

International Biometric Society

Eastern North American Region

PROGRAM & ABSTRACTS

2
0
0
4

Spring Meeting with IMS and
Sections of ASA
March 28 – 31, 2004
Hilton Pittsburgh
Pittsburgh, PA



World Leader in Biotechnology



Using science and innovation to dramatically improve people's lives.

For more than two decades, Amgen has helped lead a revolution in human therapeutics – a revolution grounded in science, powered by new technologies and conducted at the molecular level. But the achievement of Amgen's greatest ambition still lies ahead. For we are now intent on building an organization that combines the scientific knowledge, the collaborative process, and the human spirit necessary to be the best human therapeutics company in the world.

With over 5,500 employees at our corporate headquarters in Southern California, Amgen is tackling some of the world's most challenging unmet medical needs. Our employees thrive on fresh ideas, hard work and a shared commitment to excellence. Despite a vigorous growth pace, Amgen actively fosters a small-company informality. It is a collegial atmosphere that encourages people to both work hard and have fun.

As our Research & Development group grows this year, we need more qualified candidates for our **Biostatistics** area. To view a comprehensive listing of employment opportunities and to apply online, please visit the Careers page on our website at www.amgen.com/career. Our outstanding compensation package features excellent benefits and relocation assistance. Learn more about Amgen and the communities we serve at www.amgen.com. As an EEO/AA employer M/F/D/V, Amgen values and encourages diversity in the workplace.

Stop by our booth!

**Representatives are here at the event to meet with you
and discuss employment opportunities at Amgen.**



PITTSBURGH 2004: EXPANDING THE FOCUS OF BIOMETRICS

Acknowledgements	4
Officers and Committees	5
Programs and Representatives	6
Workshop for Junior Researchers	7
ENAR Student Award Committee	7
Student Award Winners	7
Future Meetings of the International Biometric Society	7
Special Thanks	8
Short Courses	9 - 10
Tutorials	11 - 12
Roundtable Luncheons	13 - 14
Program Summary	15 - 18
Notes	19
Scientific Program	20 - 51
Poster Session Summary	52 - 53
Notes	54
Abstracts	55 - 299
Poster Session	300 - 314
Notes	315
Index of Participants	316 - 325
Notes	328 - 330
Hilton Pittsburgh Floor Plan	331

ACKNOWLEDGEMENTS

EXHIBITORS

Amgen
Blackwell Publishing, Inc.
Cambridge University Press
CRC Press – Taylor and Francis Group
Cytel Software Corporation
Duxbury, a part of the Thomson Corporation
Insightful Corporation
Kluwer Academic Publishers
MedFocus, inc.
Oxford University Press
SAS Institute
SIAM (Society for Industrial and Applied Mathematics)
Smith Hanley Associates
Springer-Verlag New York, Inc.
StatSoft, Inc.
Wiley

SPONSORS

We gratefully acknowledge the support of:

Amgen
AstraZeneca
Bristol-Myers Squibb
GlaxoSmithKline
ICON Clinical Research
Inspire Pharmaceuticals, Inc.
Johnson & Johnson Pharmaceutical Research & Development
Merck Research Laboratories
Millenium Pharmaceuticals, Inc.
PPD, Inc.
Rho, Inc.
SAS Institute
Schering-Plough Research Institute
Statistics Collaborative, Inc.
The Emmes Corporation

We are indebted to Eli Lilly and Company for substantial support of the preparation of the final scientific program listing.

OFFICERS AND COMMITTEES

Executive Committee – Officers

President	Marie Davidian
Past President	Timothy G. Gregoire
President-Elect	Peter Imrey
Secretary (2003-2004)	Sharon-Lise Normand
Treasurer (2004-2005)	Joanna Shih

Regional Committee (RECOM)

President (Chair)	Marie Davidian
Six ordinary members (elected to 3-year terms): + Stacy Lindborg (RAB Chair)	

2002-2004

Julie Legler
Lance Waller

2003-2005

Stephen George
Roderick Little

2004-2006

Bruce Craig
Amita Manatunga

Regional Members of the Council of the International Biometric Society

Marie Davidian, Walter W. Piegorsch, Ron Brookmeyer, Louise Ryan, Janet Wittes

Appointed Members of Regional Advisory Board (3-year terms)

Chair: Stacy Lindborg

2002-2004

Mark Conaway
Jason Fine
Susan Halabi
Elizabeth Margosches
Sati Mazumdar
Myunghee Cho Paik
Yuko Palesch
Jane Pendergast
Dionne Price
Joanna Shih

2003-2005

Scarlett Bellamy
Christopher R. Bilder
DuBois Bowman
James J. Chen
Michael Daniels
Francesca Dominici
Montserrat Fuentes
Cynthia Garvan
Brian D. Marx
Alicia Y. Toledano

2004-2006

Hongshik Ahn
Brent Coull
Debashis Ghosh
Amy Herring
Tom Loughin
Jared Lunceford
Jeffrey Morris
Kerrie Nelson
Frank Roesch
Helen Zhang

PROGRAMS

2004 Joint Statistical Meeting

Paul Rathouz

2004 Spring Meeting – Pittsburgh, PA

Local Arrangements Chairs: Joyce Chang and Ada Youk

Program Chair: Thomas R. Ten Have

2005 Joint Statistical Meeting

Naisyin Wang

2005 Spring Meeting – Austin, TX

Program Chair: A. John Bailer

Program Co-Chair: Maura Stokes

Biometrics Editors

Brian Cullis, Mike Kenward, and Xihong Lin

Biometric Bulletin Editor

Thomas R. Ten Have

ENAR Correspondent for the *Biometric Bulletin*

Rosalyn Stone

ENAR Executive Director

Kathy Hoskins

International Biometric Society Business Manager

Claire Shanley

REPRESENTATIVES

Committee of Presidents of Statistical Societies (COPSS)

ENAR Representatives

Marie Davidian (President)

Timothy G. Gregoire (Past-President)

Peter Imrey (President-Elect)

ENAR Standing/Continuing Committee Chairs

Nominating

Timothy G. Gregoire

Sponsorship

Frank Shen

Information Technology Oversight
(ITOC)

Stewart Anderson

American Association for the Advancement of Science (Joint with WNAR)

Terms through February 22, 2005

Section N, Medical Sciences

Karen Bandeen-Roche

Section G, Biological Sciences

Zhao-Bang Zeng

Section U, Statistics

Aparna Huzurbazar

Section O, Agriculture

Bahman Shafii

National Institute of Statistical Sciences

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

Member: Marie Davidian

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

WORKSHOP FOR JUNIOR RESEARCHERS

Francesca Dominici, (Chair)
F. DuBois Bowman
Christina Kendziorski
Paul Rathouz

ENAR STUDENT AWARD COMMITTEE

Chair: Carol Redmond
Karl Broman
Xihong Lin
Sharon-Lise Normand
Walter Piegorsch
Dionne Price
Daniel Schnell
Naisyin Wang
Lisa Weissfeld
Daniel Zelterman
Dale Zimmerman

STUDENT AWARD WINNERS

VAN RYZIN AWARD WINNER

Ming Yuan, University of Wisconsin-Madison

AWARD WINNERS

Sungyoung Auh, University of Pittsburgh
Saonli Basu, University of Washington
Dachuang Cao, Purdue University
Guoqing Diao, University of North Carolina
Chiu-Hsieh Hsu, University of Michigan
Lan Huang, University of Connecticut
Xiaohong Huang, University of Minnesota
Xiaoping Jin, University of Minnesota
Erning Li, North Carolina State University

Huiling Li, Columbia University
Rongheng Lin, Johns Hopkins University
Lei Liu, University of Michigan
Li Liu, University of Illinois at Chicago
Li Qin, University of Pennsylvania
Samiran Sinha, University of Florida
Ji-Ping Wang, Northwestern University
Menggang Yu, University of Michigan
Jiameng Zhang, University of Pennsylvania
Jing Zhao, University of Pennsylvania

