Netherlands  
Geert Verbeke, Katholieke Universiteit-Leuven, Belgium

The statistical literature over the last two decades has recognized that an advisable strategy for the analysis of incomplete longitudinal responses is to perform a sensitivity analysis mainly due to the fact that the observed data do not contain enough information to distinguish between competing models. The majority of the sensitivity analysis techniques has focused on continuous responses and typically involves variations to a basic model i.e., changing assumptions about the error distribution, the mean structure, etc. In this work we focus on categorical longitudinal responses that are allowed to be missing intermittently. In particular, we develop a general class of selection models that encompasses the traditional selection models (Little, JASA: 1995, pp. 1112-1121) as well as the shared parameter models (Follmann et al., Biometrics:1995, pp. 151-168). Thus, a broader sensitivity analysis, than usual, can be undertaken since different assumptions for the missing data mechanism are possible. Furthermore and in order to avoid the algebraic complexity of marginal models for categorical responses (Fitzmaurice et al., JRSSA: 2005, pp. 723-735), we consider marginalized mixed effects models based on the work of Heagerty and Zeger (2000, StatSc, pp. 1-26) in the incomplete data context.

e-mail: spyridoula.tsonaka@med.kuleuven.be

## IDENTIFICATION STRATEGIES FOR PATTERN MIXTURE MODELS WITH COVARIATES

Chenguang Wang\*, University of Florida  
Michael J. Daniels, University of Florida

Pattern mixture modeling is a popular approach for handling incomplete longitudinal data. Such models are not identifiable by construction since the distribution of the missing data is not specified given the observed data. Identifying restrictions, such as complete case missing value (CCMV) constraints or available case missing value (ACMV) constraints etc., are one approach to mixture model identification and have been well discussed in the literature (Little, 1994; Little and Wang, 1996; Molenberghs et al., 2002; Kenward et al., 2003; Daniels and Hogan, 2008). However, identification strategies can be difficult in models with covariates; in particular,

# ABSTRACTS

baseline covariates with time-invariant coefficients. An alternative identifying restriction based on residuals is proposed and connections between the proposed constraint and the common missing at random (MAR) constraint and conducting sensitivity analysis is explored. We illustrate this approach using data from a recent clinical trial.

e-mail: cgwang@cog.ufl.edu

## A COMPARISON OF IMPUTATION METHODS IN A LONGITUDINAL CLINICAL TRIAL COUNT DATA

Mohamed Alesh\*, U.S.Food and Drug Administration

Despite substantial efforts to follow each patient, missing data remains a common problem in longitudinal clinical trials. In this presentation we compare methods for handling missing data in a clinical trial with repeated evaluations for actinic keratosis lesion counts. One approach uses a predictive mean response from an extended integer valued autoregressive INAR(1) model. A second approach is based on the inverse probability weighting for handling missing data. These two methods are contrasted with last observed value carried forward (LOCF) and the complete case analyses in a simulation study. Missing data were simulated using an approach consistent with the missing data patterns found in the original data, where missingness is expected to be related to some covariates in the model. Key Words: missing data, model dependent imputation, INAR(1) model

e-mail: Mohamed.Alesh@fda.hhs.gov

## 104. META-ANALYSIS

### META ANALYSIS OF SOIL INGESTION INTAKE FOR CHILDHOOD RISK ASSESSMENT

Edward J. Stanek III\*, University of Massachusetts-Amherst  
Edward J. Calabrese, University of Massachusetts-Amherst

Casual ingestion of soil by young children frequently is the most important source of contaminant exposure when assessing risk of contaminated sites. Estimates of soil ingestion in the US are based on several mass-balance soil ingestion studies, but the methodology has the potential for bias and high variability. We develop a conceptual framework for a soil ingestion model, and use it to estimate the distribution of soil ingestion in a Monte-Carlo exposure assessment of children. The research uses data from four previously conducted mass-balance soil ingestion studies among children in the US. We describe a basic deterministic mass-balance model, and introduce random variables to translate it into a stochastic model. We present steps used to estimate reliability in the stochastic model, and based on the reliability estimates, identify data values that are considered to have high potential to include bias. We show how this process leads to identifying data for the meta analysis, and guides analysis decisions. In so doing, we discuss limitations of data, identify critical decisions and assumptions, and present the formal meta analysis plan and results.

e-mail: stanek@schoolph.umass.edu

## METHYL BROMIDE ALTERNATIVES IN HUELVA (SPAIN): A CASE OF META-ANALYSIS APPLICATION

Dihua Xu\*, University of Maryland-Baltimore County  
Bimal K. Sinha, University of Maryland-Baltimore County  
Guido Knapp, TU Dortmund University

Methyl Bromide has been widely used globally as a pre-planting soil fumigant in the horticultural industry all over the world. Due to its ozone depleting nature, effective alternatives have been sought. Several studies in Spain have been carried out with conflicting conclusion. In this talk an appropriate statistical meta-analysis has been performed to settle the issues. Some special features of the data sets have been pointed out.

e-mail: dxu1@umbc.edu

## AN ALTERNATIVE APPROACH TO META-ANALYSIS WITH CONFIDENCE

G.Y. Zou\*, University of Western Ontario

Research results in primary studies are now usually reported in the format of effect point estimates and associated confidence intervals. The purpose of this presentation is to discuss a general meta-analytic confidence interval procedure that relies only on the individual confidence intervals reported in primary studies. It is shown that the lower margin of error for the overall effect size is given by the mean root sum squares of individual lower margins, while the upper margin of error is given by the mean root sum squares of individual upper margins. Simulation results for several common effect measures demonstrate that this simple procedure performs well, even when the effect sizes are heterogeneous, the number of studies is small, and the effect size estimates have skewed distributions.

e-mail: gzou@robarts.ca

## A COMPARISON OF META-ANALYSIS METHODS FOR RARE EVENTS IN CLINICAL TRIALS

Zhiying Xu\*, Bristol-Myers Squibb Company  
Mark Donovan, Bristol-Myers Squibb Company  
David H. Henry, Bristol-Myers Squibb Company

Rare adverse events such as deaths or cardiac disorders observed in clinical trials are crucial to assess the safety of new drugs. However, in most cases, single trials are not designed to evaluate such kind of outcomes and inherently the tests are largely underpowered. Meta-analysis is commonly used to synthesize results from individual studies to support the claimed effect with improved power. When performing the meta-analysis for clinical studies, the possibility that the control group may have less duration of follow-up than the experimental group should not be ignored. In addition, time-dependent event rates can affect the ability to adjust for follow-up imbalances. Through simulation, we evaluate the performance of multiple methods with and without consideration of time-dependent factors with respect to the accuracy and precision of the estimates. Simulations are based on a diabetes project with different event rates and relative risks.

e-mail: zhiying.xu@bms.com



## RANDOM EFFECTS META-ANALYSIS WITH TWO-COMPONENT NORMAL MIXTURES

Michael P. LaValley\*, Boston University

Use of the random effects model has become widespread in meta-analysis to account for heterogeneity of study results. The random effects model allows the variation in study results to come from two sources: 1) the within study sample variance, and 2) a between study variation in the (latent) true effect. The between study variation is usually modeled using a normal distribution. In this talk, we evaluate the use of two-component mixtures of normal distributions for the between study variation. This allows greater heterogeneity in the study results and more flexibility in modeling the between study variation. In addition, this type of analysis may be useful when there is concern that the studies included in a meta-analysis come from more than one population. Simulated and real data will be used to evaluate this approach to random effects meta-analysis.

e-mail: mlava@bu.edu

## META-ANALYSES ON THE RATE OF CHANGE OVER TIME WHEN INDIVIDUAL PATIENT DATA ARE PARTLY AVAILABLE

Chengjie Xiong\*, Washington University

Longitudinal studies are often conducted to estimate and compare the longitudinal rate of changes on biological markers. Most up-to-date scientific evidences on the rate of change can only be obtained when all existing estimates from the literature and the most up-to-date individual patient longitudinal data are jointly analyzed statistically. This article provides a unified approach of meta-analysis across a group of longitudinal studies within the framework of likelihood principle. We propose a general linear mixed effects model to conduct meta-analyses when individual patient longitudinal data are only available for some of the studies and summary statistics on the rate of changes have to be used for the rest of the studies. Through an appropriate augmentation of all available data, we show that the standard statistical procedures such as PROC MIXED/SAS can be used to find the maximum likelihood estimates to the rate of change and obtain appropriate statistical inferences. We also propose measures of heterogeneity for the rate of changes based on our proposed model of meta-analyses. Finally, we demonstrate our proposed methodology through a real life example studying the longitudinal rate of cognitive changes as a function of apolipoprotein E4 genotype among elderly individuals.

e-mail: chengjie@wustl.edu

## A META-ANALYTIC FRAMEWORK FOR COMBINING INCOMPARABLE COX PROPORTIONAL HAZARD MODELS CAUSED BY OMITTING IMPORTANT COVARIATES

Xing Yuan\*, University of Pittsburgh  
Stewart Anderson, University of Pittsburgh

In Cox proportional hazard models with censored survival data, estimates of treatment effects with some important covariates omitted

will be biased toward zero (Gail et al., 1984). This is especially problematic in meta-analyses to combine estimates of parameters from studies where different covariate adjustments are made. Presently, few constructive solutions have been provided to address this issue. We propose a meta-analytic framework for combining incomparable Cox models under both aggregated patient data (APD) and individual patient data (IPD) structures. For APD, two meta-regression models with indicators of different covariates in Cox models are proposed to adjust the heterogeneity of treatment effects across studies. Both parametric and nonparametric estimators for the pooled treatment effect and the heterogeneity variance are presented and compared. For IPD, we propose a fully augmented weighted estimator based on frailty models accommodating covariate(s) omission from different studies, and results are compared with estimations from multiple imputations method. Furthermore, while most meta-analyses focus on combining univariate effect sizes, we generalize our methods to estimate the entire vector of effect sizes simultaneously. We illustrate the advantages of our proposed analytic procedures over existing methodologies by simulation studies and real data analyses using multiple breast cancer clinical trials.

e-mail: xiyl17@pitt.edu

**WEDNESDAY, MARCH 18, 2009  
10:30 AM - 12:15 PM**

## 105. INTEGRATING GENOMIC AND/OR GENETICS DATA

### INFORMATION-INTEGRATION APPROACHES TO BIOLOGICAL DISCOVERY IN HIGH-DIMENSIONAL DATA

John Quackenbush\*, Dana-Farber Cancer Institute and Harvard University

Two trends are driving innovation and discovery in biological sciences: technologies that allow holistic surveys of genes, proteins, and metabolites and a realization that biological processes are driven by complex networks of interacting biological molecules. However, there is a gap between the gene lists emerging from genome sequencing projects and the network diagrams that are essential if we are to understand the link between genotype and phenotype. 'Omic technologies were once heralded as providing a window into those networks, but so far their success has been limited. To circumvent these limitations, we developed a method that combines 'omic data with other sources of information. Here we will present a number of approaches we have developed, including an integrated database that collects clinical, research, and public domain data and synthesizes it to drive discovery and an application of seeded Bayesian Network analysis applied to gene expression data that deduces predictive models of network response. We will demonstrate some applications of this approach using a collection of genomic assays collected on a collection of samples obtained from >130 patients with stage III/IV serous ovarian cancer.

e-mail: johnq@jimmy.harvard.edu

# ABSTRACTS

## METHODS FOR IDENTIFYING THE GENETIC VARIANTS ASSOCIATED WITH HIGH-ORDER EXPRESSION MODULES

Hongzhe Li\*, University of Pennsylvania

Many recent studies have shown that the level of expression of gene can be heritable and many eQTLs have been identified by traditional linkage/assocaiton analysis. These studies mainly focus on the effects of genetic variants on the levels of expression of genes. However, genetic variants can also affect expression coordination among genes and therefore high-order expression modules. This paper proposes a regression-on-correlation (RegCor) model to link the genetic variants to expression association between two genes in order to identify the variants that are association with high-order expression modules. A score test and a regularized MLE procedure are proposed for hypothesis testing and for selection of genetic variants. Some theoretical results, simulations and application to expression data in yeast segregants will be presented.

e-mail: hongzhe@upenn.edu

## A GRAPHICAL SOLUTION FOR THE MULTIPLE TOP-K LIST PROBLEM WITH APPLICATIONS IN MOLECULAR MEDICINE

Michael G. Schimek\*, Danube University, Krems, Austria  
Eva Budinska, Masaryk University, Brno, Czech Republic

For a fixed set of objects, let us have  $L$  lists representing their individual rank positions. For many studies in molecular medicine it is typical that the probability for consensus rankings is decreasing with increasing distance from the top rank position. An important task is to consolidate such lists, i.e. to identify those objects that are characterized by rankings of high conformity across the assessments up to position  $k$ . HALL and SCHIMEK (COMPSTAT 2008 Proceedings, p.433-444) have addressed this inference problem for the case of two assessors. They could develop a moderate deviation-based procedure for random degeneration in paired rank lists. For each pairwise comparison of the  $L$  lists a specific  $k$ -value is obtained. These  $k$ -values can substantially vary because the degree of correspondence between the partial lists is in practice not high due to various irregularities of the assessments. The integration of the thus obtained results into a common set of conforming objects is highly dependent on technical assumptions, and the choices of the tuning parameters. Therefore, it would be of great value to have a graphical representation that allows us to monitor the integration process. The development and application of such a graphical tool is described in this paper.

e-mail: michael.schimek@donau-uni.ac.at

## FINITE MIXTURE OF SPARSE NORMAL LINEAR MODELS IN HIGH DIMENSIONAL FEATURE SPACE WITH APPLICATIONS TO GENOMICS DATA

Shili Lin\*, The Ohio State University

The development of modern technology has led to the generation of high dimensional data in many areas of scientific research such

as statistical genomics. In such applications, one is often interested in studying the relationship between a response variable and a high dimensional vector of features. However, it is often the case that only a small number of the features contribute to the response variable, and this leads to the problem of sparse regression modeling. Such data may also exhibit heterogeneity. As such, each homogeneous sub-population rather than the entire population may be suitably modeled by a linear regression. In such situations, finite mixture of regression (FMR) models provide a powerful tool for learning the structure of the data. In this talk, I will outline a two-step procedure for treating this problem. The first step is to reduce the high dimensionality of the original feature space to a relatively lower dimension. Then we learn about the sparse Normal finite mixture based on a FMR modeling approach. We will demonstrate this approach with an application to a genome wide association study with single nucleotide polymorphism data. This is join work with Drs. Abbas Khalili and Jiahua Chen.

e-mail: shili@stat.osu.edu

## 106. APPLICATION OF DYNAMIC TREATMENT REGIMES

### ESTIMATING THE CAUSAL EFFECT OF LOWER TIDAL VOLUME VENTILATION ON SURVIVAL IN PATIENTS WITH ACUTE LUNG INJURY

Daniel O. Scharfstein\*, Johns Hopkins Bloomberg School of Public Health

Acute lung injury (ALI) is a condition characterized by acute onset of severe hypoxemia and bilateral pulmonary infiltrates. ALI patients typically require mechanical ventilation in an intensive care unit. Low tidal volume ventilation (LTVV), a time-varying dynamic treatment regime, has been recommended as an effective ventilation strategy. This recommendation was based on the results of the ARMA study, a randomized clinical trial designed to compare low vs. high tidal volume strategies (ARDSNetwork, 2000). After publication of the trial, some critics focused on the high non-adherence rates in the LTVV arm suggesting that non-adherence occurred because treating physicians felt that deviating from the prescribed regime would improve patient outcomes. In this paper, we seek to address this controversy by estimating the survival distribution in the counterfactual setting where all patients assigned to LTVV followed the regime. Our estimation strategy is based on Robins's (1986) G-computation formula and fully Bayesian multiple imputation to handle intermittent missing data.

e-mail: dscharf@hsph.edu

### STAR\*D, DYNAMIC TREATMENT REGIMES AND MISSING DATA

Dan Lizotte, University of Michigan  
Lacey Gunter, University of Michigan  
Eric Laber, University of Michigan  
Susan Murphy\*, University of Michigan