FUTURE MEETINGS OF THE INTERNATIONAL BIOMETRIC SOCIETY

2004 International Biometric Conference
Cairns, Australia

2006 International Biometric Conference
Montreal, Canada

2005 ENAR Spring Meetings
Austin, TX

2006 ENAR Spring Meetings
New Orleans, LA

2007 ENAR Spring Meetings
Miami, FL

SPECIAL THANKS

2004 ENAR Program Committee

Tom Ten Have (Chair), University of Pennsylvania
Stacy Lindborg (Co-Chair), Eli Lilly & Company
Kate Cowles, University of Iowa
Bruce Craig, Purdue University
Philip Dixon, Iowa State University
Susan Halabi, Duke University Medical Center
Philip Iversen, Eli Lilly & Company
Danyu Lin, University of North Carolina
Susan Murray, University of Michigan
Ed Vonesh, Baxter Healthcare
Heping Zhang, Yale University

ASA Section Representatives

Ralph D'Agostino, Jr., Wake Forest University
Michael Elliott, University of Pennsylvania
Dianne Finkelstein, Harvard University
Anna Nevius, U.S. Food and Drug Administration
Estelle Russek-Cohen, University of Maryland
Sue Sereika, University of Pittsburgh
Fei Wang, Boston University
Linda Young, University of Florida

ENAR Education Advisory Committee

Brad Carlin, University of Minnesota
David Dunson, National Institute of Environmental Health Sciences
Tim Gregoire, Yale University
Xihong Lin, University of Michigan
Eva Petkova, Columbia University
Tom Ten Have, University of Pennsylvania
Ji Zhang, Sanofi-Synthelabo

ENAR Workshop for Junior Researchers

Francesca Dominici (Chair), Johns Hopkins University
F. DuBois Bowman, Emory University
Christina Kendziorski, University of Wisconsin
Paul Rathouz, University of Chicago

ENAR Local Arrangements

Joyce Chang, University of Pittsburgh
Ada Youk, University of Pittsburgh

IMS Program Chair

Daniel Nicolae, University of Chicago

ENAR Student Awards

Carol Redmond, University of Pittsburgh

2
0
0
4

2
0
0
4

ENAR SHORT COURSES

SCI: Hierarchical Modeling and Analysis for Spatial Data (Full Day) *Sterlings*

Instructors: Sudipto Banerjee (University of Minnesota), Brad Carlin (University of Minnesota), and Alan Gelfand (Duke University)

Description:

As recently as two decades ago, the impact of hierarchical Bayesian methods outside of a small group of theoretical probabilists and statisticians was minimal at best. Suddenly, around 1990, the “Markov chain Monte Carlo (MCMC) revolution” enabled hierarchical models to become not only feasible, but the preferred approach for almost any model involving multiple levels incorporating random effects or complicated dependence structures.

In this course we will describe hierarchical modeling methods for an area of application in which they can pay substantial dividends: spatial statistics. We will begin by outlining and providing illustrative examples of the three types of spatial data: point-level (geostatistical), areal (lattice), and spatial point process. We then describe both exploratory data analysis tools and traditional modeling approaches for point-referenced data. Modeling approaches from traditional geostatistics (variogram fitting, kriging, and so forth) will be covered here. We shall then offer a similar presentation for areal data models, again starting with choropleth maps and other displays and progressing towards more formal statistical concepts, such as Brook’s Lemma and the Markov random field topics that underlie the conditional, intrinsic, and simultaneous autoregressive (CAR, IAR, and SAR) models so often used in areal data settings.

The remainder of our presentation will cover hierarchical modeling for both univariate and multivariate spatial response data, including Bayesian kriging and lattice modeling, as well as more advanced issues such as anisotropy and nonstationarity. Bayesian methods will also be suggested for modeling data that are spatially misaligned (say, with one variable measured by census tract but another by zip code), since they are particularly well-suited to sorting out complex interrelationships and constraints. We conclude with a brief discussion of spatially varying coefficient, spatio-temporal, and spatial survival models. Computer implementation for our models via the WinBUGS, R, and S+SpatialStats packages will be described throughout our presentation as appropriate.

The presenters are authors of the forthcoming book “Hierarchical Modeling and Analysis of Spatial Data” (CRC Press, 2004), upon which much of the course material will be based.

Prerequisites:

Short course participants should have an M.S. understanding of mathematical statistics at, say, the Hogg and Craig (1978) level, as well as basic familiarity with standard statistical models and computing. We will not assume any significant previous exposure to

DATE: Sunday, March 28, 2004

Full Day Fee

Members	\$190 (\$215 after 2/20)
Nonmembers	\$235 (\$260 after 2/20)

Short Course Registration

Saturday, March 27	3:00 - 5:00 p.m.
Sunday, March 28	7:00 - 8:30 a.m.

Course

Sunday, March 28 (lunch on your own)	8:30 a.m. - 5:00 p.m. (full day)
---	----------------------------------

spatial or Bayesian methods, although students with basic knowledge of these areas (say, based on the books by Cressie, 1993, and Carlin and Louis, 2000, respectively) will certainly face a gentler learning curve.

SC2: Semiparametric Regression (Full Day) *Ballroom 4*

Instructors: David Ruppert (Cornell University) and Matt Wand (University of New South Wales)

Description:

Parametric regression involves fitting a curve to a data set within the confines of a parametric families: for example, lines, parabolas and exponentials. Nonparametric regression, often called smoothing, only imposes the condition that the curve be smooth; the shape of the curve depends primarily upon the data. Semiparametric regression combines nonparametric and parametric models. For example, in the analysis of time series data on mortality and air pollution it is common for the air pollution effects to be modeled linearly, and confounders such as time, temperature and humidity to be nonparametric.

There are many methods for nonparametric estimation including kernel regression, local polynomial regression, smoothing splines, regression splines, and wavelets. However, the method of penalized splines has a number of convenient features such as:

- ease of use
- modularity, meaning that each component of the model can be developed separately and in a simple fashion
- close connections with parametric statistics so that maximum likelihood estimation and likelihood ratio tests can be used

ENAR SHORT COURSES

- ability to be implemented with standard software, e.g., SAS, S-PLUS, and WinBUGS

The course will focus on semiparametric models based on penalized splines viewed as linear mixed models (LMMs), or, for binary, Poisson, and other non-Gaussian outcomes, as generalized linear mixed models (GLMMs). Topics include

- scatterplot smoothing, especially by penalized splines
- LMMs and penalized splines as best linear unbiased predictors
- automatic selection of the degree of smoothing, especially estimation of the penalty parameter of a spline by restricted maximum likelihood
- additive models
- inference
- generalized semiparametric regression and GLMMs
- survival analysis
- spatial statistics
- measurement error
- Bayesian semiparametric regression
- programming in SAS, S-PLUS, and WinBUGS.

The presenters are authors of the text “Semiparametric Regression” (with R.J. Carroll, Cambridge University Press, 2003), which will form the basis for much of the course material.

Prerequisites:

Essential: Basic statistical theory, Regression analysis, Matrix notation and algebra. Desirable: Familiarity with mixed models.

SC3: Statistical Analysis with Missing Data (Full Day)

[*Ballroom 3*](#)

Instructors: Rod Little and Trivellore Raghunathan (both at University of Michigan)

Description:

Missing data is a pervasive problem in statistical analysis. Methods for handling the problem will be discussed by two authorities on the subject. The course will be based on the recently-released 2nd edition of Little and Rubin’s classic text, “Statistical Analysis with Missing Data”. Topics include introduction and overview, nonresponse weighting, principles of imputation, multiple imputation, maximum likelihood methods including EM, Bayesian simulation methods, nonignorable missing data, missing data software. Common missing data problems, such as regression with missing covariates and repeated-measures analysis with missing values, will also be discussed. The course will include illustrations with real data sets.

Prerequisites:

The course is intended for the applied statistician and will not be heavily theoretical. Some familiarity with basic principles of inference based on likelihoods is assumed, as well as familiarity with standard models of complete-data analysis, such as the normal linear model, multinomial models for counted data and the properties of standard statistical distributions, especially the multivariate normal.

SC4: Practical Approaches to Data Monitoring of Clinical Trials (Full Day)

[*Kings Garden North & South*](#)

Instructors: David DeMets (University of Wisconsin), Susan Ellenberg (Center for Biologics Evaluation and Research, FDA), and Thomas Fleming (University of Washington)

Description:

Data monitoring committees (DMCs) (also commonly known as data and safety monitoring boards or other similar titles) are increasingly used for monitoring accumulating data in clinical trials. Such committees, separate from trial organizing committees, steering committees, sponsors and investigators, are established to enhance the protection of patients enrolled in the trial by ensuring objective oversight of experts who have no vested interest in the trial or in the product under study.

Because the accumulating data reviewed by DMCs are considered highly confidential, few individuals have opportunities to attend or participate in DMC meetings. The objective of this workshop is to provide an overview of the DMC process, including the types of trials most in need of DMC oversight, criteria for membership, typical responsibilities and operating procedures, statistical approaches to monitoring and how they are implemented by DMCs, regulatory issues, and currently unsettled or controversial topics. The material will be illustrated by many real examples of DMC decision-making. Substantial time will be reserved for questions and discussion.

The presenters are the authors of the recent text, “Data Monitoring Committees in Clinical Trials: A Practical Perspective” (John Wiley, 2002), which will be the basis for much of the material presented.

Prerequisites:

Useful to have some experience with clinical trials and some background with sequential monitoring procedures, but the content will primarily be more conceptual than statistical.

ENAR 2004 TUTORIALS

T1: An Introduction to the R Environment

[Ballroom 4](#)

Date: Monday, March 29

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

Instructor: Peter Dalgaard (University of Copenhagen)

Description:

The R language and environment is an open source implementation of the well-known S language. It adds to the original design in various ways, but most importantly, the simple fact that it is freely available and easily deployable makes it a very attractive option for teaching, practical statistical analysis, and methodological development on many levels.

The tutorial will cover the following areas:

- (1) A simple interactive session showing basic language features.
- (2) Modeling tools. The notion of a fitted model object. Model formulas.
- (3) Using R packages. Examples of functionality provided. The CRAN repository.
- (4) Practical analysis. Input of data, workspaces, storage. Loading packages. Attaching and detaching data sets.
- (5) Graphics. Standard R graphics (the ink-on-paper model). Graphical coordinate units and layout. Standard plots and ad-hoc modifications. If time permits, also discuss Grid graphics and the Lattice package.
- (6) Ad-hoc programming: matrix algebra and optimization tools.

Professor Dalgaard is the author of the recent Springer monograph "Introductory Statistics with R."

Prerequisites:

It will be assumed that the audience has been acquainted with some statistical software, has a masters level knowledge of statistics plus some feeling for the issues involved in practical statistical analysis. The tutorial does not assume prior experience with R or S-PLUS. Some knowledge of computer programming may be advantageous, but not crucial.

T2: Experiments and Observational Studies: Casual Inference in Statistics

[Ballroom 4](#)

Date: Monday, March 29

Time: 1:45-3:30 p.m.

Instructor: Paul Rosenbaum (University of Pennsylvania)

Description:

The effect caused by a treatment is a comparison of the two responses a subject might exhibit if the treatment were applied or withheld. Because at any one moment, for any one subject, the treatment is either applied or withheld, at most one of these two responses is observed, so causal effects cannot be calculated for individuals. This fact led generations of philosophers to claim that causality is a matter of speculation. Remarkably, Sir Ronald Fisher derived a tight logic for inference about the effects of treatments on finite populations of subjects, as opposed to individual subjects, when treatments are assigned to subjects at random, as they are today in clinical trials. Beginning with the literature of the 1920s and 1930s on randomized experiments, this tutorial will review the contributions of Fisher, Neyman, Cochran, Rubin, Cornfield, Hill, Campbell, and others, and will introduce recent developments for observational studies in which randomization is not feasible. Topics include: randomization as the "reasoned basis for inference" about causal effects, propensity scores for overt biases, sensitivity analyses for hidden biases, attributable effects, and design considerations in observational studies. Concepts and techniques will be illustrated using data from several randomized trials and observational studies.

Prerequisites:

Basic knowledge of probability and statistics. It would be good, but not essential, to have seen the hypergeometric distribution, Fisher's exact test for a 2x2 table, Wilcoxon's signed rank test, and odds ratios. Familiarity or experience with empirical studies of human subjects would also be helpful.

T3: Generalized and Nonlinear Mixed Models in the SAS System: Tools and Applications

[Ballroom 4](#)

Date: Tuesday, March 30

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

Instructor: Oliver Schabenberger (SAS Institute Inc.)

Description:

This tutorial discusses tools in SAS/STAT software for fitting generalized linear and nonlinear mixed models. Classes of mixed models and estimation methods are discussed along with the associated SAS procedures and macros. The tutorial focuses on the MIXED and NL MIXED procedures, and the %GLIMMIX and %NLINMIX macros. Participants will hear about the advantages and disadvantages of a particular methodological approach and its software implementation. The tutorial also addresses common computational problems and their remedy, and general tricks useful in mixed model fitting.

Prerequisites:

This tutorial requires basic knowledge of linear mixed models and the MIXED procedure in SAS.

ENAR 2004 TUTORIALS

T4: Introduction to Joint Modeling of Longitudinal and Time-to-event Data

[Ballroom 4](#)

Date: Tuesday, March 30

Time: 1:45-3:30 p.m.

Instructor: Peter Diggle (University of Lancaster and Johns Hopkins University)

Description:

In many studies involving longitudinal follow-up of multiple subjects, the outcome variables for each subject are of two kinds: a longitudinal sequence of measurements at pre-determined observation times; and a possibly right-censored time to occurrence of an event of interest. A number of different approaches to the formulation and fitting of joint models for these two types of outcome variable have been proposed in the recent biostatistical literature. The aim of this tutorial session is to introduce participants to some of these approaches to joint modeling, and to discuss their application to practical biostatistical problems.

In particular, I will compare and contrast random effects models, which are conceptually attractive but involve computationally intensive fitting methods using non-standard software, and transformed Gaussian models, whose fitting requires only the manipulation of multivariate Gaussian probabilities.

Examples discussed in the tutorial will include the following.

1. An AIDS trial in which the longitudinal measurements are CD4 cell-counts and the event-time is progression to AIDS or death. One of several possible scientific questions is whether the CD4 cell-count sequence can be used as surrogate for time of progression.
2. A surgical intervention trial for heart patients, in which the longitudinal measurements are physical properties of current heart function and the event-time is death. The scientific focus is on the subject-level joint behaviour of measured heart function and survival hazard.
3. A randomised trial of different drug therapies for schizophrenia patients, in which the longitudinal measurements are PANSS (Positive and Negative Symptom Scale) scores, and the event-time is dropout from the trial. The scientific problem is to compare mean PANSS scores over time in the different treatment arms, allowing for possibly informative dropout.

Prerequisites:

It will be assumed that participants are familiar with standard methods of longitudinal data analysis (linear model with correlated errors) and survival analysis (simple parametric models, Cox proportional hazards model); a short summary of both topics will be given at the start of the tutorial.

ENAR 2004 ROUNDTABLES

R1: Review Times in Statistical Journals: Are Changes of Our Culture Feasible?

Discussion Leader: Xihong Lin, University of Michigan, Coordinating Editor, Biometrics

Lengthy review times have been recognized as an issue for statistical journals for many years. This is of particular concern for junior researchers who are under the time pressure of tenure and promotion. There are significant differences in review times and editorial structures between statistical journals and leading health sciences journals. However, the need and the feasibility for improving review times in statistical journals have been controversial. Arguments include timely publication of statistical papers is less urgent than medical journals, and statistical papers require much longer time to review. Required changes would challenge our long-standing culture, and no consensus has been reached in our profession. We will discuss these issues and possible strategies statistical journals can adopt to improve review times. Topics for discussion include the editorial board structure, the editorial process, and the review process. The Biometrics review time statistics and editorial process will be shared. Suggestions for improvements will be sought.