We illustrate the combined use of Q-Learning and multiple imputation for use in constructing dynamic treatment regimes using

the clinical trial, STAR\*D (Sequenced Treatment Alternatives to Relieve Depression). In this trial each individual was rerandomized to new treatments each time the individual failed to respond to treatment. Q-Learning is a generalization of regression for use in constructing dynamic; this method is not a likelihood based method. Missing data occurs in a variety of ways in the STAR\*D data set including missing data due to study dropout and missing data due to missed clinical visits and missing data due to missed clinical assessments. We discuss open questions that arise in applying Q-Learning to multiply imputed data sets.

e-mail: samurphy@umich.edu

## A PROSTATE CANCER TRIAL WITH RE-RANDOMIZATION: HOW WE SPENT A DECADE STUDYING TWELVE DYNAMIC TREATMENT REGIMES

Peter F. Thall\*, University of Texas, M.D. Anderson Cancer Center

In 1999, I designed a clinical trial that aimed to evaluate four combination chemotherapies for advanced prostate cancer. The design included a multi-stage treatment strategy constructed by an oncologist, Randy Millikan, that re-randomized a patient to a new combination if his initial therapy failed. Consequently, each patient received one of twelve possible dynamic treatment regimes, although we were unaware of this terminology when we began the trial. In this talk, I will review the design published in Statistics in Medicine in 2000, the elementary statistical analyses of the trial results published in JNCI in 2007, our response to an interesting letter to the editor of JNCI criticizing our analyses, how the subsequent desire to account for informative drop-outs led us to refine the definition of patient outcome, and the results of a more recent analysis of the resulting, more refined data, including inverse probability of treatment weighted estimation.

e-mail: rex@mdanderson.org

## 107. DATA SHARING: AN EXAMPLE OF CONFLICTING INCENTIVES

### DATA SHARING: AN EXAMPLE OF CONFLICTING INCENTIVES

Thomas A. Louis\*, Johns Hopkins Bloomberg School of Public Health

Requiring NIH grant and contract recipients who generate or consolidate data to provide publicly available, documented datasets "within a reasonable time" entails potential benefits and drawbacks for science and policy, and poses challenges for researchers and the research enterprise. I conclude that properly implemented, on balance the policy will be beneficial, but care is needed to realize these benefits. The policy needs to specify time frames that allow researchers to understand and document their datasets and to publish principal findings. Funding and infrastructure must be in place to ensure long term access; confidentiality must be protected; in some situations user certification may be necessary. Less direct, but no less important are ensuring that the requirements do not reduce engagement by the best researchers in high profile studies and that reward systems in academe

and elsewhere respect and reflect the new environment. Potential benefits of the policy include energizing cultural and system changes such as increased standardization of data definitions and database configurations, and a movement towards reproducible research. Importantly, ready access will benefit researchers as they formulate and design new studies and will increase the quality and impact of individual studies and research syntheses.

e-mail: tlouis@jhsph.edu

## ACCESS TO DATA FROM PUBLICLY FUNDED STUDIES: OPPORTUNITIES AND UNINTENDED CONSEQUENCES

Constantine Gatsonis\*, Brown University

Recently enacted policies for making data from publicly funded studies accessible to other investigators are already having their impact. Data, such as clinical information, imaging studies, and biospecimen analysis results, are beginning to be made available to investigators outside the original study teams as well as to manufacturers of drugs and devices. The new policies may well signify a sea change in biomedical research. On the one hand, they create new possibilities for augmenting the value of studies. On the other, they create a host of scientific, regulatory, and technical problems. These problems include the risks inherent in third party analyses of data from complex clinical studies, the need to protect the privacy of participants, the lack of a regulatory framework for these data releases, and the lack of attention to resource implications. Even if particular problems can be addressed, the emphasis on quick and broad access to the data by outside investigators is likely to affect the incentives for investigators, with possibly unintended consequences. In this presentation, we will discuss the above issues using the experience of a collaborative group as an example.

e-mail: gatsonis@stat.brown.edu

## NIH MANDATE ON SHARING RESEARCH DATA: THE REGULATORY LANDSCAPE

Jane Pendergast\*, The University of Iowa

The mandated sharing of research data generated through NIH support is intended to broaden the scope of investigation of these data -- thus generating more information and expedited translation of findings into useful knowledge, products, and procedures to improve human health. Yet other federal legislation restricts the sharing of health data, with the goal of protecting the privacy of research participants and keeping their health data secure and confidential. Biostatisticians/statisticians are often responsible for the creation, documentation, maintenance, and security of such datasets and are caught in the middle of these two mandates - making data available while keeping it protected from others. Just what are the rules and investigator rights behind these two seemingly contradictory mandates? How do they impact informed consent documents? Will the benefits of data sharing outweigh the costs? This presentation will address these issues with a goal of clarifying the regulatory landscape in today's research environment.

e-mail: jane-pendergast@uiowa.edu

# ABSTRACTS

## 108. MAPPING SPATIAL DATA INTO THE FUTURE

### TESTING AND MODELING THE CROSS-COVARIANCE FUNCTIONS OF MULTIVARIATE RANDOM FIELDS

Marc G. Genton\*, Texas A&M University

There is an increasing wealth of multivariate spatial and multivariate spatio-temporal data appearing. First, we propose a methodology to evaluate the appropriateness of several types of common assumptions on cross-covariance functions in the spatiotemporal context. The methodology is based on the asymptotic joint normality of sample space-time cross-covariance estimators. Specifically, we address the assumptions of symmetry, separability and linear models of coregionalization. Second, we address the problem of constructing valid parametric cross-covariance functions. We propose a simple methodology for developing flexible, interpretable, and computationally feasible classes of cross-covariance functions in closed form. We discuss estimation of these models and perform a small simulation study to demonstrate our approach. We illustrate our methodology on a trivariate pollution dataset from California. This talk is based on joint works with Tatiyana Apanasovich, Bo Li, and Michael Sherman.

e-mail: genton@stat.tamu.edu

### ON CONSISTENT NONPARAMETRIC INTENSITY ESTIMATION FOR INHOMOGENEOUS SPATIAL POINT PROCESSES

Yongtao Guan\*, Yale University

A common nonparametric approach to estimate the intensity function of an inhomogeneous spatial point process is through kernel smoothing. When conducting the smoothing, one typically uses events only in a local set around the point of interest. The resulting estimator, however, is often inconsistent since the number of events in a fixed set is of order one for spatial point processes. In this paper, we propose a new covariate-based kernel smoothing method to estimate the intensity function. Our method defines the distance between any two points as the difference between their associated covariate values. Consequently, we determine the kernel weight for a given event of the process as a function of its new distance to the point of interest. Under some suitable conditions on the covariates and the spatial point process, we prove that our new estimator is consistent for the true intensity. To handle the situation with high-dimensional covariates, we also extend sliced inverse regression, which is a useful dimension-reduction tool in standard regression analysis, to spatial point processes. Simulations and an application to a real data example are used to demonstrate the usefulness of the proposed method.

e-mail: yongtao.guan@yale.edu

## PROBABILISTIC QUANTITATIVE PRECIPITATION FORECASTING USING A TWO-STAGE SPATIAL MODEL

Tilmann Gneiting\*, University of Washington

Veronica J. Berrocal, Duke University

Adrian E. Raftery, University of Washington

Short-range forecasts of precipitation fields are required in a wealth of agricultural, hydrological, ecological and other applications. Forecasts from numerical weather prediction models are often biased and do not provide uncertainty information. Here we present a postprocessing technique for such numerical forecasts that produces correlated probabilistic forecasts of precipitation accumulation at multiple sites simultaneously. The statistical model is a spatial version of a two-stage model that describes the distribution of precipitation with a Bernoulli-Gamma mixture. Spatial correlation is captured by assuming that two Gaussian processes drive precipitation occurrence and precipitation amount, respectively. A site-specific transformation function retains the marginal right-skewed distribution of precipitation while taking the numerical forecast into account. The two-stage spatial model was applied to forecasts of daily precipitation accumulation over the Pacific Northwest in 2004, at a prediction horizon of 48 hours. The predictive distributions from the two-stage spatial model were calibrated and sharp, and outperformed reference forecasts for spatially composite and areally averaged quantities.

e-mail: tilmann@stat.washington.edu

### THE OTHER WORLD OF LARGE SPATIAL DATA SETS

Douglas Nychka\*, National Center for Atmospheric Research

Many problems in analyzing the Earth's climate depend on large spatial data sets. These problems typically break standard and exact methods for spatial statistics and approximate approaches are needed. This talk will use a suite of regional climate simulations to illustrate some techniques for introducing sparsity into covariance models and the use of conditional simulation for inference.

e-mail: nychka@ucar.edu

## 109. STRATEGIES FOR SUCCESSFUL STATISTICAL CONSULTING/COLLABORATION IN DRUG AND VACCINE DISCOVERY AND DEVELOPMENT

### STATISTICAL CONSULTING: EARNING YOUR PLACE ON THE TEAM

Mandy L. Bergquist, GlaxoSmithKline

As pharmaceutical companies face increasing patent expirations and dramatically shrinking budgets, they expect more from every employee. Statisticians can rely on government regulation for continued employment, or they can rise to the challenge. Even in a difficult environment, statisticians can extend their influence and move into larger industry roles by becoming more than role players

on the matrix teams to which they contribute. Good technical skills are valuable, but not enough to succeed. This presentation draws on the speaker's experience as an internal statistical consultant at a major pharmaceutical company to provide practical advice for statisticians who want to earn a respected position on the team.

e-mail: mlbergquist@yahoo.com

## **FROM POPULATION TO CELL TO ANIMAL TO HUMAN**

Borko D. Jovanovic\*, Northwestern University Medical School  
Raymond C. Bergan, Northwestern University Medical School

We discuss the scientific process and the statistical consulting process related to discovery of a pathway upstream from cell detachment and invasion, consequently leading to prostate cancer metastasis. In particular, the process involved the analysis of epidemiologic data, cell line microarray data, cell motility observations data, RTPCR and Northern Blot data, mouse metastasis count data and Phase II prevention clinical trial for patients following radical prostatectomy. A single research team and a single statistician have been collaborating in this process for the last eight years.

e-mail: borko@northwestern.edu

## **THE ROLE OF THE STATISTICAL CONSULTANT IN STRATEGIC PLANNING AND DEVELOPMENT IN THE PHARMACEUTICAL AND BIOTECHNOLOGY INDUSTRIES**

Bruce E. Rodda\*, Strategic Statistical Consulting LLC and University of Texas School of Public Health

Many statistical consultants play a role in the design of clinical trials and even more participate in the analysis of these trials. However, a clinical trial is rarely performed in isolation, but is an element of a broader, more complete, corporate strategy. The statistical consultant can play an important role in formulating and developing the strategy that is necessary to provide the framework for an efficient, focused, and successful set of trials. This presentation will present and discuss opportunities available to a consulting statistician in this critically important responsibility.

e-mail: Bruce\_Rodda@msn.com

## **PRIORITIZING AREAS OF THERAPEUTIC INTEREST TO TARGET PRODUCT PROFILES: LESSONS AND EXAMPLES IN INFECTIOUS DISEASES AND VACCINE DEVELOPMENT**

Nicole C. Close\*, EmpiriStat, Inc.

A review from the statistician's role and viewpoint will be given on the statistical input provided and needed when reviewing strategies to prioritize areas of therapeutic interest for an organization. Also, a large amount of time is spent with Regulatory Professionals in developing target product profiles and often times statisticians are not

fully engaged or involved in this process. Tips will be given on how to engage your statistician in these areas as well as how the statistician must engage the organizational team. Examples are drawn from infectious disease areas and vaccine development.

e-mail: nclose@empiristat.com

## **110. CLINICAL TRIAL DESIGN**

### **SAFETY AND ACCURACY OF PHASE 1 ONCOLOGY TRIALS: TRADITIONAL METHOD VS. ROLLING 6 VS. CRM**

Arzu Onar-Thomas\*, St. Jude Children's Research Hospital  
Zang Xiong, St. Jude Children's Research Hospital

Various dose-finding designs are available for Phase I pediatric oncology trials and recently the list was augmented by one with the addition of the Rolling 6 Design. The newcomer is quite similar to the empirically-based Traditional Design (also known as the 3+3 or up-and-down method), with the exception that it allows cohorts of up to 6 patients to be registered at one dose level. The intent is to decrease the duration of the trial and limited simulations indicate that Rolling 6 can achieve this goal without notable increase in toxicity. In this talk we will present extensive simulation results which compare the performance of the Rolling 6 design to that of the Traditional Method as well as to results from a frequentist version of the continuous reassessment method. The comparisons are based on accuracy, sample size and toxicity associated with each approach. The advantages/disadvantages of using each design will be highlighted within the context of challenges unique to pediatric oncology trials such as BSA based dosing.

e-mail: arzu.onar@stjude.org

### **DOSE-ESTABLISHMENT DESIGNS FOR PHASE I/II CANCER IMMUNOTHERAPY TRIALS**

Karen Messer\*, Department of Medicine and Moores University of California at San Diego Cancer Center  
Loki Natarajan, Department of Medicine and Moores University of California at San Diego Cancer Center  
Edward Ball, Department of Medicine and Moores University of California at San Diego Cancer Center  
Thomas Lane, Department of Medicine and Moores University of California at San Diego Cancer Center

A standard Phase I oncology trial is designed to find the maximum tolerated dose in a setting where serious drug-related toxicity is expected. However, agents investigated for cancer immunotherapy may hope to show efficacy without increasing the rate of adverse events. Our Phase I/II dose-establishment design is suitable when the therapeutic dose is expected to be well below the maximum tolerated dose. This design incorporates a likelihood ratio test of a safety hypothesis in Phase I, using a standard 3+3 group sequential enrollment scheme. Phase I serves as an interim safety analysis before proceeding to Phase II. The Phase I/II data are combined for an upper confidence limit on the toxicity rate at the therapeutic dose. We give an example in a Phase I/II trial of immunotherapy in leukemia.

# ABSTRACTS

The number of patients enrolled in Phases I and II depends on the sequential toxicity outcomes of the trial, and both Phases contribute to estimating the toxicity rate and efficacy outcome at the target dose. We show how to compute the power, expected sample size, and expected number of dose limiting toxicities, as well as the MLE and exact small sample confidence intervals for the toxicity rate at the therapeutic dose.

e-mail: kmesser@ucsd.edu

## AN ITERATIVE PHASE I/II CLINICAL TRIAL DESIGN INCORPORATING GENOMIC BIOMARKERS

Ashok Krishnamurthy\*, School of Public Health and Information Sciences, University of Louisville

The goal of a Phase I clinical trial is to determine the maximum tolerated dose (MTD) that corresponds to some given acceptable level of toxicity known as a dose-limiting toxicity. Effectiveness of the MTD is tested in a Phase II trial, where the purpose is to find the minimum effective dose (MED). The designs fall into two classes: Nonparametric rule-based designs (ex: 3 + 3 ) and the Bayesian model-guided designs (ex: CRM). A well known limitation to the generalizability of Phase I clinical trial is the heterogeneity of the patient population. Ignoring the differences in subgroups and running a single Phase I clinical trial would result in an inaccurate estimate of the MTD. We propose an iterative Phase I/II design that incorporates genomic biomarker information to classify patients as Biomarker Positive (B+) or Negative (B-). We present and examine a method for obtaining separate MTD estimates for subgroups (B+/B-). Performance is evaluated by considering three possible MTD re-estimation techniques: the logistic regression; the isotonic regression and the constrained logistic regression. The design then branches into two secondary Phase I trials to obtain a redefined MTD for subgroups. We can then conclude a drug may be appropriate for a defined subgroup.

e-mail: a0kris04@louisville.edu

## REINFORCEMENT LEARNING DESIGN FOR CANCER CLINICAL TRIALS

Yufan Zhao\*, University of North Carolina at Chapel Hill  
Michael R. Kosorok, University of North Carolina at Chapel Hill  
Donglin Zeng, University of North Carolina at Chapel Hill

We develop reinforcement learning trials for discovering individualized treatment regimens for life-threatening diseases such as cancer. A temporal-difference learning method called Q-learning is utilized which involves learning an optimal policy from a single training set of finite longitudinal patient trajectories. Approximating the Q-function with time-indexed parameters can be achieved by using support vector regressions or extremely randomized trees. Within this framework, we demonstrate that the procedure can extract optimal strategies directly from clinical data without relying on the identification of any accurate mathematical models, unlike approaches based on adaptive design. We show that reinforcement learning has tremendous potential in clinical research because it can select actions that improve outcomes by taking into account delayed effects even when the relationship between

actions and outcomes is not fully known. To support our claims, the methodology's practical utility is illustrated in a simulation analysis. For future research, we will apply this general strategy to studying and identifying new treatments for advanced metastatic stage IIIB/IV non-small cell lung cancer, which usually includes multiple lines of chemotherapy treatment.

e-mail: yzhao@bios.unc.edu

## PROPENSITY SCORE MATCHING IN RANDOMIZED CLINICAL TRIAL

Zhenzhen Xu\*, University of Michigan  
John D. Kalbfleisch, University of Michigan

Cluster randomization trials with relatively few clusters have been widely used in recent years for evaluation of health care strategies. On average, randomized treatment assignment achieves balance in both known and unknown confounding factors between treatment groups, however, in practice investigators can only introduce a small amount of stratification and cannot balance on all the important variables simultaneously. The limitation arises especially when there are many confounding variables and in small studies. Such is the case in the INSTINCT trial designed to investigate the effectiveness of an education program in enhancing the tPA use in stroke patients. In this paper, we introduce a new randomization design, the balance match weighted (BMW) design, which applies the optimal matching with constraints technique to a prospective randomized design and aims of to minimize the mean squared error of the treatment effect estimator. When true confounding effects are known, we construct the BMW design to produce treatment effect estimators with minimal MSE; these results suggest that, even when the confounding effects are unknown, the BMW design with appropriately chosen parameters can generate treatment effect estimators with substantially improved MSE properties. The simulation study shows that, under various confounding scenarios, the BMW design can reduce the MSE for the treatment estimator by 10% to as much as 80% compared to a completely randomized or matched-pair design. We illustrate these methods in proposing a design for the INSTINCT trial.

e-mail: zzxu@umich.edu

## EVALUATING PROBABILITY OF SUCCESS FOR CLINICAL TRIALS WITH TIME-TO-EVENT ENDPOINTS

Di Li\*, Abbott Laboratories  
Todd A. Busman, Abbott Laboratories  
Martin S. King, Abbott Laboratories

Traditional clinical trial design and sample size planning is usually based on achieving a targeted statistical power to detect a specified treatment effect. Since statistical power is conditional on the specified treatment effect, trial planning based on statistical power leaves out the uncertainty of the estimate of the true underlying treatment effect. A high statistical power does not guarantee a high probability of achieving the desired outcome. The probability of success is the unconditional probability of a successful trial. In simple cases, it can be calculated as the expected power averaged over some prior

distribution for the unknown true treatment effect. The primary efficacy endpoint in oncology clinical trials is often a time-to-event outcome, e.g., overall survival or time to disease progression. This presentation illustrates evaluation of the probability of success in trial planning with time-to-event endpoints through simulations in a Bayesian framework.

e-mail: di.li@abbott.com

## DESIGN OF PREDICTIVE POWER METHOD WITH TWO ENDPOINTS

Kiya R. Hamilton\*, University of Alabama at Birmingham  
Leslie A. McClure, University of Alabama at Birmingham

Stochastic curtailment methods, such as conditional power and predictive power, can be used to assess the chance that a significant difference between two treatment groups may occur at the end of the study, given the data collected up to an interim time point. However, stochastic curtailment methods do not currently allow for simultaneously testing more than one endpoint at a time. We extend the predictive power procedure to include two endpoints. Our focus is on the case where the two endpoints are independent. We will describe the methodology that has been developed to assess the predictive power for two independent outcomes, and how it may be extended to the case where the correlation is non-zero. We will compare our method with other group sequential methods, by assessing the average sample size and whether a 'correct decision' was made by the end of the study. We will also discuss some of the issues that may arise when monitoring multiple endpoints simultaneously. This approach is illustrated with real data from the Nitric Oxide Trial.

e-mail: krhamilton\_7@hotmail.com

## 111. GENETIC STUDIES WITH RELATED INDIVIDUALS

### ANALYSIS OF TWIN DATA USING SAS

Rui Feng\*, University of Alabama at Birmingham  
Gongfu Zhou, Yale University  
Meizhuo Zhang, Yale University  
Heping Zhang, Yale University

Twin studies are essential for assessing disease inheritance. Data generated from twin studies are traditionally analyzed using specialized computational programs. For many researchers, especially those who are new to twin studies, understanding and using those specialized computational programs can be a daunting task. Given that SAS is the most popular software for statistical analysis, we suggest the use of SAS procedures for twin data may be a helpful alternative and demonstrate that we can obtain similar results from SAS to those produced by specialized computational programs. This numerical validation is practically useful, because a natural concern with general statistical software is whether it can deal with data that are generated from special study designs such as twin studies and whether it can test a particular hypothesis. We conclude through our extensive simulation that SAS procedures can be used easily as a very convenient alternative to specialized programs for twin data analysis.

e-mail: rfeng@ms.soph.uab.edu

## HAPLOTYPING INHERITED HUMAN DISEASES WITH A FAMILY-BASED DESIGN

Qin Li\*, University of Florida  
Arthur Berg, University of Florida  
Rongling Wu, University of Florida and Penn State University

The understanding of the genetic etiology of inherited diseases is one of the main focuses in medical genetic research. Recent genetic analyses suggest that genes may trigger their effects on inherited diseases through different expression of haplotypes constructed by alleles at multiple loci. We derive a statistical model for characterizing risk haplotypes associated with an inherited disease. The model is founded on a family-based design composed of the father, the mother, and their offspring. Each member in a family is genotyped for a panel of loci, although the disease can be phenotyped only for the offspring. A two-level hierarchical likelihood is formulated to estimate population genetic parameters with parental information and quantitative genetic parameters (including haplotype effects) and the recombination fraction with offspring data. A complex structure of the EM algorithm is derived, providing precise and efficient estimates of all the underlying parameters. Simulation studies are performed to test the statistical behavior of the model under different sampling strategies (few families vs. large size or many families vs. small size), different heritabilities and a range of genetic parameters. The new model will provide a timely tool for detecting risk haplotypes for inherited diseases from an increasing amount of family-based data.

e-mail: qli@biostat.ufl.edu

## IMPUTING MISSING DATA IN CASE-PARENT TRIAD STUDIES

Tracy L. Bergemann\*, University of Minnesota

A commonly used design in genetic association studies is the case-parent triad design. Generally, samples are drawn from an affected offspring, manifesting a disease or phenotype of interest, as well as from the parents. The trio genotypes may be analyzed using a variety of available methods, but we focus on log-linear models because they test for genetic association and additionally estimate the relative risks of transmission. The models need to be modified to impute missing genotypes. Furthermore, instability in the parameter estimates can arise when certain kinds of genotype combinations do not appear in the dataset. In this research, we kill two birds with one stone. We propose a new method to simultaneously impute missing genotype data and account for genotype combinations with zero counts. This approach solves a zero-inflated Poisson (ZIP) regression likelihood. The maximum likelihood estimates yield relative risks and a likelihood ratio test determines the significance of genetic association. We compared the ZIP regression approach to previously proposed methods in both simulation studies and a dataset investigating the risk of orofacial clefts. The ZIP likelihood estimates relative risks with less bias than other methods. Further, the new method preserves the appropriate type I error rate more carefully.

e-mail: berge319@umn.edu

# ABSTRACTS

## EVALUATING THE HETEROGENEITY OF POLYGENIC VARIANCE COMPONENT BY SEX AND AGE ON CARDIOVASCULAR RISK FACTORS IN BRAZILIAN FAMILIES

Suely R. Giolo\*, Federal University of Parana, Brazil and Heart Institute, University of Sao Paulo, Brazil  
Julia M. Soler, University of Sao Paulo, Brazil  
Alexandre C. Pereira, Heart Institute, University of Sao Paulo, Brazil  
Mariza de Andrade, Mayo Clinic  
Jose E. Krieger, Heart Institute, University of Sao Paulo, Brazil

In family studies it is important to evaluate the impact of genes and environmental factors on traits of interest as well as to investigate genes that can explain differences and similarities between individuals, particularly when comparing heritability between subgroups of individuals such as young and old or males and females. In order to evaluate the evidence for heterogeneity in genetic and environmental sources of variance in males and females and also in young and old individuals on six quantitative cardiovascular risk factors (diastolic and systolic blood pressure, LDL and HDL-cholesterol, fasting blood glucose and triglycerides). We used variance components models allowing for heterogeneity and three survival censored traits (age of diagnosis of hypertension, diabetes and cholesterol) using the random-effects Cox proportional hazards model. We used information of 81 families, involving 1,675 people of the Baependi family heart study. We found evidence of sex and/or age differences in variance components for some of the traits analyzed. In some cases such differences affected the genetic variance, in others the environmental variance, and in other cases both of these variances. For visualizing such sources of heterogeneity we will present some useful graphics.

e-mail: mandrade@mayo.edu

## A REGULARIZED REGRESSION APPROACH FOR DISSECTING GENETIC CONFLICTS THAT INCREASE DISEASE RISK IN PREGNANCY

Yuehua Cui, Michigan State University  
Shaoyu Li\*, Michigan State University

Many human diseases developed during pregnancy could be caused by the direct effects of both maternal and fetal genes, and/or by the indirect effects due to genetic conflicts. Genetic conflicts exist when the effects of fetal genes are opposed by the effects of maternal genes, or when there exists conflict between the maternal and paternal genes within the fetal genome. Dissecting genetic conflict effects that increase disease risk during pregnancy presents statistical challenges. In this talk, we consider a unified framework to model and test the genetic conflicts via a regularized regression approach. Our model is developed considering real situation in which the paternal information is often completely missing, an assumption that fails most current family-based study. A mixture model based penalized logistic regression is proposed for data sampled from a natural population. We develop a variable selection procedure to select significant genetic features. Simulation studies show that the model has high power and good false positive control under reasonable sample size and disease allele frequency. Our model provides a powerful tool for dissecting genetic conflicts that increase disease risk during pregnancy.

e-mail: lishaoyu@stt.msu.edu

## LINKAGE MAP CONSTRUCTION IN INTEGRATED CROSSES

Emma Huang\*, CSIRO Mathematical and Information Sciences  
Andrew George, CSIRO Mathematical and Information Sciences

The integrated cross is an exciting new experimental design which enables genomic regions housing genes of commercial significance to be detected with far greater precision than previously possible. By using four or eight parents rather than a traditional biparental design and breeding generations through to fixation, these crosses represent an abundance of genetic diversity with the potential for high mapping resolution. CSIRO is currently conducting the world's first integrated cross in wheat. This promises to be instrumental in unlocking the genetic secrets of agronomic, disease and quality traits of one of the world's most important domesticated crops. While there are similarities between this project and the Collaborative Cross in mice, a key difference is in the lack of physical maps or genome sequence for wheat. Thus a crucial element for the success of the project is the production of highly accurate DNA marker maps. We have developed statistical methods and computational tools to address this challenge. We use two- and three-point haplotype probabilities to group and order loci within linkage groups. These algorithms and software have been tested through extensive simulations and will be applied to 4-way and 8-way cross data as it becomes available.

e-mail: b.emma.huang@gmail.com

## TESTING FOR FAMILIAL AGGREGATION OF FUNCTIONAL TRAITS

Yixin Fang\*, Georgia State University  
Yuanjia Wang, Columbia University

In genetic epidemiology, the first and foremost task is testing for familial aggregation; if no familial aggregation is found, it is unnecessary to conduct further genetic analysis. For functional traits, we propose a test statistic, which is actually the leading functional principal component of heritability. The p-value can be obtained by a permutation procedure or a theorem of Johnstone and Forrester (2004). The methods are applied to the cholesterol data from Framingham Heart Study.

e-mail: matyxf@langate.gsu.edu



## **112. VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA**

### **PAIRWISE VARIABLE SELECTION FOR HIGH-DIMENSIONAL MODEL-BASED CLUSTERING AND ITS APPLICATION TO MICROARRAY DATA**

Jian Guo\*, University of Michigan  
Elizaveta Levina, University of Michigan  
George Michailidis, University of Michigan  
Ji Zhu, University of Michigan

Gene (variable) selection for high-dimensional model-based clustering is an important yet challenging problem in microarray analysis. Existing variable selection methods for model-based clustering select informative variables in an “one-in-all-out” manner; that is, a variable is selected if at least one pair of clusters are separable by this variable and is removed if all clusters are nonseparable by this variable. In many situations, however, biologists are also interested in knowing which clusters are separable and which clusters are nonseparable for each informative variable. To address this problem, we propose a pairwise variable selection method for high-dimensional model-based clustering. We provide some evidence that our new method performs better than the  $\$l_1\$$ -norm approach and offers better interpretation.

e-mail: guojian@umich.edu

### **FAST FSR VARIABLE SELECTION WITH INTERACTIONS**

Dennis D. Boos\*, North Carolina State University  
Hugh B. Crews, SAS Institute Inc.  
Leonard A. Stefanski, North Carolina State University

The Fast FSR approach of Boos et al. (2009, Biometrics) is extended to handle forward selection for second-order models under various hierarchy restrictions between main effects and second-order terms (squares and interactions). New easy-to-use SAS macros are presented that implement these FSR variable selection approaches for linear regression, logistic regression, and Cox regression, and are available at <http://www4.stat.ncsu.edu/~boos/var.select/hugh.crews.software.html>.

e-mail: boos@stat.ncsu.edu

### **VARIABLE SELECTION IN HIGH-DIMENSIONAL VARYING COEFFICIENT MODELS VIA THE ADAPTIVE GROUP LASSO**

Fengrong Wei\*, University of Iowa  
Jian Huang, University of Iowa

Nonparametric varying coefficient models are important in studying the time-dependent effects of variables. In this paper, we propose an adaptive group Lasso approach to variable selection and estimation in sparse, high-dimensional varying coefficient regression based on a spline approximation to the models. Under appropriate conditions, we show that the group Lasso selects a model of the right order of dimensionality, selects all variables whose coefficients are greater

than certain threshold level, and is estimation consistent. An interesting aspect of our results is that the logarithm of the number of variables can be of the same order as the sample size for certain random dependant designs. However, the group Lasso is in general not selection consistent and tends to also select variables that are not important in the model. We use the adaptive group Lasso to improve the selection results. We show that under suitable conditions, the adaptive group Lasso has an oracle selection property, in the sense that it can correctly select important variables with probability converging to one. In contrast, group Lasso do not possess such oracle property. Both methods are illustrated by simulation studies and a real data example.

e-mail: fengrong-wei@uiowa.edu

### **ON THE STUDY OF LAD-LASSO IN HIGH-DIMENSIONAL SETTINGS**

Xiaoli Gao\*, Oakland University  
Jian Huang, University of Iowa

Penalized regularization methods can achieve variable selection and estimation simultaneously by using specially designed penalty functions. The LAD regression is an interesting and robust alternative to the LS method. We propose a penalized LAD regression with the Lasso penalty (LAD-Lasso) and study the consistency properties of LAD-Lasso under some conditions in ‘large p, small n’ settings. In this presentation, we summarize the theoretical properties of LAD-Lasso on both estimation and model evaluation aspects in the high-dimensional case. The finite sample behavior of this estimator is studied and compared to LS-Lasso via simulation studies.