R2: Current Issues in Space-Time Modeling of Environmental and Health Data

Discussion Leader: Montserrat Fuentes, North Carolina State University

This roundtable focuses on the statistical challenges to model, predict and estimate the complex spatial-temporal structure of environmental and human health data. One of the main challenges is to calculate a Gaussian likelihood for large space-time problems, and modern approaches to this problem use spectral methods and algorithms to approximate the determinant and the inverse of a large matrix. However, the impact these approximations have on inference is not yet clear. Another issue is how to model a complex space-time dependency structure, and current work involves extensions of separable models and coregionalization models (in the multivariate case). These are popular models, but are they realistic? What is the best software to handle large spatial-temporal problems? What are the advantages/drawbacks of SAS, R and all the R packages available to handle space-time data? This roundtable provides an opportunity to propose and discuss strategies for addressing all these important issues and review the recent literature for space-time data, in particular the Bayesian methods to capture different sources of uncertainty.

R3: The “Other” Biometrics: Hijacking of Our Name, or an Opportunity?

Discussion Leader: Marie Davidian, North Carolina State University

The terms “Biometrics” and “Biometry” have been used since early in the twentieth century to refer to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. In the past few years, the term “biometrics” has been used in the popular media and in computer science, engineering, and other fields to refer to the area of technology devoted to identification of individuals using physiological or behavioral characteristics, such as retinal or iris scans or fingerprints. Academic programs focusing on these topics are emerging. A Google search on the word “Biometrics” yields our journal website in the top position, but there are numerous web sites devoted to this “other” biometrics, including those for professional societies and publications having to do with this new use of the term. Biometrics editors are besieged with requests by companies dealing in this technology to advertise in Biometrics and have even been asked to provide interviews to the media, who seem oblivious to long-standing meaning of the term. In his President’s Corner column in the first issue of Biometric Bulletin in 2003, IBS President Norman Breslow posed the question of how our discipline and societies such as IBS should respond to what some see as the “hijacking” of “Biometrics.”

What should our profession do? Change our name? Reach out to the practitioners of this new field (who undoubtedly face statistical challenges)? How would we do this? What are the consequences if we do nothing? This roundtable will provide a forum for brainstorming about this issue.

R4: Using Statistical Analysis of Clinical Data to Detect Fraud and Better Inform Site Monitoring in Clinical Trials

Discussion Leader: Lisa LaVange, Inspire Pharmaceuticals, Inc.

The use of data during an ongoing clinical study to facilitate effective conduct of the trial and increase quality control of the final data, without compromising study integrity through unintentional unmasking of treatments, will be discussed. Examples include the use of clinical data to trigger “for cause” site monitoring during the study and the examination of data patterns, measures of variance, and multivariable correlations after study completion in order to flag areas of possible data fraud. The discussion will include the practical aspects of involving the team statistician in a non-traditional role on a clinical trial.

R5: Communicating Uncertainty in a Policy Context

Discussion Leader: Tom Louis, Johns Hopkins University

In communicating the results of an environmental risk assessment (e.g., the health effects of air pollution), a medical policy assessment (e.g., recommendations on cancer screening) or behavioral risk assessment (e.g., diet and heart disease risk), we must communicate valid, credible and understandable information. The most

ENAR 2004 ROUNDTABLES

understandable and high-impact messages avoid communicating uncertainty, but science and ethics require that we communicate relevant uncertainties. Indeed, in some contexts uncertainty is the principal finding. In this roundtable we will discuss what to include in an uncertainty assessment, what to communicate and how to communicate the assessment (to policy makers, to the press and to individuals). In most settings a careful titration is necessary to produce an “honest” assessment with short-term impact and longer-term credibility. Discussion will focus on issues and examples and will not involve technical issues.

R6: Design-based and Model-based Inference: Instilling an Appreciation for the Difference

Discussion Leader: Tim Gregoire, Yale University, Current ENAR President

Following a planned experiment or a sample survey, inference may be based on the randomization distribution of all estimates possible under the design or on the distribution presumed under a stipulated model for the response variate. Because the two bases of inference may at times lead to nearly identical results, students and practitioners have difficulty appreciating the distinction between them. In the context of sampling, too, model-assisted estimation may serve further to impede a firm understanding of the role of models. This roundtable discussion will provide an opportunity to discuss these inferential issues as well as techniques that can be used to help instill a clear comprehension of their importance.

R7: Filling the Pipeline: Undergraduate Statistics Education

Discussion Leader: Joel Greenhouse, Carnegie Mellon University

Biostatistics is a fundamental scientific component of biomedical and public health research. Yet the current demand for biostatisticians far exceeds the supply and the gap is expected to continue to widen. One solution is to create stimulating undergraduate programs that will attract students to our field. How do we generate interest among undergraduates in the biostatistical sciences, and once we have their attention, what are the essential elements of an undergraduate curriculum that best prepare students for work in industry, government or for graduate school? Discussion will focus on these and other important issues.

R8: Current Issues and Topic of Statistical Interest at the FDA: An Open Discussion

Discussion Leader: Robert O'Neill, FDA Office of Biostatistics

This roundtable discussion with Dr. O'Neill is intended to be an open dialogue with participants on any topics of statistical interest to them. Current issues that roundtable participants raise will be addressed as time permits.

R9: Strategies to Handle Incomplete Data

Discussion Leader: Geert Molenberghs, Limburgs Universitair Centrum, Belgium and Current IBS President

The issue of incomplete data arises in a wide number of research areas, including longitudinal data in clinical and epidemiological research as well as survey research. A lot of research activity has been developed in this field for over a quarter of a century. Nevertheless, the impression is generated that a number of sub-debates have been ongoing (within the industry, the regulatory authorities, the academic world) with often very little interaction between them. Reasons for this will be explored, as well as strategies to increase the efficiency of the debate.

R10: Casual Inference for Infectious Diseases and Vaccine Studies

Discussion Leader: M. Elizabeth Halloran, Emory University

The potential outcome approach to causal inference has become widespread in a variety of settings. One of the common assumptions is that the potential outcomes in each individual are independent of the treatment assignments and outcomes in the other individuals. This assumption is obviously violated in infectious diseases, and sharp definitions of causal effects of vaccination and vaccine programs are difficult. Another issue in vaccine studies is that focus is often on the effect of vaccine on post-infection outcomes. In this case, the estimated effect could be subject to selection bias. This roundtable will discuss various approaches to these questions.

R11: Biostatistics Graduate Education in the 21st Century: Challenges and Opportunities

Discussion Leader: Clarence E. (Ed) Davis, University of North Carolina at Chapel Hill

The modern biostatistics department is at a critical juncture in fulfilling its mission to train the next generation of biostatistical scientists. The demand for PhD and Master's level biostatisticians continues to escalate; thus, encouraging talented individuals to pursue graduate training in the field is an essential requirement. As institutions of higher learning face new budget constraints, demands on faculty to secure outside funding are ever-higher, limiting the time they may devote to recruiting and curriculum development. The acceleration of new advances in biomedical technologies and the facility to collect massive amounts of data raise the question of whether current graduate biostatistics curricula provide appropriate preparation for the kinds of challenges graduates will face while at the same time providing significant new opportunities for broadening the scope of the training we provide. This roundtable will provide a forum for discussing these and other challenges and opportunities facing academic graduate programs in biostatistics.

PROGRAM SUMMARY

SATURDAY, MARCH 27

9:00 a.m.–9:00 p.m.

Kings Garden South

3:00 p.m.–5:00 p.m.

Ballroom Foyer

SUNDAY, MARCH 28

7:30 a.m.–6:30 p.m.