e-mail: gao2@oakland.edu

### **VARIABLE SELECTION IN THE KERNEL MACHINE FRAMEWORK**

Michael C. Wu\*, Harvard University  
Tianxi Cai, Harvard University  
Xihong Lin, Harvard University

The complexity of high-throughput biomedical data requires the use of flexible methods that can account for nonlinearity and complex interactions in building the predictive models. Kernel machine methods, e.g. support vector machines, meet these criteria and are frequently applied to build predictive models from genomics data. However, KM methods are still subject to considerable noise and decreased prediction accuracy when few predictors are related to the outcome. Variable selection is necessary. We propose a statistical framework for integrating regularization and variable selection with kernel machines that still maintains the flexible dual formulation. Connections with existing variable selection procedure are examined.

e-mail: mwu@hsph.harvard.edu

# ABSTRACTS

## VARIABLE SELECTION FOR ORDINAL RESPONSE MODELS WITH APPLICATIONS TO HIGH DIMENSIONAL DATA

Kellie J. Archer\*, Virginia Commonwealth University  
Andre Williams, Virginia Commonwealth University

Although prediction accuracy is often a goal in predictive modeling, often it is of interest to identify features most predictive of the observed response. For high-dimensional datasets such as those arising from gene expression microarray studies, many have used penalized models for automatic identification of important features. Others have used variable importance measures from random forests for identifying important features within a high-dimensional dataset. Penalized models and random forests have been described and applied to continuous, nominal class, and survival responses. However, in a large number of biomedical applications, the response to be predicted may be inherently ordinal. Examples of ordinal responses include TNM stage (I, II, III, IV), drug toxicity (none, mild, moderate, severe), and response to treatment (complete response, partial response, stable disease, progressive disease). Herein, we propose a method for L1 penalized ordinal response models. We also describe measures of variable importance from bootstrapped classification trees using our proposed ordinal impurity function. Results will be presented for both simulated and genomic datasets.

e-mail: kjarcher@vcu.edu

## CHOOSE AN OPTIMAL RIDGE PARAMETER IN PENALIZED PRINCIPAL-COMPONENTS BASED ON HERITABILITY

Man Jin\*, American Cancer Society  
Yuanjia Wang, Columbia University  
Yixin Fang, Georgia State University

To analyze high-dimensional data, a ridge penalized principal-components approach based on heritability was applied to avoid over-fitting. The optimal regularization parameter can be chosen by cross-validation. Due to computational intensity, a generalized cross-validation formula was developed in this paper to select the optimal parameter. From the simulation studies in four settings, the penalized principal-components of heritability analysis had substantially larger coefficients for the traits with genetic effect than for the traits with no genetic effect, while the non-regularized analysis failed to identify the genetic traits. Thus the penalized principal-components approach based on heritability can effectively handle large number of traits with family structure.

e-mail: man.jin@cancer.org

## 113. CLUSTERED DATA METHODS

### INFERENCE FOR VARIANCE COMPONENTS IN GENERALIZED LINEAR MIXED MODELS FOR POOLED BINARY RESPONSE

Joshua M. Tebbs\*, University of South Carolina  
Peng Chen, University of South Carolina  
Christopher R. Bilder, University of Nebraska-Lincoln

We investigate likelihood-based tests for variance components in generalized linear mixed models for pooled binary response. Pooled binary responses are commonly observed in group testing applications, where individual specimens (such as blood or urine samples) are tested in pools. Our variance component tests can be used to assess heterogeneity among clusters (e.g., testing sites), while preserving the anonymity of individual subjects. We illustrate our methods using chlamydia and gonorrhea data collected by the state of Nebraska as part of the Infertility Prevention Project.

e-mail: tebbs@stat.sc.edu

### ASSOCIATION MODELS FOR CLUSTERED DATA WITH BIVARIATE MIXED RESPONSES

Lanjia Lin\*, Florida State University  
Debajyoti Sinha, Florida State University  
Stuart Lipsitz, Harvard Medical School

We consider fully specified models and associated likelihood and Bayesian analysis of clustered data with mixed bivariate responses. We present a novel bivariate random effects model which induces associations among the binary outcomes within a cluster, among the continuous outcomes within a cluster, between a binary and a continuous outcome from different subjects within a cluster, as well as the direct association between the binary and continuous outcomes within the same subject. For the ease of interpretation of the regression effects, the marginal model of the binary response probability integrated over the random effects preserves the logistic form and the marginal regression function of the continuous response preserves the linear form. We implement maximum likelihood estimation of model parameters using standard software such as PROCNLIMIXED of SAS, as well as fully parametric Bayesian analysis. We extend our fully parametric model to accommodate a semiparametric Bayesian model using a Dirichlet mixture of normal for the continuous response. Posterior computations for both parametric and semiparametric models are implemented via Markov Chain Monte Carlo sampling methods. We illustrate our methodology by analyzing a developmental toxicity study of ethylene glycol in mice using the three methods.

e-mail: lanjia@stat.fsu.edu



## **ESTIMATION METHODS FOR AN AUTOREGRESSIVE FAMILIAL CORRELATION STRUCTURE**

Roy T. Sabo\*, Virginia Commonwealth University  
N. R. Chaganty, Old Dominion University

In this paper we apply an autoregressive correlation structure to the analysis of familial clustered data in the one-parent case with homogeneous intra-class variance. We use the quasi-least squares procedure to derive estimators of the correlation parameters and compare them with maximum likelihood and moments estimators. Asymptotically, the quasi-least squares estimators are nearly as efficient as the maximum likelihood estimators. The small-sample case is analyzed through simulation, and the quasi-least squares estimators are found to be more robust than the maximum likelihood estimators. A simple example is given to show the application of the estimation procedures. We also allude to the heterogeneous intra-class variance case, where the quasi-least squares procedure has certain advantages over the maximum likelihood procedure.

e-mail: rsabo@vcu.edu

## **INFERENCE FOR MARGINAL LINEAR MODELS WITH CLUSTERED LONGITUDINAL DATA WITH POTENTIALLY INFORMATIVE CLUSTER SIZES**

Ming Wang\*, Emory University  
Maiying Kong, University of Louisville  
Somnath Datta, University of Louisville

Clustered longitudinal data are often collected as repeated measures on subjects arising in clusters. Examples include periodontal disease study, where the measurements related to the disease status of each tooth are collected over time for each patient. Under such situations, generalized estimating equations (GEE) may lead to invalid inferences. We investigate the performance of three competing proposals of fitting marginal linear models to clustered longitudinal data, namely, generalized estimating equations (GEE), within-cluster resampling (WCR), and cluster-weighted generalized estimating equations (CWGEE). We show by simulations and theoretical calculations that, when the cluster size is informative, GEE provides biased estimators, while both WCR and CWGEE achieve unbiasedness under a variety of ‘working’ correlation structures for temporal measurements within each subject. Statistical properties of confidence intervals have been investigated using the probability-probability plots. Overall, CWGEE appears to be the recommended choice for marginal parametric inference with clustered longitudinal data that achieves similar parameter estimates and test statistics as WCR while avoiding Monte Carlo computation. We illustrate our analysis using a temporal dataset on periodontal disease which clearly demonstrates the need for CWGEE over GEE.

e-mail: wm\_pk@hotmai.com

## **QUASI-LEAST SQUARES WITH MIXED CORRELATION STRUCTURE**

Jichun Xie\*, School of Medicine, University of Pennsylvania  
Justine Shults, School of Medicine, University of Pennsylvania

In this paper, we focus on a two-stage nonparametric approach, quasi-least squares (QLS), which yields a consistent estimator for the correlation parameter in longitudinal data analysis. Our work is motivated by a desire to appropriately model the correlation and regression parameters in an analysis of families in a longitudinal Ophthalmology study in Old Order Amish (OOA). To argue the validity of QLS for the OOA analysis, we first prove some properties of QLS for linear correlation structures, that previously have only been proven on a case by case basis for particular structures: namely, that the stage one QLS estimator always exists and is feasible and that the stage two estimator is unique. Based on our general proofs, we then discuss application of QLS for analysis of longitudinal data that are missing completely at random (MCAR) and that are missing at random (MAR). We then implement QLS for familial longitudinal data with families of varying sizes, which is a special case of data with mixed correlation structures. We conduct simulations to assess the performance of QLS in estimation of the correlation parameter. We then conduct an analysis of the Ophthalmology study in the Old Order Amish, which reveals insights regarding the intra-familial correlations that are of scientific interest.

e-mail: jichun@mail.med.upenn.edu

## **GENERALIZED VARYING COEFFICIENT SINGLE-INDEX MIXED MODEL**

Jinsong Chen\*, Virginia Polytechnic Institute and State University  
Inyoung Kim, Virginia Polytechnic Institute and State University  
George R. Terrell, Virginia Polytechnic Institute and State University

Generalized linear mixed model (GLMM) is widely used for the analysis of correlated data/clustered data, and time independent functional form of the predictor in this model is usually assumed. However, the linear model is not complex enough to capture the underlying relationship between the response and its associated covariates. And also time-independent functional form may be too restrictive to represent true fundamental covariate effects. Therefore, in this paper, we generalize this model to nonparametric single-index mixed model and also allow this model to have varying coefficients. We call this model a generalized varying coefficient single-index mixed model (GVSIMM). We propose a penalized likelihood approach to estimate varying single-index coefficient. Using bootstrapping approach and asymptotic theory, we make a inference for parameters. Simulation studies are performed to compare our GVSIMM with GLMM. The study of the association between daily air pollutants and daily mortality in various counties of North Carolina is applied to demonstrate the advantage of our approaches.

e-mail: jschen24@hotmail.com

# ABSTRACTS

## ON NONPARAMETRIC TESTS FOR PARTIALLY CORRELATED DATA WITH APPLICATION TO PUBLIC HEALTH ISSUES

Hani M. Samawi\*, Georgia Southern University  
Lili Yu, Georgia Southern University  
Robert Vogel, Georgia Southern University  
Laura H. Gunn, Georgia Southern University

Correlated or matched data is frequently collected under many study designs in applied sciences such as the social, behavioral, economic, biological, medical, epidemiologic, health, public health, and drug developmental sciences. Challenges with respect to availability and cost commonly occur with matching observational or experimental study subjects, thus researchers frequently encounter situations where the observed sample consists of a combination of correlated and uncorrelated data. This paper discusses and proposes testing procedures to handle data when partially correlated data is available. Theoretical as well as numerical investigation will be provided. The proposed testing procedures will be applied to real data. These procedures will be of special importance in meta-analysis where partially correlated data is a concern when combining results of various studies.

e-mail: hsamawi@georgiasouthern.edu

## 114. ESTIMATION IN SURVIVAL MODELS

### SEMIPARAMETRIC INFERENCE OF LINEAR TRANSFORMATION MODELS WITH LENGTH-BIASED CENSORED DATA

Jane Paik\*, Columbia University  
Zhiliang Ying, Columbia University

In this article we propose an estimation method for the regression parameters and the transformation function in semiparametric linear transformation models when dealing with biased sampling in censored data. This paper is the first to propose a method for length-biased sampling in linear transformation models. Existing estimation procedures for censored data using linear transformation models yield biased estimators for regression parameters of interest. Our approach is motivated by the unified estimation procedure proposed by Chen et al. (2002) which made use of the martingale integral representation. The proposed estimators for the regression parameters are proven to be consistent and asymptotically normal. The variance-covariance matrix has a closed form which can be consistently estimated.

e-mail: jane@stat.columbia.edu

### INTERCEPT ESTIMATION FOR THE SEMIPARAMETRIC ACCELERATED FAILURE TIME MODEL

Ying Ding, University of Michigan  
Bin Nan, University of Michigan

There is a rich literature on the slope parameter estimation for a linear regression model for censored survival data when the error distribution is unspecified. The estimation of intercept, however, has not been thoroughly studied mostly because that the follow up time is usually finite in practice so the intercept, directly related to the mean survival time, would be inevitably underestimated. Motivated by the results of consistent estimation of the mean survival time from Susarla and Van Ryzin (1980) and Stute and Wang (1993), in this paper we show that the intercept can be consistently estimated when the support of some covariates with nonzero coefficients is unbounded. Without the commonly assumed regularity condition on bounded covariates, we also show that the slope estimators obtained from the rank-based estimating equation with Gehan weights are consistent and asymptotically normal, which provides a crucial condition for the consistency of the intercept estimator. The theoretical finding is further verified for finite samples by a simulation study. Simulation also shows that, when both models are correct, the accelerated failure time model yields reasonable mean square errors for survival time prediction and outperforms the Cox model for censored data, particularly with heavy censoring. An illustrative real data example is provided.

e-mail: yingding@umich.edu

### EFFICIENT ESTIMATION IN THE PARTLY LINEAR ADDITIVE HAZARDS REGRESSION MODEL WITH CURRENT STATUS DATA

Xuewen Lu\*, University of Calgary  
Peter X.-K. Song, University of Michigan  
John D. Kalbfleisch, University of Michigan

We study efficient estimation in the partly linear additive hazards regression model with current status data. We consider the use of polynomial splines to approximate the cumulative baseline hazard function with monotone constraints and nonparametric regression functions without constraints. Both regression coefficients and nuisance parameters are estimated simultaneously using the method of maximum likelihood estimation. The estimator of the finite-dimensional vector of regression parameters is shown to be asymptotically normal and achieves the semiparametric information bound. Rates of convergence for the estimators of the nonparametric components are investigated. To implement estimation, we treat the model as a generalized linear model and apply the iterative weighted least squares method. We conduct simulation studies to examine the finite sample performance of the estimators and computational challenges in the proposed methods.

e-mail: lux@math.ucalgary.ca



## **EFFICIENT ESTIMATION FOR THE PROPORTIONAL ODDS MODEL WITH BIVARIATE CURRENT STATUS DATA**

Bin Zhang\*, University of Missouri  
Xingwei Tong, Beijing Normal University  
Jianguo Sun, University of Missouri

Efficient estimation approach provides a useful tool for semiparametric analysis of failure time data if the main interest is estimation of regression parameters (Huang, 1996; Martinussen and Scheike, 2002). This paper discusses the application of this approach to regression analysis of bivariate current status data arising from the proportional odds model (Yang and Prentice, 1999; Rabinowitz et al., 2000). Current status data arise if each study subject is observed only once and often occur in fields including demographical studies and tumorigenicity experiments. For the analysis, the copula model is assumed for the joint survival function of two related failure time variables which are supposed to follow the proportional odds model marginally. Simulation studies indicate that the estimates of regression parameters derived perform well in practical situations and the methodology is illustrated using a set of data from a tumorigenicity experiment.

e-mail: bz38d@mizzou.edu

## **SEMIPARAMETRIC CURE RATE MODELS FOR CURRENT STATUS DATA**

Guoqing Diao\*, George Mason University

In this research we study a class of semiparametric cure rate models for the analysis of current status data. This class includes the commonly used mixture cure rate model and proportional hazards cure model as special cases. We show that the nonparametric maximum likelihood estimators for the regression parameters of these models are consistent, asymptotically normal, and asymptotically efficient. We conduct extensive simulation studies to evaluate the performance of the proposed method. An illustration with a real study is provided.

e-mail: gdiao@gmu.edu

## **ACCELERATED HAZARDS MIXTURE CURE MODEL**

Jiajia Zhang\*, University of South Carolina  
Yingwei Peng, Queen's University

We propose a new cure model for survival data with a surviving or cure fraction. The new model is a mixture cure model where the covariate effects on the proportion of cure and the distribution of the failure time of uncured patients are separately modeled. Unlike the existing mixture cure models, the new model allows covariate effects on the failure time distribution of uncured patients to be negligible at time zero and to increase as time goes by. Such a model is particularly useful in some cancer treatments when the treat effect increases gradually from zero, and the existing models usually cannot handle this situation properly. We develop a rank based semiparametric estimation method to obtain the maximum likelihood estimates of the parameters in the model. We compare it with existing models and

methods via a simulation study, and apply the model to a breast cancer data set. The numerical studies show that the new model provides a useful addition to the cure model literature.

e-mail: jzhang@gwm.sc.edu

## **115. BIOLOGICS, PHARMACEUTICALS, MEDICAL DEVICES**

### **PROPOSED METHODOLOGY FOR SHELF LIFE ESTIMATION**

Michelle Quinlan\*, University of Nebraska-Lincoln  
James Schwenke, Boehringer Ingelheim Pharmaceuticals, Inc.  
Walt Stroup, University of Nebraska-Lincoln

As part of the research efforts of the Product Quality Research Institute (PQRI) Stability Shelf Life Working Group, statistical methodology is being developed to directly estimate the shelf life of a pharmaceutical product. The proposed methodology presented here estimates shelf life based on the overall mean response among batches of a pharmaceutical product for a stability limiting characteristic. Incorporating random batch effects into the statistical model and using calibration techniques allow for direct estimation of shelf life as opposed to indirect methods of pooling data or relying on a worst batch scenario. The estimated shelf life is the storage time corresponding to the point where the predicted mean response intersects the specification limit or acceptance criteria. A lower interval estimate is constructed about the calibrated point to determine the labeled shelf life. Results of a computer simulation will be presented to validate the statistical methodology and to demonstrate the advantages over the current approach for estimating shelf life as defined by ICH guidelines. An example using real life data will be presented to highlight the proposed methodology for shelf life estimation.

e-mail: mquinlan22@yahoo.com

## **EXTENDING GROUP SEQUENTIAL METHODS TO OBSERVATIONAL MEDICAL PRODUCT SAFETY SURVEILLANCE**

Jennifer C. Nelson\*, Group Health Center for Health Studies; University of Washington  
Andrea Cook, Group Health Center for Health Studies; University of Washington  
Shanshan Zhao, Group Health Center for Health Studies; University of Washington

Conducting large-scale, proactive, and rapid post-marketing medical product safety surveillance is important for detecting rare adverse events potentially not identified in pre-licensure studies. To detect adverse events as early as possible after the introduction of a new product, continuous monitoring methods such as Wald's classical and Kulldorff's maximized sequential probability ratio test (SPRT) have been proposed. Although continuous monitoring is advantageous for rapid detection, such frequent monitoring may not be feasible or desirable in some instances. And while SPRT-based methods can be applied on a less frequent basis, they may not be optimal in terms of statistical power. A more natural methodology for testing

# ABSTRACTS

on a periodic basis is group sequential interim monitoring. Group sequential methods are well developed and widely used in clinical trials for monitoring drug safety and efficacy. However, their use in observational settings has not been considered. We will discuss the issues that arise when extending group sequential testing methods to observational safety surveillance settings. We will also present results of a simulation study that evaluates the performance of continuous SPRT-based monitoring methods in an observational study setting compared to several standard group sequential designs.

e-mail: nelson.jl@ghc.org

## A CONDITIONAL MAXIMIZED SEQUENTIAL PROBABILITY RATIO TEST FOR PHARMACOVIGILANCE

Lingling Li\*, Department of Ambulatory Care and Prevention, Harvard Medical School  
Martin Kulldorff, Department of Ambulatory Care and Prevention, Harvard Medical School

The importance of post-marketing surveillance for drug and vaccine safety is well recognized, as rare but serious adverse events may not be detected in pre-approval clinical trials. In such surveillance, it is natural to use a sequential test, as we prefer to detect the severe adverse events as soon as possible. Various sequential probability ratio tests (SPRT) have been widely applied in real time vaccine and drug safety surveillance, including Wald's classical SPRT with a single alternative and the maximized SPRT (MaxSPRT) with a composite alternative. These methods require that the expected number of events under the null is known as a function of time  $t$ . In practice, the expected counts are usually estimated from historical data. When we do not have a large sample size from the historical data, the SPRTs will be biased due to the variance in the estimate of the expected number of events. We present a conditional maximized sequential probability ratio test (CMaxSPRT), which adjusts for the uncertainty in the expected counts. Our test incorporates the randomness and variability from both the historical data and the surveillance population. Evaluation of the power performance of CMaxSPRT under different scenarios will be presented.

e-mail: lingling07.li@gmail.com

## SELECTION AND EVALUATION OF BIOMARKERS USING INFORMATION THEORETIC APPROACH

Abel Tilahun\*, Universiteit Hasselt  
Dan Lin, Universiteit Hasselt  
Suzy Van Sanden, Universiteit Hasselt  
Ziv Shkedy, Universiteit Hasselt  
Ariel Alonso, Universiteit Hasselt  
Geert Molenberghs, Universiteit Hasselt

The selection and evaluation of biomarkers plays a vital role in the discovery and development of new drugs. This calls for new and efficient statistical methods that can be used in different scenarios for the selection and evaluation purposes. While joint models have been proposed for several settings involving a combination of normally and non-normally distributed outcomes, they are cumbersome in the sense

of computationally complex and of producing validation measures that are, unlike in the Gaussian case, not of an R-squared type (Van Sanden et al. 2007). A way to put these problems to rest is by employing information theory, already applied in the continuous case (Alonso and Molenberghs 2007). In this paper, the information-theoretic approach is applied to the selection and evaluation of genomic markers. Its use is illustrated using case studies and its performance, relative to existing methods, is also assessed.

e-mail: abel.tilahuneshete@uhasselt.be

## PATIENT FOCUSED METHOD FOR ASSESSING IN VITRO DRUG COMBINATIONS USING GROWTH RATES

Maksim Pashkevich\*, Eli Lilly and Company  
Philip Iversen, Eli Lilly and Company  
Harold Brooks, Eli Lilly and Company

We propose a new method that allows screening oncology drug combinations using data from in-vitro studies to select agents that have the promise of showing a synergistic effect in-vivo. In contrast to known approaches that define combination effect either on concentration scale or on percent inhibition scale, we use the growth rate of treated cells as a primary indicator of treatment activity. The developed method is based on a novel mathematical model that describes the growth of cancer cells that are subject to treatment with a combination of compounds. Mathematically, the model assumes a multi-compartment cell population with transition rates between compartments modeled according to biochemical reaction properties, and cells in each compartment growing according to exponential law. This translates to a linear system of ordinary differential equations, whose solution is accurately approximated by a closed-form expression using rapid equilibrium assumptions. Special cases of the aforementioned model represent situations when the combination effect is absent or when the considered drugs act as the same compound. Akaike information criterion and the likelihood ratio test are used to distinguish between different mechanisms of action for the considered compounds, and to test if a significant combination effect is being observed.

e-mail: pashkevich\_maksim@lilly.com

## SPLIT-PLOT DESIGNS IN SERIAL DILUTION BIOASSAY USING ROBOTS

Jeff Buzas\*, University of Vermont  
Carrie Wager, Precision Bioassay  
David Lansky, Precision Bioassay

Serial dilution bioassay is routinely used for relative potency estimation of lot release and in stability studies of biotechnology products. Typically, samples and dilution order are separately assigned (ideally at random) to rows and columns of 96 well plates, resulting in a strip-plot design. Robots are increasingly used to implement serial-dilution designs, and robots with individual tip control can implement split-plot designs, i.e. designs where dose order need not be the same for each sample compound. For a given split-plot design, there are many possible paths a robot can take to fill wells. We show the shortest path

is equivalent to the shortest common supersequence (SCS) problem, and describe an algorithm for finding the SCS useful in bioassay applications. We also describe an algorithm for the reverse process: We describe how to generate split-plot designs that can be filled in nearly the same number of steps as strip-plot designs.

e-mail: buzas@cem.s.uvm.edu

## LOG-RANK TEST WEIGHT SELECTION FOR HAZARD RATIO WITH A CHANGE-POINT

Jun Wu\*, Bristol Myers Squibb  
Howard Stratton, School of Public Health, SUNY Albany

The anti-cancer effect of immunotherapy is delayed because it needs to activate the immune system to mount cytotoxic attack. Weighted log-rank family tests are typically used to compare overall survival with non-constant hazard ratios. The power and size of tests with weight selection aided by change-point Cox model will be assessed via simulation based on FDA published data from Phase III clinical trials of a cancer vaccine.

e-mail: jun.wu@bms.com



# Access Professional Resources with the American Statistical Association

Consider membership in the American Statistical Association, the premier professional organization for statisticians. The ASA is here for you, whether you practice in an academic, business, or government setting.

## **ASA MEMBERS ENJOY:**

**A SUBSCRIPTION** to *Amstat News*,\* the ASA's monthly membership magazine, full of upcoming events and job opportunities

**NETWORKING** through the ASA's regional chapters and special-interest sections

**MEMBERS-ONLY FEATURES** of the ASA web site, such as member forums, publisher discounts, and an enhanced member directory

**OPPORTUNITIES** to expand career horizons with the ASA's Career Placement Service at the Joint Statistical Meetings and the ASA's JobWeb at <http://jobs.amstat.org>

**DISCOUNTED** registration fees for the annual Joint Statistical Meetings and Continuing Education courses

**ONLINE ACCESS\*\*** to the Current Index to Statistics (CIS), *Journal of the American Statistical Association* (JASA), *Journal of Business & Economic Statistics* (JBES), and *The American Statistician* (TAS)

\* Family and Developing Country memberships do not include these specific benefits.

\*\* Family and Life-Retired memberships do not include these specific benefits.



# **SAS® Publishing and JMP® Statistical Discovery Software**

**Visit the SAS® booths to learn how SAS can help you  
leverage world-class business analytics and get ahead!**

## **Build your skills with SAS® Publishing.**

SAS® Press titles deliver expert advice from SAS® users worldwide.

SAS® Documentation provides the information you need to use SAS products and solutions. Whether you are new to SAS or a seasoned user, SAS Publishing can help you prepare for the SAS certification exams. Attendees receive 20% off all SAS Publishing titles!

**[support.sas.com/bookstore](http://support.sas.com/bookstore)**

## **Discover dynamic data visualization with JMP® statistical discovery software.**

Interactive, comprehensive, and highly visual, JMP is the desktop software from SAS that lets you work with your data to explore relationships, see hidden trends, and dig into areas that interest you.

JMP® Genomics, customized for vast life sciences data sets, brings the power of SAS to desktop genomic data analysis.

**[www.jmp.com](http://www.jmp.com)**



Statistical Discovery™ From SAS.



# INDEX

Aban, Chichi	7f	Barr, Christopher D	34
Abebe, Fisseha	32	Barry, William T	3f
Abecasis, Goncalo R.	44	Basu, Sanjib	70
Ahn, Chul	2d	Basu, Saonli	44
Ahn, Jeongyoun	93	Bathke, Arne C	46
Albert, Paul S.	25, 70	Bautista, Dianne	102
Albrechtsen, Anders	30	Beavers, Daniel P	21
Allen, Genevera I.	40	Bekaert, Maarten	5b
Allison, David B.	77	Bekele, B. Nebiyou	64
Almeida, Jonas S.	100	Bekeley, Benjamin N	88
Alonso, Ariel	21, 115	Bell, Michelle L	85
Alosh, Mohamed	23, 103	Benson, Constance	80
An, Di	57	Berg, Arthur	20, 80, 111
Anderson, Michael P.	32	Bergan, Raymond C	109
Anderson, Stewart J.	92, 104	Bergemann, Tracy L	89, 111
Andrade, Dalton F.	3b	Bergquist, Mandy L	109
Andrade, Mariza de	3b, 111	Berrocal, Veronica J	10a, 79, 108
Andrei, Adin-Cristian	25, 70	Berry, Donald A	7c, 97
Andrei, Alina	100	Berry, Scott M	R8, 28
Andridge, Rebecca R.	57	Betensky, Rebecca A	40
Apanasovich, Tatiyana V.	34	Betts, Keith A	92
Archer, Kellie J.	112	Beunckens, Caroline	39
Arshad, Hasan	4b	Bhattacharjee, Samsiddhi	87
Arterburn, David	34	Bhattacharya, Sourab	10a
Asparouhov, Tihomir	59	Bilder, Christopher R	113
Assam, Pryseley N.	21	Billard, Lynne	101
Atlas, Mourad	99	Boden, William E	8d
Augustin, Nicole H.	91	Boehnke, Michael	20, 77
Aukema, Brian	9d	Boerwinkle, Eric	77
Bai, Yun	9e	Bokov, Alex F	7b
Bailer, A. John	80	Boland, Richard C	2c
Bailey, Kent R	37	Bondell, Howard D	96
Baiocchi, Michael	61	Boos, Dennis D	112
Ball, Edward	110	Borkowski, John	99
Bamlet, William	67	Bowman, DuBois	SC5, 56
Bancroft, Tim	48	Boyle, James P	10e, 11d
Bandeen-Roche, Karen	65	Braun, Thomas M	1b
Bandos, Andriy I	68	Breitkreutz, Ashton	63
Bandyopadhyay, Dipankar	1e, 14a	Brignell, Christopher J	9k
Banerjee, Sudipto	SC3, 42, 72, 85, 102	Brinkman, Ryan	76
Banks, David L.	15	Brinton, John T	82
Bao, Jieqiong	23	Broglio, Kristine	7c
Barhak, Jacob	10b	Brooks, Harold	115
Barker, Lawrence	10e, 11d	Brown, Hendricks C	22
Barnes, Sunni A	2c	Brown, Patrick E	9c, 9f
Barnhart, Huiman	81	Brunelli, Steven	61
		Budinska, Eva	105
		Burch, Christina	33

Busman, Todd A	110	Chen, Min	43
Buzas, Jeff	115	Chen, Peng	113
Cadwell, Betsy L	10e, 11d	Chen, Ping	37
Caffo, Brian	56	Chen, Qingxia	72
Cai, Bo	46	Chen, Qixuan	57
Cai, Jianwen	25, 27	Chen, Rui	47
Cai, Tianxi	18, 83	Chen, Shufeng	89
Cai, Tony	48	Chen, Shuo	56
Calabrese, Edward J	104	Chen, Shyh-Huei	36
Cao, Hongyuan	48	Chen, Yian Ann	100
Cao, Jing	69, 112	Chen, Zehua	19
Cao, Weihua	23	Cheng, Cheng	55
Cao, Xueyuan	55	Cheng, Dunlei	2c
Cardon, Zoe	78	Cheng, Jing	61
Cardot, Herve	95	Cheng, Nancy F	10f
Carey, Vincent	SC1	Cheng, Yasheng	60
Carlin, Brad	SC6, 26	Cheng, Yu	25
Carriere, Keumhee Chough	91	Cheng, Yu-Jen	58
Carroll, Raymond J	34, 89, 100	Cheung, Ken	64
Casanova, Ramon	4d	Chi, Yueh-Yun	47
Casella, George	77	Chinchilli, Vernon M	81
Chaganty, N. Rao	59, 79, 113	Cho, Hyunsoon	69
Chakraborty, Avishek	78	Choi, Hyungwon	63
Chakraborty, Bibhas	88	Choi, Jaeun	27
Chakraborty, Dev P	68	Choi, Suh-yeon	15
Chakraborty, Hrishikesh	2f, 24	Chowdhury, Dhuly	2f
Chaloner, Kathryn	12a	Chowdhury, Rafiqul I	47
Chan, Wenyaw	82	Christakis, Nicholas A	16
Chandrasekhar, Rameela	1d	Christensen, Brock C	40
Chang, Chung	103	Christman, Mary C	78
Chang, Howard H	103	Chuang, Ya-Hsiu	78
Chang, Maryon M	80	Chung, Hwan	90
Charles, Janelle K	14e	Chung, Moo K	60
Chatterjee, Nilanjan	87, 89	Chung, Yeojin	101
Chauhan, Chand K	33	Chung, Yeonseung	11a
Chen, Baojiang	51	Clement, Meagan E	56
Chen, Cuixian	13a, 13b	Close, Nicole C	109
Chen, Din	35	Cohen, Margaret A	9a
Chen, Dung-Tsa	55	Cook, Andrea J	34, 78
Chen, James J	55	Cook, Andrea	115
Chen, Jiahua	19	Cook, Dianne	15
Chen, Jinbo	87	Cook, Richard	38, 51
Chen, Jinsong	113	Cooper, Robin L	36
Chen, Li	11e	Coull, Brent A	40, 52, 102
Chen, Lin S	67	Couper, David	56
Chen, Lin-An	82	Covault, Jonathan	98
Chen, Ling	8b	Cowen, Mark	31
Chen, Meng	20	Craig, Bruce A	7e, 90