Ballroom Foyer

8:30 a.m.–5:00 p.m.

Sterlings

Ballroom 4

Ballroom 3

Kings Garden North & South

3:00 p.m.–5:00 p.m.

Ballroom 2

4:00 p.m.–7:00 p.m.

Rivers

4:30 p.m.–6:30 p.m.

Traders

8:00 p.m.–11:00 p.m.

Ballroom 1

MONDAY, MARCH 29

7:30 a.m.–8:30 a.m.

Sterlings 1

7:30 a.m.–5:00 p.m.

Ballroom Foyer

7:30 a.m.–5:00 p.m.

Black Diamond

9:00 a.m.–5:00 p.m.

Traders

8:30 a.m.–5:30 p.m.

Ballroom 2

8:30 a.m.–10:15 a.m.

Ballroom 4

Ballroom 3

Kings Garden North

Benedum

Kings Garden South

Sterlings 2 & 3

Brigade

Le Bateau

Heinz

Duquesne

Rivers

King Terrace

Workshop for Junior Researchers

Conference Registration

Conference Registration

Short Courses

SCI Hierarchical Modeling and Analysis for Spatial Data

SC2: Semiparametric Regression

SC3: Statistical Analysis with Missing Data

SC4: Practical Approaches to Data Monitoring of Clinical Trials

Exhibits Open

ENAR Executive Committee (closed)

Placement Service

Social Mixer and Poster Session

Student Breakfast

Conference Registration

Speaker Ready Room

Placement Service

Exhibits Open

Tutorial

TI: An Introduction to the R Environment

Scientific Program

1. Advances in Latent Variable Methodology: Bridging the Gap Between Social Science and Biomedical Perspectives

2. Evaluating National Security Surveillance Tools

3. Recent Developments in Phase I Trial Design

4. Statistics and Neuroscience

5. Special Contributed Session: Statistical Review in the Medical Literature

6. Contributed Papers: Microarray Normalization and Expression Indices

7. Contributed Papers: Missing Data for Longitudinal Trials

8. Contributed Papers: Applications in Statistical Genetics

9. Contributed Papers: Clinical Trials and Epidemiologic Methods

10. Contributed Papers: Bayesian Clinical Trial Design and Analysis

11. Contributed Papers: Spatial Modeling I

PROGRAM SUMMARY

10:15 a.m.–10:30 a.m.

Exhibit Hall-Ballroom 2

10:30 a.m.–12:15 p.m.

Sterlings 2 & 3

Ballroom 3

Kings Garden North

Le Bateau

Kings Garden South

Brigade

Rivers

Heinz

Kings Terrace

Duquesne

Benedum

12:15 p.m.–1:30 p.m.

Ballroom 1

12:30 p.m.–4:30 p.m.

Allegheny

1:45 p.m.–3:30 p.m.

Ballroom 4

Sterlings 2 & 3

Kings Garden North

Kings Garden South

Ballroom 3

Benedum

Duquesne

Brigade

Heinz

Rivers

Kings Terrace

Le Bateau

3:30 p.m.–3:45 p.m.

Exhibit Hall-Ballroom 2

3:45 p.m.–5:30 p.m.

Kings Garden North

Kings Garden South

Sterlings 2 & 3

Ballroom 3

Benedum

Le Bateau

Rivers

Brigade

Duquesne

Heinz

Kings Terrace

Break

Scientific Program

12. Estimation and Inference in Mixed-Effects Models

13. Methodological Issues in Preserving Confidentiality in Public Use Datasets

14. Modeling Treatment Use & Effectiveness in Mental Illness

15. Environmental Statistics

16. Special Contributed Session: Addressing the Looming Shortage of Statisticians

17. Contributed Papers: Survival Methods with Random Effects, Noncompliance, or Longitudinal Data

18. Contributed Papers: Epidemiologic Methods

19. Contributed Papers: Linkage Disequilibrium and QTL Analysis

20. Contributed Papers: Large Scale Multiple Testing

21. Contributed Papers: Causal Inference

22. Contributed Papers: Clinical Trials: Adaptive Design and Optimization

Roundtable Luncheons

Regional Advisory Board (RAB) Luncheon Meeting (By Invitation Only)

Tutorial

T2: Experiments and Observational Studies: Causal Inference in Statistics

Scientific Program

23. Challenging Issues in the Analysis of Longitudinal/Clustered Data

24. Statistical Issues in Emerging Diseases

25. Statistical Models for Complex Data in Agricultural and Environmental Studies

26. The Future of Biostatistical Science

27. Bayesian Methods in Survival Analysis

28. Contributed Papers: Analysis of Imaging Data

29. Contributed Papers: Clinical Trials: Non-Parametric Methods

30. Contributed Papers: Linkage Analysis

31. Contributed Papers: Survey Research Data

32. Contributed Papers: Gene Selection Methods

33. Contributed Papers: Generalized Linear Models

Break

Scientific Program

34. Nonparametric Regression Techniques Applied to Analysis of Survival Data

35. Statistics in Genetics

36. Assessing Efficacy in a Changing Background

37. The Causes and Consequences of HIPAA: A Statistician's Perspective (HIPAA: Health Insurance Portability and Accountability Act)

38. Information Related Problems in Biology

39. Contributed Papers: Missing Data Methods

40. Contributed Papers: Survival Analysis I

41. Contributed Papers: Environmental and Ecological Applications

42. Contributed Papers: Longitudinal Data Analysis I

43. Contributed Papers: Clustering and Classification Methods

44. Contributed Papers: Mixture Models and QT Corrections

PROGRAM SUMMARY

6:30 p.m.–7:30 p.m.
Kings Garden South

President’s Reception (By Invitation Only)

TUESDAY, MARCH 30

7:30 a.m.–5:00 p.m.
Ballroom Foyer

Conference Registration

7:30 a.m.–5:00 p.m.
Black Diamond

Speaker Ready Room

9:00 a.m.–5:00 p.m.
Traders

Placement Service

8:30 a.m.–5:30 p.m.
Ballroom 2

Exhibits Open

8:30 a.m.–10:15 a.m.
Ballroom 4

Tutorial

T3: Generalized and Nonlinear Mixed Models in the SAS System: Tools and Applications

Sterlings 2 & 3
Kings Garden North
Ballroom 3
Kings Garden South
Benedum

Scientific Program

- 45. What More Needs to be Done in the Graduate Training of Statisticians?
- 46. Methods for Combining Gene Expression Datasets from Different Studies
- 47. Recent Developments in Survival Analysis
- 48. New Advances in Nonparametric Statistics
- 49. Special Contributed Session: Combining Data from Disparate Surveys to Improve Small Area Estimation
- 50. Contributed Papers: Disease Detection and Surveillance
- 51. Contributed Papers: Bayesian Survival Analysis
- 52. Contributed Papers: Diagnostic and Screening Tests
- 53. Contributed Papers: Sequence Analysis and Proteomics
- 54. Contributed Papers: Sequential Trials, Interim Decision Making, and Multiplicity

Duquesne
Rivers
Brigade
Kings Terrace
Le Bateau

10:15 a.m.–10:30 a.m.
Exhibit Hall-Ballroom 2

Break

10:30 a.m.–12:15 p.m.
Ballroom 1

Presidential Invited Address

12:30 p.m.–4:30 p.m.
Sterlings 1

Regional Committee (RECOM) Luncheon Meeting (By Invitation Only)

1:45 p.m.–3:30 p.m.
Ballroom 4

Tutorial

T4: Introduction to Joint Modeling of Longitudinal and Time-to-event Data

Kings Garden North
Kings Garden South
Ballroom 3
Benedum
Sterlings 2 & 3
Duquesne
Le Bateau
Kings Terrace
Brigade
Rivers
Heinz

Scientific Program

- 55. Statistical Methods in Computer Vision
- 56. New Developments on Estimation Equations
- 57. Bayesian Methods in Genomics
- 58. Nonparametric Function Estimation
- 59. Stochastic Modeling and Inference in Science
- 60. Contributed Papers: Survival Analysis II
- 61. Contributed Papers: Experimental Design
- 62. Contributed Papers: Advances in QTL Analysis
- 63. Contributed Papers: Health Service and Epidemiologic Methods
- 64. Contributed Papers: Applications of Random Effects Models
- 65. Contributed Papers: Bioassays and Cell Kinetics

3:30 p.m.–3:45 p.m.
Exhibit Hall-Ballroom 2

Break

PROGRAM SUMMARY

3:45 p.m.–5:30 p.m.