# INDEX

Crainiceanu, Ciprian M	45, 52, 58, 93	Dominici, Francesca	34, 90, 103
Crews, Hugh B	112	Donev, Alexander N	75
Crews, Kristine R	55	Dong, Yuexiao	101
Crofton, Kevin	35	Donoho, David L	19
Crooks, Kevin	17	Donovan, Mark	104
Cropsey, Karen L	5a	Downing, James R	55
Cui, Lu	91	Dragalin, Vladimir	54
Cui, Xiangqin	3a	Dryden, Ian L	9k
Cui, Yuehua	20, 32, 89, 111	Du, Pang	47
Cushman, Mary	7i	Du, Yining	1f
Cutter, Gary R	7f	Dubnicka, Suzanne R	32
Cyril, Rakovski	41	Dukic, Vanja M	86
Czogiel, Irina	9k	Dunson, David B	11a, 35, 86
Daggy, Joanne K	90	Dupuis, Josée	96
Dahl, David B	48, 63	Ebrahimi, Nader	55
D'Angelo, Gina	23	Edwards, Lloyd J	34
Daniels, Michael J	R10, 39, 103	Edwards, Sharon	9g
Daoud, Yahya A	2c	Egleston, Brian L	5a, 22
Das, Sourish	98	Ehrlich, Melanie	100
Datta, Somnath	25, 113	Eisen, Ellen A	92
Datta, Susmita	99	El-Kamary, Samer S	91
Datto, Michael	3f	Elliott, Michael R	5c, 57, 84
Daumer, Martin	70	Elmi, Angelo	93
Davidian, Marie	R3, 22, 23, 70	Eloyan, Ani	59
Davis, Justin W	43	Enders, Felicity B	82
Day, Ryan	63	Entsuah, Richard	88
De Andrade, Mariza	3b, 111	Ericsson, Tore	85
De Oliveira, Victor	78	Ewart, Susan	4b
DeGruttola, Victor	80	Fan, Jianqing	11g, 19, 50
Delaigle, Aurore	95	Fan, Ruzong	77
Deng, Dianliang	100	Fan, Yingying	50
Deng, Yihao	59, 79	Fang, Hong-Bin	93, 100
Derado, Gordana	56	Fang, Yixin	111, 112
DeSantis, Stacia M	40	Fardo, David	43
DeVito, Mike	35	Farjah, Farhood	90
DeVol, Ed	2c	Fedorov, Valerii V	75
Dey, Dipak K	32, 98	Feldman, Harold I	61
DeYoe, Edgar A	6f	Feng, Huaibao	54
Di, Chongzhi	45, 93	Feng, Rui	111
Diao, Guoqing	114	Feng, Yang	11g
DiCasoli, Carl M	7d	Feng, Yijia	54
Diggle, Peter J	71	Ferreira, Marco A	78
Ding, Rui	3d	Fiecas, Mark	73
Ding, Ying	92, 114	Fine, Jason P	53
Divers, Jasmin	4d	Finley, Andrew O	SC3, 42, 85
Dixon, Dennis O	65	Fleet, James C	67
Dixon, Philip M	15	Flum, David R	90
Djira, Gemechis D	35	Fochesatto, Javier	102

Freisthler, Bridget J	66	Goode, Ellen L	67
French, Benjamin	90	Gorelick, Marc	9i
Friderici, Karen	4b	Gorman, Dennis M	66
Fridley, Brooke L	67	Gottardo, Raphael	76
Fu, Haoda	54	Govindarajulu, Usha S	92
Fuentes, Montse	59	Grab, Josh D	4d
Fuentes, Montserrat	66, 85	Graubard, Barry I	57
Furrer, Reinhart	SC4	Gray, Simone	9g
Gail, Mitchell H	4c, 20, 89, 94	Green, Bonnie L	81
Gamst, Anthony	80	Greenhouse, Joel	16
Gan, Jianjun	57	Gregoire, Timothy G	42
Gangnon, Ronald E	34, 102	Greven, Sonja	34
Ganguli, Bhaswati	92	Gruenewald, Paul J	66
Gansky, Stuart A	10f, 57	Gu, Dongfeng	89
Gao, Jingjing	81	Gu, Jessie	94
Gao, Xiaoli	112	Gu, Wen	94
Gao, Xin	5c	Gu, Xuemin	54
Gardiner, Joseph C	31, 90	Gu, Yu	72
Garrett-Mayer, Elizabeth	1e	Guan, Shanhong	45
Gastwirth, Joseph L	92	Guan, Weihua	77
Gatsonis, Constantine	91, 107	Guan, Yongtao	108
Gaydos, Brenda L	28	Guerry, DuPont	45
Gelfand, Alan E	9g, 10a, 59, 78	Gumpertz, Marcia	R9
Gelfond, Jonathan A	7b, 100	Gunn, Laura H	113
Gennings, Chris	35	Gunst, Richard F	6d, 6e
Genton, Marc G	34, 108	Gunter, Lacey	106
George, Andrew	111	Guo, Jian	112
George, Nysia I	43	Guo, Mengye	48
Ghosal, Subhashis	7d	Guo, Wenge	48
Ghosh, Arpita	44, 98	Guo, Wensheng	93
Ghosh, Debashis	99	Guo, Ying	SC5, 58, 60
Ghosh, Jayanta K	7e	Gur, David	68
Ghosh, Joyee	46	Gurka, Matthew J	14c
Ghosh, Kaushik	82	Haber, Michael	81
Ghosh, Malay	98	Hahne, Florian M	76
Ghosh, Pulak	70, 82	Hall, Penny	78
Ghosh, Sujit K	7d, 59	Hall, Peter	33, 95
Giacolletti, Katherine E.D.	35	Hamilton, Kiya R	110
Gibbons, Robert D	81	Han, Younghun	82
Gimotty, Phyllis A	45	Hanfelt, John J	79
Gingras, Anne-Claude	63	Hanson, Timothy E	102
Giolo, Suely R	3b, 111	Harel, Ofer	23, 98
Giurcanu, Mihai C	33	Haribhai, Dipecia	8h
Glueck, Deborah H	21, 24, 82	Harlow, Sioban	8a
Gneiting, Tilmann	108	Harrell, Frank E., Jr.	SC2
Gobakken, Terje	42	Harrington, David P	92
Goldsmith, Linda Jane	69	Hart, Jeffrey D	12c
Gonen, Mithat	68	Hartwell, Tyler D	24

# INDEX

Haubrich, Richard	80	Huang, Xuelin	97
Hauer-Jensen, Martin	2a	Huang, Yijian	53
He, Chong Z	102	Hubbard, Rebecca A	81
He, Jing	44	Huber, Dudley A	77
He, Xuanyao	34, 102	Huda, Shahriar S	47
He, Xuming	53	Huebner, Marianne	4b
He, Yi	54	Hung, H.M. James	38, 49
Heagerty, Patrick J	18, 29, 68, 90	Hunter, David	73
Heckman, Carolyn J	5a	Hurst, Greg	99
Hedeker, Donald	47	Hyrien, Ollivier	47
Heitjan, Daniel F	31, 37, 48, 58	Ibrahim, Joseph G	56, 60, 69, 72
Hendry, Martin	93	Ionita-Laza, Iuliana	43
Henry, David H	104	Isaman, Deanna J.M	10b
Henrys, Peter	9c	Ishak, Khajak	86
Herring, Amy H	46, 59, 66, 85	Islam, M.A.	47
Heyse, Joseph F	35	Ivanova, Anastasia	64
Hilbe, Joseph M	62	Iversen, Philip	115
Hill, Elizabeth G	21	James, Gareth	50
Hillis, Stephen L	68, 81	Janes, Holly	94
Hiriote, Sasiprapa	81	Jang, Woncheol	93
Hoffmann, Raymond G	6f, 8h, 9i, 12d, 56	Jenkins, Gregory	67
Hofmann, Heike	15	Jeon, Youngsook	88
Hogan, Joseph W	39	Jeong, Kwonho	21
Holsinger, Kent E	32	Ji, Hongkai	74
Hossain, M.D. M	102	Ji, Yuan	64
Hou, Wei	4a	Jiang, Aixiang	101
Houseman, Andres	40	Jiang, Jiancheng	50
Houseman, E. Andres	40	Jin, Jiashun	19
Hsing, Tailen	95	Jin, Man	112
Hsu, Chiu-Hsieh	23, 103	Jo, Booil	22
Hsu, Fang-Chi	36	Joffe, Marshall M	22, 61
Hsu, Jason C	24	Johnson, Dallas E	24
Hsu, Li	67	Johnson, Lynn	38
Hsu, Ying-Lin	55	Johnson, Timothy D	R6, 6a, 6b, 6c
Hsuan, Francis	54	Johnson, Valen	69
Hu, Chen	7g	Jolly, Anna K	99
Hu, Chengcheng	80	Joo, Yongsung	79
Hu, Fan	2d	Jovanovic, Borko D	109
Hu, Feifang	88, 97	Jung, Inkyung	9b
Hu, Joan X	8i	Juraska, Michal	7i
Hu, Ming	100	Kai, Bo	101
Hu, Yijuan	44	Kaizar, Eloise	86
Hua, Lei	101	Kalbfleisch, John D	41, 92, 110, 114
Huang, Bevan E	8h, 9i	Kalendra, Eric	59, 66
Huang, Emma	111	Kang, Jian	6c
Huang, Jian	112	Kang, Sangwook	25
Huang, Liping	46	Kargas, Margaret R	40
Huang, Xiaodong	99	Karimpour-Fard, Anis	24

Karmaus, Wilfried	4b	Kwon, Deukwoo	3d
Katki, Hormuzd A	45	Lababidi, Samir	55
Keighley, John D	24, 91	Laber, Eric	106
Keles, Sunduz	74	Lacey, Michelle R	100
Kelsey, Karl T	40	Lamba, Jatinder	55
Kennedy, Richard E	3a	Lan, Kong	21
Kenward, Michael G	27, 39	Lan, Kuang-Kuo G	49
Khodursky, Arkady	100	Lan, Ling	25
Kim, Daeyoung	82	Landes, Reid D	2a
Kim, Dong-Yun	89	Lane, Thomas	110
Kim, Eunhee	21, 68	Lange, Christoph	43, 96
Kim, Inyoung	69, 99, 113	Lange, Kenneth	77
Kim, Jaejik	101	Langefeld, Carl D	4d
Kim, Jong-Min	3d	Langholz, Bryan	41
Kim, Mi-Ok	25, 91	Lansky, David	115
Kim, Sinae	48, 55	Larget, Bret	32
Kim, Sungduk	46	Larsen, Brett	63
Kim, Wonkuk	24, 45	Laud, Purushottam, W	20
Kim, Yang-Jin	80	LaValley, Michael P	59, 104
King, Martin S	110	LaVange, Lisa	38
Kishko, Michael	80	Lawrence, Michael	15
Klein, Andreas G	22	Lawson, Andrew B	R7, 102
Klein, Gary	16	Lazar, Ann A	47
Klein, John P	37	Lazar, Nicole A	60
Klein, Martin	32	Lazev, Amy B	5a
Knapp, Guido	104	LeBlanc, Michael	87
Kodell, Ralph L	2a	Lee, Jack J.	54
Koech, Wilson A	79	Lee, JungBok	79
Koh, Ohn Jo	6e	Lee, Juyoun	10c
Kolaczyk, Eric D	73	Lee, Keunbaik	79
Kolm, Paul	8d, 90	Lee, Mei-Ling Ting	51
Kong, Lan	21	Lee, Mihee	33
Kong, Maiying	113	Lee, Minjae	21
Kong, Xiaoxiao	99	Lee, Sang Han	93
Kooperberg, Charles	87	Lee, Shih-Yuan	82
Korn, Edward L	57	Lee, Yoonkyung	24
Kort, Eric J	55	Lensing, Shelly Y	2a
Kosorok, Michael R	48, 110	Leon Novelo, Luis G	88
Kottas, Athanasios	63	Leonard, Stefanksi A	70
Kou, Samuel	83	Leonov, Sergei	75
Kranzler, Henry R	98	Levina, Elizaveta	112
Krieger, Jose E	3b, 111	Li, Biao	89
Krishnamurthy, Ashok	110	Li, Bin	11f, 36, 66
Kuan, Pei Fen	74	Li, Bing	101
Kubilis, Paul S	78	Li, Bo	108
Kublin, Edgar	91	Li, Di	110
Kulldorff, Martin	9b, 115	Li, Erning	27
Kumar, K. Sree	2a	Li, Gengxin	20

# INDEX

Li, Haihong	88	Lipsitz, Stuart R	98
Li, Hongfan	89	Lipsitz, Stuart	113
Li, Hongzhe	30, 105	Little, Roderick	8a, 23, 57, 58
Li, Huilin	20	Liu, Aiyyi	33, 45
Li, Jia	3e	Liu, Anna	47
Li, Lingling	115	Liu, Ching-Ti	96
Li, Mingyao	44	Liu, Chunling	33, 45
Li, Pei	102	Liu, Chunyan	91
Li, Qin	111	Liu, Jun S	30
Li, Qizhai	45, 87	Liu, Lei	31, 47
Li, Runze	19, 46	Liu, Lian	88
Li, Ruosha	79	Liu, Li-yu D	3c
Li, Shaoyu	111	Liu, Lyrica Xiaohong	103
Li, Shi-Hwan	12d	Liu, Nianjun	89
Li, Shun H	8h, 9i	Liu, Pei	89
Li, Shun-Hwa	6f	Liu, Qing	54
Li, Wenjun	57	Liu, Suyu	54
Li, Xiaobo	20, 77	Liu, Tzu-Hsin	55
Li, Yan	57	Liu, Xiaole Shirley	74
Li, Yao	20, 80	Liu, Xue-Cheng	11c
Li, Ye	9f	Liu, Yin	99
Li, Yehua	27, 95	Liu, Yufeng	11b
Li, Yi	7h, 34, 72	Liu, Zhenyu	36
Li, Yimei	37, 56, 60	Liu, Ziyue	93
Li, Yisheng	64, 97, 101	Lizotte, Dan	106
Li, Zhiguo	2e	Lo, Kenneth	76
Liang, Faming	56	Lobach, Iryna V	89, 99
Liang, Hua	101	Logan, Brent R	37, 113
Liang, Kung-Yee	45	Long, Qi	23
Liang, Yuanyuan	91	Looney, Stephen	79
Liao, Xiaomei	7h	Lotz, Meredith J	92
Lim, Changwon	35	Louden, Christopher L	36
Lin, Dan1	15	Louis, David N	40
Lin, Danyu	41, 44, 87	Louis, Germaine M	46
Lin, Guixian	53	Louis, Thomas A	92, 107
Lin, Haiqun	22	Love, Tanzy M	78
Lin, Lanjia	113	Lu, Bo	81, 91
Lin, Rongheng	80	Lu, Shuya	3e
Lin, Shili	105	Lu, Wenbin	46, 53
Lin, Weili	60	Lu, Xiaomin	70
Lin, Xihong	R5, 27, 44, 112	Lu, Xuewen	114
Lin, Yanzhu	67	Lund, Steven P	55
Lindborg, Stacy	26	Luo, Li	77
Linder, Ernst	78	Luo, Zhehui	31
Lindquist, Martin A	60	Lyon, Roger	11c
Lindsay, Bruce G	82, 101	Ma, Shuangge	55
Linkletter, Crystal	73	Ma, Yan	12b
Lipscomb, John	35	Mahaffey, Kenneth	22

Mai, Yabing	92, 114	Moosman, Ann	69
Maity, Arnab	44	Morlock, Laura	90
Majumdar, Anandamayee	102	Morton, Leigh A	2h
Malloy, Betty J	92	Müller, Peter	23, 67, 88, 100
Manatunga, Amita	23, 99	Muller, Keith E	14c, 24, 34, 47, 56
Mandrekar, Sumithra J	1a	Munoz Maldonado, Yolanda	91
Manjourides, Justin	34	Murphy, Susan A	2e, 36, 69, 88, 106
Mao, Meng	51	Murray, Jared	78
Marcus, Sue	12b	Murray, Nicholas M	9j
Marion, Miranda C	4d	Murray, Susan	103
Maron, David J	8d	Musio, Monica	9l
Marron, J. S.	11b, 33	Muthen, Bengt O	22, 59
Marshall, Scott L	35	Myers, Jessica A	90
Marsit, Carmen J	40	Naesset, Erik	42
Martin, Clyde F	1f, 9j, 14e	Nan, Bin	8a, 41, 69, 114
Mason, Michael J	74	Nason, Martha	76
May, Susanne	80	Natarajan, Loki	58, 110
Mazumdar, Sati	78, 92	Neale, David B	77
McBean, Alexander M	102	Nelson, Heather H	40
McClure, Leslie A	110	Nelson, Jennifer C	115
McCulloch, Charles E	R1	Nelson, Ross F	42
McHenry, M. Brent	98	Nesvizhskii, Alexey I	63
McKenzie, David	1c	Nettleton, Dan	48, 55
McNally, James W	57	Neuhaus, Anneke	70
McNally, Richard J	1c	Neuhaus, John	29
McPeek, Mary Sara	67	Neustifter, Benjamin B	33
McRoberts, Ronald E	42	Neves, Carlos E	3b
Mehrotra, Devan V	22	Nichols, Thomas E	6b, 6c
Mehta, Shraddha S	7e	Nick, Todd G	91
Mengersen, Kerrie L	4e	Nicolae, Dan L	30
Merl, Daniel	63	Nielsen, Rasmus	30
Mermelstein, Robin J	47	Ning, Jing	97
Mertens, Karl	5b	Noe, Douglas A	80
Messer, Karen	88, 110	Normand, Sharon-Lise T	16, 107
Michailidis, George	67, 112	Nur, Darfiana	4e
Miclaus, Kelci	T3	Nychka, Douglas	SC4, 108
Miglioretti, Diana L	81	Obenchain, Valerie	87
Miller, Kelly	79	Ogle, Kiona	17
Miller, Sam	75	Oh, Sunghee	43
Miranda, Marie L	9g, 10a, 59	O'Finley, Andrew O	SC3
Miyahara, Sachiko	88	O'Hair, Joel C	6d
Miyakawa, Ayumi A	3b	O'Leary, Daniel	7i
Modarres, Reza	36	Oluyede, Broderick O	70
Mogg, Robin	22	O'Malley, A. James	16
Mohapatra, Gayatri	40	Ombao, Hernando	73
Molenberghs, Geert	R2, 21, 39, 69, 115	Onar-Thomas, Arzu	110
Montgomery, Robert	12d	O'Neill, Robert T	38
Moore, Janet	24	Paciorek, Christopher J	52

# INDEX

Padmanabhan, Krishna	54, 110	Presnell, Brett D	33
Pagano, Marcello	34	Proschan, Michael A	49
Paik, Jane	114	Psaty, Bruce M	7i
Pajewski, Nicholas M	20, 77	Pullenayegum, Eleanor M	31
Paladini, Carlos A	13e	Pungpapong, Vitara	67
Palta, Mari	58	Qaqish, Bahjat F	62
Pan, Qing	45, 92	Qian, Lei	14d
Pan, Wei	55, 67	Qian, Min	36
Pan, Yi	79	Qiao, Xingye	11b, 82
Papp, Jeanette C.	77	Qin, Guoyou	29
Pararai, Mavis	58	Qin, Jing	61
Park, Cheolwoo	93	Qin, Rui	1a
Park, Do-Hwan	51	Qin, Zhaohui S	63, 100
Park, Yong Seok	37	Qu, Annie	18
Parry, Sam	93	Quinlan, Judith A	28
Pashkevich, Maksim	115	Quinlan, Michelle	115
Patki, Amit	77	Quintana, Fernando A	88
Paul, Debashis	102	Radchenko, Peter	50
Payment, Pierre	79	Raftery, Adrian E	108
Peddada, Shyamal D	35, 48	Raghunathan, Trivellore E	8c, 8f, 9e, 84
Pedroza, Claudia	90	Ramos, Paula S	4d
Penalva, Luiz	100	Rao, Youlan	24
Pendergast, Jane	107	Rappold, Ana G	78
Pendergast, Mary K	107	Ratcliffe, Sarah	93
Peng, Gang	77	Rathbun, Stephen L	33
Peng, Hanxiang	79	Rathouz, Paul J	29
Peng, Jie	50	Redden, David T	2h
Peng, Limin	53	Reddy, Sasiragh P	2b
Peng, Roger D	T1, 85, 103	Reese, Peter P	7a
Peng, Yingwei	114	Regnier, Fred E	99
Pennello, Gene A	26, 65	Reich, Brian	59, 66, 85
Pepe, Margaret S	94	Reich, Nicholas	80
Pereira, Alexandre C	111	Reilly, Cavan	99
Peter, Gary F	77	Resnicow, Ken	2b
Peters, Ulrike	67	Ribeiro, Raul C	55
Pfeiffer, Ruth M	4c, 94	Richard, Otupei J	9b
Philip, Loni P	102	Richards, Adam J	100
Pieper, Karen S	22	Ringham, Brandy M	21
Platt, Robert W	86	Rizopoulos, Dimitris	70, 103
Pletcher, Scott D	7b	Robinson, Lucy F	60
Poisson, Laila M	67, 99	Rockette, Howard E	68
Polak, Joseph F	7i	Rodda, Bruce E	109
Polhamus, Daniel G	13e	Rohrer, Baerbel	100
Ponicki, William R	66	Rosen, Ori	93
Portnoy, Stephen	53	Rosenbaum, Paul R	45, 61
Pounds, Stanley B	20, 48, 55	Rosner, Gary L	26, 67, 81
Prado, Raquel	63	Roth, David L	7f
Preisser, John S	62	Rowe, Daniel B	6f

Roy, Anuradha	36	Shih, Ya-Chen T	31
Rui, Changxiang	78	Shim, Heejung	32
Sabo, Roy T	59, 79, 113	Shkedy, Ziv	115
Saha, Paramita	68	Shojaie, Ali	67
Saha, Sourish	75	Short, Margaret	66, 102
Sain, Stephan	SC4	Shults, Justine	7a, 62, 113
Salkowski, Nicholas J	70	Shun, Zhenming	54
Samawi, Hani M	79, 104, 113	Shyr, Yu	101
Samworth, Richard	19, 50	Siddique, Juned	81
Sánchez, Brisa N	2g, 14f	Siega-Riz, Anna Maria	59
Sargent, Daniel J	1a, 97	Sima, Cami	68
Sarkar, Sanat K	48	Simpson, Doug G	59
Sattar, Abdus M	8e, 21, 47	Simpson, Pippa M	6f, 8h, 9i, 12d
Scharfstein, Daniel O	39, 106	Simpson, Sean L	34
Schaubel, Douglas E	41	Sinha, Bimal K	9h, 32, 104
Schauberger, Eric	4b	Sinha, Debajyoti	72, 98, 113
Schildcrout, Jonathan S	29	Sitlani, Colleen	18
Schimek, Michael G	105	Slate, Elizabeth H	21
Schisterman, Enrique	45	Slavkovic, Aleksandra	10c
Schlesinger, Abigail	16	Slud, Eric V	92
Schmidler, Scott C	63	Small, Dylan	61
Schmoyer, Denise	99	Smith, Carmen J	12a
Schoenberg, Frederic P	34	Smith, Davey M	80
Schubert, Christine	68	Smith, Ning	72
Schucany, William R	2d, 6d, 6e	Smith, Richard	102
Schumacher, Martin	91	Soler, Julia M	111
Schwartz, Scott L	59	Soler, Julia P	3b
Schwenke, James	115	Solomon, Cam	7i
Scott, Laura J	20	Solorzano, Eleanne	32
Scribner, Richard	66	Soltani, Ahmad Reza	82
Seaman III, John W	21	Somasundaran, Mohan	80
Seffens, William	32	Song, Jiuzhou	100
Segawa, Eisuke	47	Song, Juhee	12c
Sen, Pranab K	35	Song, Peter X.K.	9e, 62, 93, 114
Serie, Daniel	67	Song, Seongho	32
Shao, Jun	54, 58	Sotres-Alvarez, Daniela	59
Shardell, Michelle D	24, 91	Sotto, Cristina	39
Sharp, Julia	43, 99	Speckman, Paul L	69
Shen, Dinggang	56	Speed, Terence	43
Shen, Haipeng	33	Speigelman, Donna	7h
Shen, Junshan	37	Spertus, John A	8d
Shen, Lei	77	Spiegelman, Donna	92
Shen, Xiaotong	55	Spinka, Christine	89
Shete, Sanjay	77	Spino, Cathie	2g
Shi, Weiliang	36	Stahl, Goran	42
Shi, Xiaoyan	56	Staicu, Ana-Maria	93
Shih, Joanna H	70	Stamey, James D	21
Shih, Tina	23, 31	Stanek III, Edward J	10g, 13c, 57, 104, 115

# INDEX

Stanhope, Stephen	9i	Thurston, Sally	78
Stanwyck, Elizabeth A	9h	Tian, Guo-liang	33
Stefanski, Leonard A	112	Tibshirani, Rob	40
Stein, Bradley	16	Tilahun, Abel E	21, 115
Stock, Shannon	80	Tiwari, Hemant K	44, 77
Stork, LeAnna G	35	Tiwari, Ram C	34, 72, 82
Stratton, Howard	115	Tobias, Randall D	75
Strawderman, Robert L	31	Tokdar, Surya T	7e
Stroup, Walt	115	Tolle, Jon	33
Styner, Martin	56	Tong, Xingwei	114
Su, Haiyan	101	Tosteson, Tor D	58
Su, Shu-chih	56	Tracey, Jeff	17
Sui, Yunxia	43	Trippa, Lorenzo	68
Sun, (Tony) Jianguo	8b, 37, 47, 51, 114	Tsai, Chih-Ling	46
Sun, Jianping	89	Tsai, Jerry W	63
Sun, Jie (Rena)	92	Tsai, Wei-Yann	103
Sun, Liuquan	51	Tseng, George C	3e, 43
Sun, Wenguang	48	Tseng, Yi-Kuan	51
Sun, Yanhui	7f	Tsiatis, Anastasios A	22, 23, 70
Sundaram, Rajeshwari	46	Tsodikov, Alex	7g, 79, 82, 103, 114
Swamy, Geeta	10a	Tsonaka, Roula	103
Swartz, Michael D	46, 77	Tsumagari, Koji	100
Swearingen, Christopher J	14a	Tu, Xin M	13d
Tadesse, Mahlet G	T2, 40	Tyers, Mike	63
Tan, Ming	100	Tzeng, Jung-Ying	96
Tan, Zhiqiang	92	Van Meter, Emily M	1e
Tang, Hui	43	Van Sanden, Suzy	115
Tang, Man Lai	33	VanderWeele, Tyler J	84
Tang, Wan	12b	VanDyke, Rhonda D	36
Tang, Xiaoqin	90	Vannucci, Marina	48, 93
Tarima, Sergey	11c	Vansteelandt, Stijn	5b, 84, 96
Tassone, Channing	11c	Vaughan, Roger D	2b
Taylor, Andrew	23	Verbeke, Geert	39, 69, 103
Taylor, Jeremy M.G.	72, 93	Vickers, Andrew J	94
Tebbs, Joshua M	113	Viele, Kert	36
Telesca, Donatello	67, 100	Villagran, Alejandro	93
Ten Have, Thomas R	22, 61	Vogel, Robert L	79, 113
Terrell, George R	113	Von Wilpert, Klaus	9l
Teuschler, Linda K	35	Wager, Carrie	115
Thall, Peter F	106	Wager, Tor D	6b, 6c, 60
Thomas, Duncan C	52	Wagler, Amy E	14b
Thomas, Fridtjof	20	Wahba, Grace	36
Thomas, Laine E	70	Wahed, Abdus S	88, 106
Thomas, Ly	41	Waldmann, Patrik	85
Thomasson, Arwin M	7a	Wall, Melanie M	70
Thometz, John	11c	Wallace, Dorothy	9j
Thompson, Theodore J	10e, 11d	Waller, Lance A	66, 71
Thompson, Wesley K	93	Wang, Chenguang	39, 103

Wang, Chen-pin	22	Williams, Calvin L	33
Wang, Chi	92	Williams, Calvin	8h
Wang, Chia-Ning	8a	Wills-Karp, Marsha	4b
Wang, Cuiling	24	Wilson, Charles J	13e
Wang, Hansheng	46	Wing, Jeffrey J	2g
Wang, Hao	58	Wittes, Janet	38
Wang, Hongkun	90	Wolf, Bethany J	21
Wang, Huixia Judy	53	Wolfinger, Russ	T3
Wang, Jane-Ling	51	Wolfson, Julian	41
Wang, Lan	19, 53, 101	Wood, Simon N	9l
Wang, Li	18	Woodard, Dawn	81
Wang, Lianming	35	Woodward, Wayne A	6d, 6e
Wang, Libo	67	Wrensch, Margaret	40
Wang, Mei-Cheng	25	Wright, Fred A	44
Wang, Ming	113	Wright, Stephen E	80
Wang, Nae-Yuh	27	Wu, Dongfeng	81
Wang, Naisyin	27, 43, 105	Wu, Hao	74
Wang, Ning	1f	Wu, Jincao	6a
Wang, Pei	50	Wu, Jun	115
Wang, Shufang	79	Wu, Louie R	20
Wang, Sue-Jane	38, 49	Wu, Meihua	14f
Wang, Tao	89	Wu, Michael C	40, 112
Wang, Xia	98	Wu, Rongling	4a, 20, 111
Wang, Xiaodong	79	Wu, Rongling	80, 89
Wang, Xinlei	100	Wu, Xiaoru	91
Wang, Xueqin	96	Wu, Yichao	11g, 19
Wang, Yaping	89	Wu, Yuan	25
Wang, Yishi	13a	Wu, Zhijin	43
Wang, Yu	91	Xia, Amy	26
Wang, Yuanjia	111, 112	Wurttele, Eve	15
Wang, Zuoheng	67	Xiang, Qinfang	24
Wathen, Kyle	88	Xiao, Guanghua	100
Watson, Charity N	33	Xiao, Rui	20
Wei, Fengrong Wei	69, 112	Xiao, Yuanhui	48
Wei, Peng	67	Xiao, Zhiguo	58
Weintraub, William S	8d, 90	Xie, Jichun	33, 113
Weiss, Robert	14d	Xie, Jieru	69
Weissfeld, Lisa	8e, 21	Xie, Yang	43
Wheeler, Matthew W	59, 66	Xiong, Chengjie	23, 104
White, Kristin L	67	Xiong, Momiao	77
Wiemels, Joseph	40	Xiong, Zang	110
Wiencke, John K	40	Xu, Bo	13c
Wikle, Christopher K	R4, 17	Xu, Dihua	104
Wilding, Gregory E	1d	Xu, Lei	6b
Wileyto, Paul E	37	Xu, Ruifeng	58
Wilkins, Kenneth J	8g	Xu, Zhenzhen	110
Willan, Andrew R	31	Xu, Zhiying	104
Williams, Andre	112	Xue, Lan	18