Sterlings 2 & 3

Ballroom 3

Kings Garden North

Kings Garden South

Benedum

Duquesne

Heinz

Kings Terrace

Brigade

Le Bateau

Rivers

Scientific Program

66. Recent Advances in Tumor Growth Modeling

67. Functional and Longitudinal Data Analysis

68. Bioinformatics in Genomics Studies

69. Modeling and Surveillance of Sequences of Health Events

70. Analysis of Intensively Collected Data

71. Contributed Papers: Methods for Missing or Unbalanced Data

72. Contributed Papers: Genetics and Drug Screening

73. Contributed Papers: Survival Analysis: Methods for Multiple Endpoints

74. Contributed Papers: Categorical Data

75. Contributed Papers: Nonparametric Methods for Longitudinal Data and Bayesian Analysis

76. Contributed Papers: Surrogate Endpoints, Screening and Assessing Agreement

5:30 p.m.–6:00 p.m.

Brigade

ENAR Business Meeting (Open to all members)

6:00 p.m.–9:30 p.m.

Tuesday Night Event (Dinner at Heinz Field) – Depart Hotel Lobby Promptly at 6:00

WEDNESDAY, MARCH 31

7:30 a.m.–9:00 a.m.

Traders

Planning Committee (closed)

7:30 a.m.–12:00 noon

Ballroom Foyer

Conference Registration

7:30 a.m.–12:00 noon

Black Diamond

Speaker Ready Room

8:00 a.m.–12:00 noon

Ballroom 2

Exhibits Open

8:30 a.m.–10:15 a.m.

Sterlings 2 & 3

Scientific Program

77. Current Developments and Statistical Challenges in

Developing Syndromic Surveillance Systems for Anti-Bioterrorism

78. Novel Statistical Approaches to High Dimensional Analyses

79. Non-Inferiority/Equivalence Trials

80. Contributed Papers: Statistical Issues in Clinical Trials

81. Contributed Papers: Gene Expression Analysis

82. Contributed Papers: Non- and Semi-Parametric Models

83. Contributed Papers: Spatial Modeling II

84. Contributed Papers: Random Effects Models

85. Contributed Papers: Markov Chain Models

Ballroom 3

Ballroom 4

Kings Garden North

Brigade

Duquesne

Benedum

Kings Garden South

Rivers

10:15 a.m.–10:30 a.m.

Exhibit Hall-Ballroom 2

Break

10:30 a.m.–12:15 p.m.

Ballroom 3

Ballroom 4

Sterlings 2 & 3

Kings Garden North

Rivers

Brigade

Kings Garden South

Duquesne

Benedum

Le Bateau

Scientific Program

86. Assessing Agreement Measurements: Current Issues and New Developments

87. Applications of Resampling Methods to Estimating Functions

88. Statistical Genetics - Modeling Interaction and Multilocus Analyses

89. Contributed Papers: Bayesian Computing and Analysis

90. Contributed Papers: Nonparametric Methods

91. Contributed Papers: Measurement Error Methods

92. Contributed Papers: Longitudinal Data II

93. Contributed Papers: Genetic Maps, Linkage Disequilibrium and Haplotype Analysis

94. Contributed Papers: Methods for Multiple Endpoints

95. Contributed Papers: Biostatistical Theory and Estimation

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

SCIENTIFIC PROGRAM

Times may change slightly prior to the meetings. Please check the on-site program for final times. Asterisks (*) indicate paper presenters. Student Travel Award Winner presentations appear in **boldface**.

Monday, March 29

8:30–10:15 a.m.

I. ADVANCES IN LATENT VARIABLE METHODOLOGY: BRIDGING THE GAP BETWEEN SOCIAL SCIENCE AND BIOMEDICAL PERSPECTIVES Ballroom 3

Sponsor: ENAR

Organizer: David B. Dunson, National Institute of Environmental Health Sciences

Chair: Amy H. Herring, University of North Carolina at Chapel Hill

- 8:30 Latent Variables in the Social Sciences
Kenneth A. Bollen*, University of North Carolina, Chapel Hill
- 8:55 Bayesian Latent Variable Methods in Biomedical Applications
David B Dunson*, Biostatistics Branch, NIEHS, NIH
- 9:20 Integrating Biometric and Psychometric Latent Variable Modeling
Bengt O Muthen*, UCLA
- 9:45 Latent Variable Models: Causal Estimation, Extrapolation and Speculation
Scott L Zeger*, Johns Hopkins University; Kathryn A Ziegler, Johns Hopkins University
- 10:10 Floor Discussion

2. EVALUATING NATIONAL SECURITY SURVEILLANCE TOOLS Kings Garden North

Sponsor: ENAR

Organizer: Peter B. Imrey, Cleveland Clinic

Chair: David Banks, Duke University

- 8:30 Nonparametric Measures of Dependence and Aggregated Algorithms in Biometric Recognition Problems
Andrew L Rukhin*, NIST and UMBC
- 9:00 Syndromic Surveillance: Is it Worth The Effort?
Michael A Stoto*, RAND; Matt Schonlau, RAND; Louis T Mariano, RAND
- 9:30 Assessing Performance and Weighing Evidence from National Security Screening Tools
Peter B Imrey*, Ph.D., Cleveland Clinic Foundation; David L Faigman, J.D., Hastings College of the Law
- 10:00 Floor Discussion

3. RECENT DEVELOPMENTS IN PHASE I TRIAL DESIGN Benedum

Sponsors: ENAR/ASA Biopharmaceutical Section

Organizer: Susan Murray, University of Michigan

Chair: Adin-Cristian Andrei, University of Michigan

- 8:30 Determining a Maximum Tolerated Schedule of a Cytotoxic Agent
Thomas M Braun*, University of Michigan; Peter F Thall, M.D. Anderson Cancer Center; Zheng Yuan, University of Michigan
- 9:00 Modeling Dose-Toxicity Relationship in Phase I Clinical Trials
Ken Cheung*, Columbia University
- 9:30 Dose-Finding with Two Agents in Phase I Oncology Trials
Peter F Thall*, M.D. Anderson Cancer Center
Randall E Millikan, M.D. Anderson Cancer Center; Peter Mueller, M.D. Anderson Cancer Center; Sang-Joon Lee, M.D. Anderson Cancer Center
- 10:00 Floor Discussion

4. STATISTICS AND NEUROSCIENCE Kings Garden South

Sponsor: IMS

Organizer: Rob Kass, Carnegie Mellon University

Chair: Rob Kass, Carnegie Mellon University

- 8:30 Maximum Likelihood Estimation for Multivariate Point Process Models of Neural Spiking Activity
Emery Brown*, Harvard University
- 9:00 Finding Consistent Activation Across fMRI Tasks: Using pFDR to Test a Union Of Nulls
Thomas E Nichols*, University of Michigan
- 9:30 Statistical Models and Issues for Neural Prostheses
Valerie Ventura, Carnegie Mellon University; Cari Kaufman*, Carnegie Mellon University
- 10:00 Floor Discussion