# INDEX

Yan, Ke	6f, 8h, 9i, 12d	Zhang, Min	22
Yanez, David	7i	Zhang, Min	67, 99
Yang, Harry	91, 97	Zhang, Nanhua	2b
Yang, Mei	20	Zhang, Peng	69
Yang, Yan	59	Zhang, Rongmei	75
Yang, Yarong	55	Zhang, Ruitao	57
Ye, Wen	10b	Zhang, Song	23, 58
Yeatman, Timothy	55	Zhang, Wei	1a
Yeh, Ru-Fang	40	Zhang, Wei	8d
Yi, Grace Y	38, 51	Zhang, Xinyan	37
Yi, Nengjun	44	Zhang, Ying	25, 101
Yin, Guosheng	54	Zhang, Yiyun	46
Ying, Zhiliang	114	Zhang, Yong	8f
Yoo, Jae Keun	79	Zhang, Yu	30
Yoon, Frank B	45	Zhang, Yufen	1a
Yu, Kai F	45, 101	Zhang, Zhiwei	88
Yu, Kai	45, 87	Zhang, Zhongfa	55
Yu, Lili	113	Zhang, Zugui	90
Yu, Menggang	18	Zhao, Hongwei	90
Yu, Qin	12b	Zhao, Hongyu	32, 99
Yu, Qingzhao	11f, 66	Zhao, Meng	103
Yu, Shuli	10g	Zhao, Qi	89
Yu, Tianwei	99	Zhao, Shanshan	115
Yuan, Xing	104	Zhao, Weiyan	89
Yuan, Ying	23, 39, 54	Zhao, Xiaoyue	43
Zeger, Scott	34	Zhao, Xingqiu	51
Zeng, Donglin	27, 41, 44, 68, 72, 110	Zhao, Yichuan	103
Zeng, Zhao-Bang	89	Zhao, Yufan	110
Zerbe, Gary O	47	Zheng, Jin	44
Zhang, Biao	61	Zheng, Qi	32, 80
Zhang, Bin	8i,	Zheng, Yanbing	9d
Zhang, Bin	114	Zheng Yingye	18
Zhang, Dabao	67, 99	Zhong, Ming	77
Zhang, Daowen	18	Zhou, Bingqing	37
Zhang, Hao	46	Zhou, Gongfu	111
Zhang, Heping	96	Zhou, Haibo	29
Zhang, Heping	111	Zhou, Mai	46
Zhang, Hongmei	57	Zhou, Nengfeng	50
Zhang, Huaiye	69	Zhou, Qing	74
Zhang, Hui	13d, 35	Zhou, Tianyue	33
Zhang, Jiajia	114	Zhu, Bin	93
Zhang, Jing	30	Zhu, Hong	25
Zhang, Jing	80, 102	Zhu, Hongjian	97
Zhang, Lanju	91, 97	Zhu, Hongtu	56, 60, 69
Zhang, Li	89, 111	Zhu, Ji	50, 112
Zhang, Lingsong	36	Zhu, Jian	8c
Zhang, Mei-Jie	37	Zhu, Jun	9d, 17
Zhang, Meizhuo	111	Zhu, Li	66

Zhu, Liang	47
Zhu, Wensheng	96
Zhu, Yanni	55
Zhu, Zhengyuan	102
Zou, Fei	44
Zou, G. Y.	104
Zubovic, Yvonne M	33
Zucker, David	7h

## Advancing science. Serving patients.



Amgen, a biotechnology pioneer, discovers, develops and delivers innovative human therapeutics. Our medicines have helped millions of patients in the fight against cancer, kidney disease, rheumatoid arthritis and other serious illnesses.

With a deep and broad pipeline of potential new medicines, Amgen remains committed to advancing science to dramatically improve people's lives.

To learn more about Amgen, our vital medicines, our pioneering science, and our career opportunities, visit [www.amgen.com/careers](http://www.amgen.com/careers).

**AMGEN®**

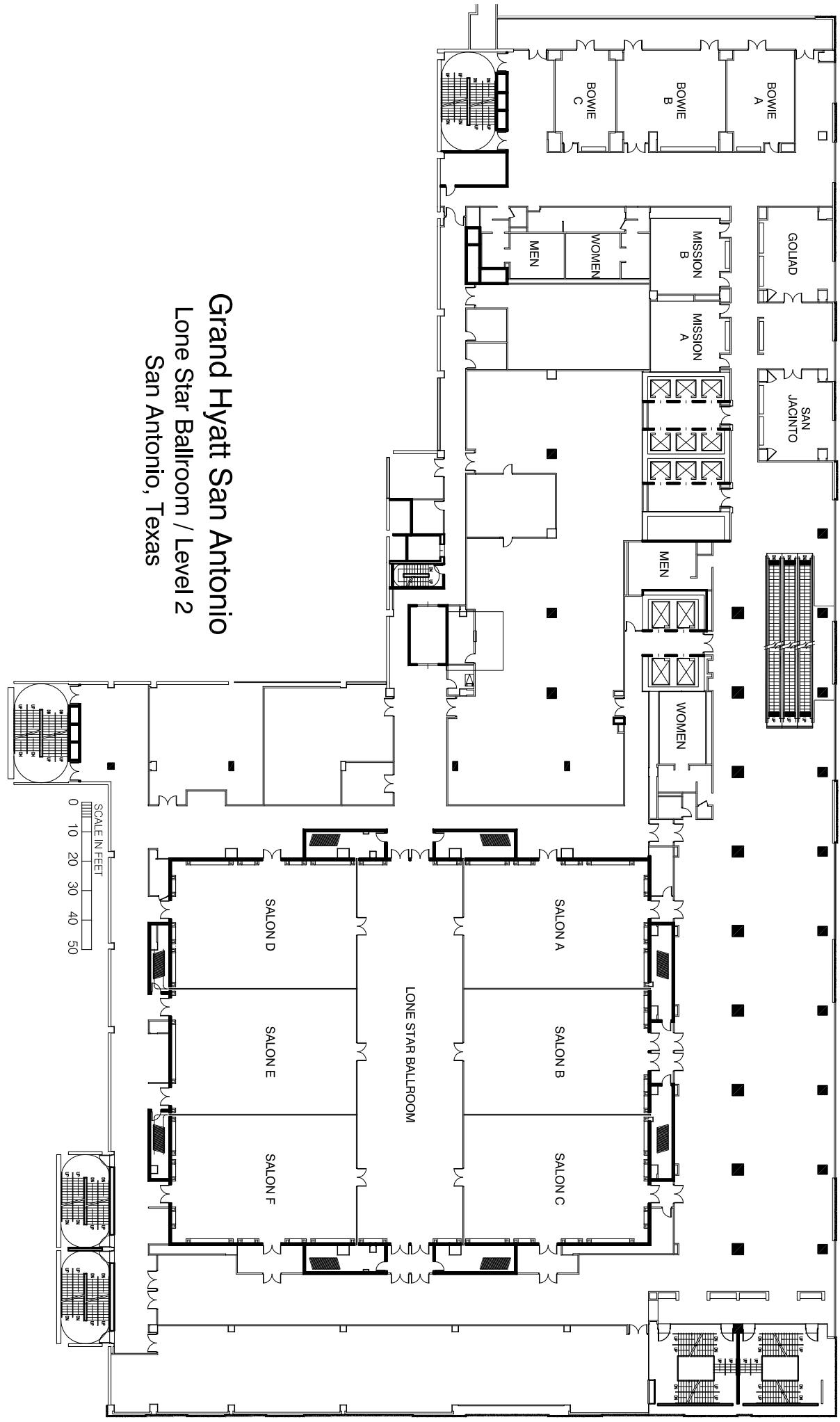
Pioneering science delivers vital medicines™

[www.amgen.com/careers](http://www.amgen.com/careers)

As an EEO/AA employer, Amgen values a diverse combination of perspectives and cultures. M/F/D/V.

EVERY EFFORT HAS BEEN MADE TO ENSURE THE ACCURACY OF ALL INFORMATION CONTAINED ON THIS FLOOR PLAN.  
HOWEVER THIS FLOOR PLAN IS TENTATIVE UNTIL ACTUAL INSPECTION OF THE FACILITY HAS BEEN COMPLETED.

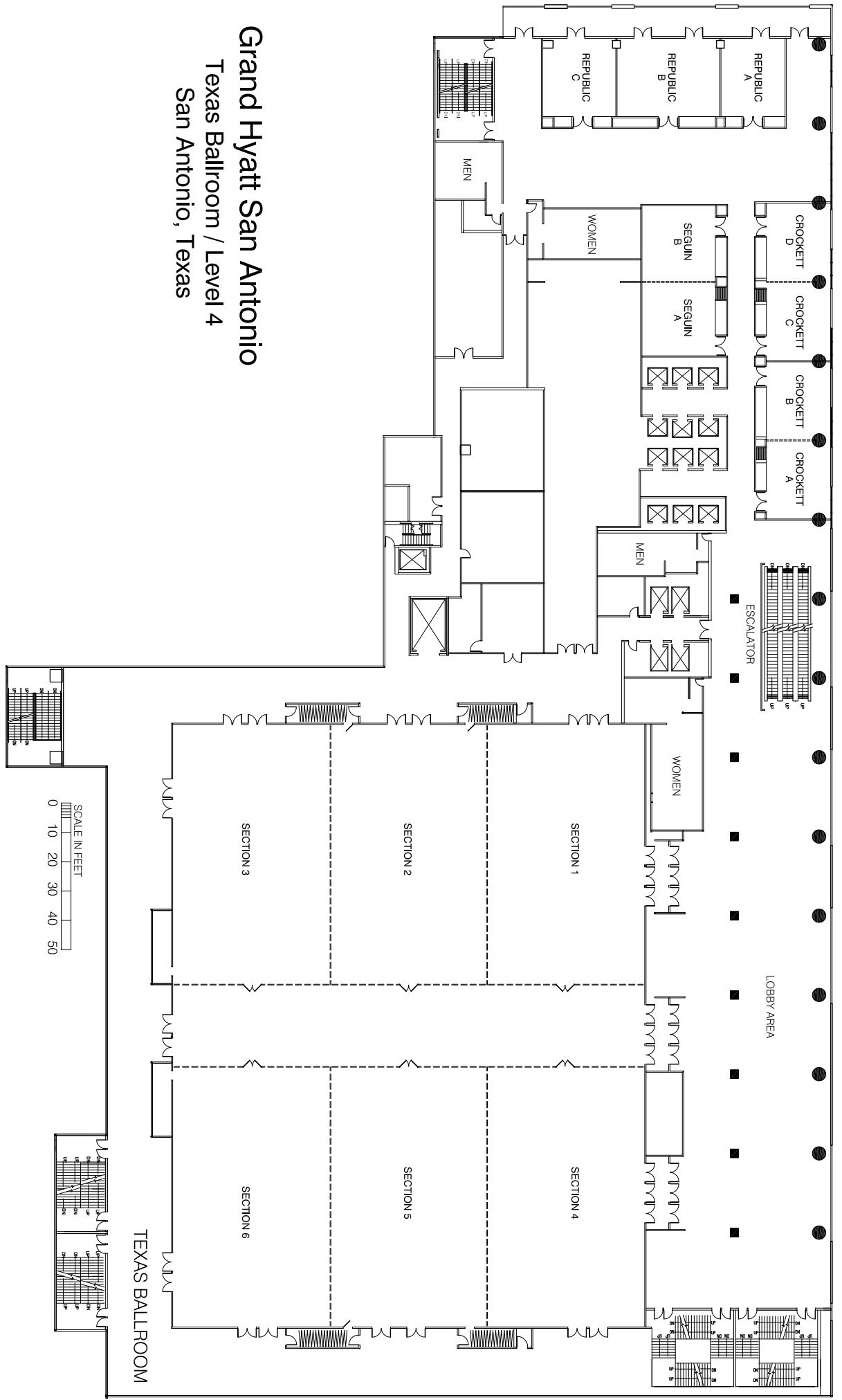
**F R E E M A N**



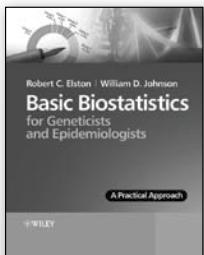
EVERY EFFORT HAS BEEN MADE TO ENSURE THE ACCURACY OF ALL INFORMATION CONTAINED ON THIS FLOOR PLAN.  
HOWEVER THIS FLOOR PLAN IS TENTATIVE UNTIL ACTUAL INSPECTION OF THE FACILITY HAS BEEN COMPLETED.

**F R E E M A N**

**Grand Hyatt San Antonio**  
**Texas Ballroom / Level 4**  
**San Antonio, Texas**



# DISCOVER THE WORLD OF STATISTICS AT ITS BEST FROM Wiley-Blackwell



## Basic Biostatistics for Geneticists and Epidemiologists

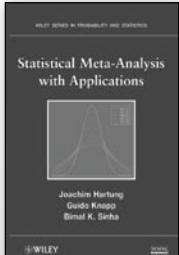
A Practical Approach

**ROBERT C. ELSTON and WILLIAM D. JOHNSON**

December 2008 • 384 pages

Paperback • 978-0-470-02490-4 • \$50.00

Hardback • 978-0-470-02489-8 • \$130.00

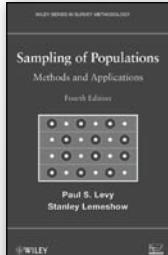


## Statistical Meta-Analysis with Applications

**JOACHIM HARTUNG, GUIDO KNAPP and BIMAL K. SINHA**

August 2008 • 248 pages

Hardback • 978-0-470-29089-7 • \$94.95



## Sampling of Populations

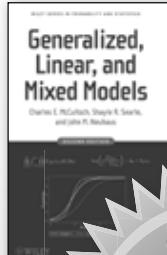
Methods and Applications

Fourth Edition

**PAUL S. LEVY and STANLEY LEMESHOW**

August 2008 • 576 pages

Hardback • 978-0-470-04007-2 • \$130.00



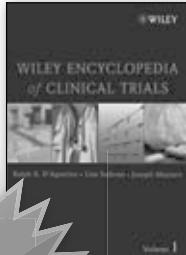
## Generalized, Linear, and Mixed Models

Fourth Edition

**CHARLES E. MCCULLOCH, SHAYLE R. SEARLE, and JOHN M. NEUHAUS**

June 2008 • 384 pages

Hardback • 978-0-470-07371-1 • \$99.95



**Save  
20% Off  
Books on  
Display**

## Wiley Encyclopedia of Clinical Trials

4-Volume Set

**Edited by: RALPH D'AGOSTINO, LISA SULLIVAN, and JOSEPH MASSARO**

November 2008 • 2524 pages

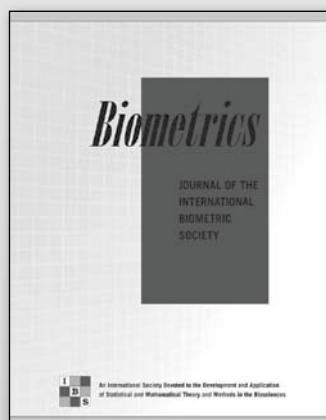
Hardback • 978-0-471-35203-7

\$1,400.00 Special Offer! \$1,100.00!

For more titles, visit our booth at ENAR, or browse and order online at [www.wiley.com/statistics](http://www.wiley.com/statistics).

To order, in North America call 1-877-762-2974 or in Rest of World call +44 (0) 1243 843294.

Save 20% at the Wiley booth at ENAR, or order online using promotion code 97337.



# Biometrics

Published on behalf of the  
International Biometric Society

**Co-Editors: THOMAS A. LOUIS, GEERT MOLENBERGHS, DAVID ZUCKER**

**Executive Editor: MARIE DAVIDIAN**

*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 a scientific application that sets scientific or policy objectives, motivates methods development, and demonstrates the operations of new methods.

**Never miss out again on the latest breakthroughs in biometric methodology and practice!**

Keep up to date with E-mail Table of Contents Alerts delivered direct to your desktop – FREE! Simply visit the *Biometrics* homepage and click on ‘Sign up for e-alerts’ and we’ll alert you each time and issue is published.

[www.blackwellpublishing.com/BIOM](http://www.blackwellpublishing.com/BIOM)



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 IBS is the parent society of ENAR, the Eastern North American Region of the IBS.

 **WILEY-BLACKWELL**



# NOTES