SCIENTIFIC PROGRAM

5. SPECIAL CONTRIBUTED SESSION: STATISTICAL REVIEW IN THE MEDICAL LITERATURE [Sterlings 2 & 3](#)

Sponsor: ENAR

Organizer: Russell Localio, University of Pennsylvania

Chair: Lisa Weissfeld, University of Pittsburgh

- 8:30 Reviewing Clinical Research
James H Ware*, Harvard School of Public Health
- 9:00 Statistical Review in the Medical Literature
Russell Localio*, University of Pennsylvania
- 9:30 Why All Biotatisticians Should Review for Medical
Journals, or, How to Retain Your Limbs While
Swimming with Sharks
Steven Goodman*, Johns Hopkins University
- 10:00 Discussant: Joel Greenhouse, Carnegie Mellon
University

6. CONTRIBUTED PAPERS: MICROARRAY NORMALIZATION AND EXPRESSION INDICES [Brigade](#)

Sponsor: ENAR

Chair: Kyoungmi Kim, University of Alabama, Birmingham

- 8:30 Optimal Use of Probe Level Information in
Affymetrix Genechip Analysis
Jianbo Li*, Cleveland Clinic Foundation
- 8:45 Normalization of Microarray Data Using Mixture of
Splines
Yongsung Joo, University of Florida; George Casella,
University of Florida; James Booth, University of
Florida; Keunbaik Lee*, University of Florida
- 9:00 A Two-Way Semi-Linear Model for Normalization
and Significant Analysis
Deli Wang*, The University of Iowa; Jian Huang, The
University of Iowa; Cuihui Zhang, Rutgers University
- 9:15 Evaluation of Affymetrix Probe Level Summarization
Algorithms Using Quality Control Genes
Yaomin Xu*, Cleveland Clinic Foundation; Xuejun
Peng, Cleveland Clinic Foundation
- 9:30 Some Issues on Cross-Validation in Microarray Data
Analysis
Wenjiang M. Fu*, Texas A & M University
- 9:45 A Case Study on Choosing Normalization Methods
and Test Statistics for Microarray Data
Yang Xie, University of Minnesota; Wei Pan,
Univeristy of Minnesota; Keyong Jeong, University of
Minnesota; Arkady Khodursky, Univeristy of Minnesota;
Brad Carlin, University of Minnesota

- 10:00 Adaptive SVD Approaches of Estimating Expression
Indexes for Oligonucleotide Arrays
Jianhua Hu, University of North Carolina at
Chapel Hill; Fred Wright, University of North Carolina
at Chapel Hill; Fei Zou, University of North Carolina
at Chapel Hill

7. CONTRIBUTED PAPERS: MISSING DATA FOR LONGITUDINAL TRIALS [Le Bateau](#)

Sponsor: ENAR

Chair: Geert Molenberghs, Limburgs Universitair Centrum

- 8:30 A Pattern-Selection Approach for the Analysis of
Longitudinal Data with Informative Missingness
Soomin Park*, Eli Lilly and Company; Lei Shen,
Ohio State University
- 8:45 Analysis of the Bias Caused by Using the Available
Case Method with Longitudinal Data
Lin Wang*, University of Wisconsin Madison; Mari
Palta, University of Wisconsin Madison; Jun Shao,
University of Wisconsin Madison
- 9:00 Nonparametric Multiple Imputation of Incomplete
Longitudinal Data Via Functional Mixed Model
Trivellore E Raghunathan, University of Michigan;
Yulei He*, University of Michigan
- 9:15 Estimating Treatment Effects from Longitudinal
Clinical Trial Data with Missing Values: Comparative
Analyses Under Different Model Assumptions
Patricia R. Houck*, University of Pittsburgh; Sati
Mazumdar, University of Pittsburgh; Tulay Koru-
Sengul, University of Pittsburgh; Gong Tang,
University of Pittsburgh; Benoit Mulsant, University
of Pittsburgh; Bruce Pollock, University of
Pittsburgh; Charles F. Reynolds, University of
Pittsburgh
- 9:30 Multiple Imputation for Longitudinal Data with
Informative Missingness
Wei Deng*, Ohio State University; Lei Shen,
Ohio State University
- 9:45 Semiparametric Models for Longitudinal Bivariate
Ordinal Outcomes Subject to Informative Dropouts
David Todem*, Michigan State University; Kyung
Mann Kim, University of Wisconsin-Madison
- 10:00 Application of Pattern Mixture Models to Non-
ignorable Missing Data Using Pseudo Maximum
Likelihood Estimation
Changyu Shen, University of Pittsburgh; Lisa Weissfeld,
Univeristy of Pittsburgh

SCIENTIFIC PROGRAM

8. CONTRIBUTED PAPERS: APPLICATIONS IN STATISTICAL GENETICS [Heinz](#)

Sponsor: ENAR

Chair: Heping Zhang, Yale University

- 8:30 Effect of Measurement Error in Reported Family History on Predictions From Mendelian Models of Disease Genes
Hormuzd A Katki*, National Cancer Institute, NIH, DHHS; Giovanni Parmigiani, Johns Hopkins University
- 8:45 **Efficient Algorithms for Allele Sharing Methods to Detect Linkage Using Extended Pedigrees**
Saonli Basu*, University of Washington, Seattle
- 9:00 Extreme Sibling Pair Designs for Quantitative Trait Locus Mapping
Jin P Szatkiewicz*, University of Pittsburgh; Eleanor Feingold, University of Pittsburgh
- 9:15 A Hidden Markov Modeling Approach for Admixture Mapping of Complex Diseases
Chun Zhang*, University of California Davis; Kun Chen, University of California Davis; Michael F. Seldin, University of California Davis; Hongzhe Li, University of California Davis
- 9:30 A Method for Using Complete and Incomplete Trios to Identify Genes Related to a Quantitative Trait
Emily O. Kistner*, NIEHS; Clarice R Weinberg, NIEHS
- 9:45 Stochastic Search Gene Suggestion: A Bayesian Hierarchical Model for Gene Mapping
Michael D Swartz*, Rice University, M. D. Anderson Cancer Center; Marek Kimmel, Rice University
- 10:00 Floor Discussion

9. CONTRIBUTED PAPERS: CLINICAL TRIALS AND EPIDEMIOLOGIC METHODS [Duquesne](#)

Sponsor: ENAR

Chair: Larry Gould, Merck Research Laboratories

- 8:30 Recurrent Episodes Analysis
Jun Yan*, University of Iowa; Jason P. Fine, University of Wisconsin-Madison
- 8:45 A Phase II Trial Design Based on Median Time To Event
Theodore G. Karrison*, University of Chicago; Dezheng Huo, University of Chicago

- 9:00 Piecewise Constant Cross-Ratio Estimates for Bivariate Survival Data
Bin Nan*, University of Michigan; Xihong Lin, University of Michigan; Lynda D. Lisabeth, University of Michigan Sioban D. Harlow, University of Michigan
- 9:15 First-Year Survival of Infants Born with Congenital Heart Defects in Arkansas (1993-1998): A Survival Analysis Using Registry Data
Mario A. Cleves, University of Arkansas for Medical Sciences; Sadia Ghaffar, University of Arkansas for Medical Sciences; Weizhi Zhao*, University of Arkansas for Medical Sciences; Bridget S. Mosley, University of Arkansas for Medical Sciences; Charlotte A. Hoobs, University of Arkansas for Medical Sciences
- 9:30 Sample Size and Power of Randomized Designs
Feifang Hu*, University of Virginia
- 9:45 A General Imputation Method for Causal Inference and G-Computation Formula
Zhuo Yu*, Bristol-Myers Squibb; Mark van der Laan, University of California at Berkeley
- 10:00 Comparison of Estimation Techniques For Medical Costs with Censored Data
Corina M. Sirbu*, Michigan State University; Joseph C. Gardiner, Michigan State University

10. CONTRIBUTED PAPERS: BAYESIAN CLINICAL TRIAL DESIGN AND ANALYSIS [Rivers](#)

Sponsor: ENAR

Chair: Gary Rosner, M.D. Anderson Cancer Center

- 8:30 Bayesian Optimal Design for Phase II Screening Trials
Meichun Ding*, Rice University; Gary L. Rosner, M.D. Anderson Cancer Center; Peter Mueller, M.D. Anderson Cancer Center
- 8:45 Adjustable Bayesian Two-Stage Design
Ming Li*, Vanderbilt University Medical Center; Yu Shyr, Vanderbilt University Medical Center
- 9:00 A Bayesian Approach for Designing Futility Rules in Large Clinical Trials
Ben L. Trzaskoma*, Eli Lilly & Company; Andreas Sashegyi, Eli Lilly & Company
- 9:15 Sample Size Re-Estimation Using Posterior Predictive Power in Randomized Clinical Trials with Binary Outcomes
Hyung Woo Kim*, MD Anderson Cancer Center; Donald Berry, MD Anderson Cancer
- 9:30 Predictive Probability in Phase II Cancer Clinical Trials
Diane D. Liu*, M.D. Anderson Cancer Center; J. Jack Lee, M.D. Anderson Cancer Center

SCIENTIFIC PROGRAM

- 9:45 Bayesian Approach to Individual Bioequivalence
Pulak Ghosh*, Georgia State University; Ravindra Khattree, Oakland University
- 10:00 Two-Sample Confidence Intervals for Threshold Violation Probabilities: Application to Noninferiority Testing
Alan H. Feiveson*, Johnson Space Center; Linda C Shackelford, Johnson Space Center

II. CONTRIBUTED PAPERS: SPATIAL MODELING I

Sponsor: ENAR [Kings Terrace](#)
Chair: Sudipto Banerjee, University of Minnesota

- 8:30 Generalized Multiple Spatial Factor Model
Xuan Liu*, University of Minnesota; Melanie M. Wall, University of Minnesota
- 8:45 Confidence Intervals for the Ratio of Intensities
Traci L. Leong*, Emory University; Andrew N. Hill, Emory University; Lance A. Waller, Emory University
- 9:00 Weighted U-Statistics from Non-Identically Distributed Data with Application to the Regression of Interpoint Distances
Dionne A. Graham*, Harvard School of Public Health; Marcello Pagano, Harvard School of Public Health
- 9:15 Exploring Goodness-of-Fit and Spatial Correlation Using Components of Tango's Index of Spatial Clustering
Monica C. Jackson*, Emory University; Lance A. Waller, Emory University
- 9:30 Decoupling Local and Global Behavior in Spatial Models
Tilman J. Gneiting*, University of Washington; Martin Schlather, Universite du Luxembourg
- 9:45 Maximum Efficiency Robust Tests for Focused Clustering
Pablo E. Bonangelino*, George Washington University; Thomas A. Louis, Johns Hopkins University
- 10:00 Test for Isotropy for Spatial Point Processes
Yongtao Guan*, University of Miami; Michael Sherman, Texas A&M University; James A. Calvin, Texas A&M University

Monday, March 29
10:15–10:30 a.m.

BREAK [Exhibit Hall-Ballroom 2](#)

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

12. ESTIMATION AND INFERENCE IN MIXED-EFFECTS MODELS [Sterlings 2 & 3](#)

Sponsors: ENAR/ASA Section on Statistics and the Environment/
ASA Biopharmaceutical Section
Organizer: Edward F. Vonesh, Baxter Healthcare Corporation
Chair: Geert Verbeke, Biostatistical Centre, Katholieke Universiteit Leuven

- 10:30 Laplace Approximation Based Methods in Nonlinear Mixed Effects Models
Lei Nie*, University of Maryland at Baltimore County; Edward F. Vonesh, Baxter Healthcare Corporation; Jinglin Zhong, Merck & Co.; Dibyen Majumdar, University of Illinois at Chicago
- 10:55 Hypothesis Testing in Mixed-Effects Models
Edward F. Vonesh*, Applied Statistics Center, Baxter Healthcare Corporation
- 11:20 Influence and Residual Diagnostics in Mixed Models
Oliver Schabenberger*, SAS Institute Inc.
- 11:45 Score Tests for Variance Components
Geert Molenberghs*, Limburgs Universitair Centrum, Diepenbeek, Belgium Geert Verbeke, Katholieke Universiteit Leuven, Leuven, Belgium
- 12:10 Floor Discussion

13. METHODOLOGICAL ISSUES IN PRESERVING CONFIDENTIALITY IN PUBLIC USE DATASETS [Ballroom 3](#)

Sponsor: ASA Section on Survey Research Methods
Organizer: Michael Elliott, University of Pennsylvania
Chair: Michael Elliott, University of Pennsylvania

- 10:30 Inferentially Valid Synthetic Data Sets for Disclosure Limitation
Trivellore E. Raghunathan*, University of Michigan
- 11:15 Disclosure Limitation Via Partially Synthetic Multiply Imputed Datasets
Jerome P. Reiter*, ISDS, Duke University
- 12:00 Discussants: Donald Rubin, Harvard University; Alan Zaslavsky, Harvard University

SCIENTIFIC PROGRAM

14. MODELING TREATMENT USE & EFFECTIVENESS IN MENTAL ILLNESS *Kings Garden North*

Sponsor: ENAR

Organizer: Sharon- Lise Normand, Harvard School of Medicine

Chair: Sharon-Lise Normand, Harvard School of Medicine

- 10:30 A Naturalistic Study of Serious Adverse Effects of Atypical Antipsychotics in a Large Multi-Ethnic Publicly Insured US Population
Marcela Horvitz-Lennon*, Harvard Medical School; Rusty Tchernis, Harvard Medical School; Larry Zaborsky, Harvard Medical School; Sharon-Lise Normand, Harvard Medical School and Harvard School of Public Health
- 11:00 Evaluating the Use of Refreshment Samples for Making Inferences in the Presence of Non-Ignorable Missing Data
Douglas E. Levy*, Harvard Medical School; Sharon-Lise Normand, Harvard Medical School; and Harvard School of Public Health
- 11:30 Accounting for Treatment-Noncompliance and Missing Outcomes in Randomized Trials
A. James O'Malley*, Harvard Medical School; Sharon-Lise Normand, Harvard Medical School, Harvard School of Public Health
- 12:00 Discussants: Ann Hohmann, NIMH; Robert Rosenheck, M.D., Yale School of Medicine

15. ENVIRONMENTAL STATISTICS *Le Bateau*

Sponsors: IMS/ASA Section on Statistics and the Environment

Organizer: Tilmann Gneiting, University of Washington

Chair: Tilmann Gneiting, University of Washington

- 10:30 Spatial Association Between Speciated Fine Particles and Human Health Effects
Montserrat Fuentes*, North Carolina State University; Hae-Ryoung Song, North Carolina State University; Sujit Ghosh, North Carolina State University; David Holland, U.S. EPA
- 10:55 Longitudinal and Spatial Analysis of Short-Term Respiratory Health Effects of Air Pollution
Paul Rathouz*, University of Chicago; Vanja Dukic, Dept. of Health Studies and CISES; Dana Draghicescu, CISES; Edward Naureckas, Dept. of Medicine and CISES; Xiaoming Bao, CISES; John Frederick, Dept. of Geophysical Sciences and CISES; Alexis Zubrow, CISES

- 11:20 Functional Data Analysis Methods in the Environmental Sciences
Wendy Meiring*, University of California, Santa Barbara
- 11:45 Separable Approximations Of Space-Time Covariances
Marc G. Genton*, North Carolina State University
- 12:10 Floor Discussion

16. SPECIAL CONTRIBUTED SESSION: ADDRESSING THE LOOMING SHORTAGE OF STATISTICIANS *Kings Garden South*

Sponsor: ASA Section on Statistical Education

Organizer: Bonnie LaFleur, Food and Drug Administration

Chair: Bonnie LaFleur, Food and Drug Administration

- 10:30 The Potential for AP Statistics to Lead to More Statisticians
Linda J. Young*, University of Florida
- 10:55 Two Programs to Attract Students to Statistics/ Biostatistics
Dennis D. Boos*, North Carolina State University
- 11:20 Creating and Sustaining a Dynamic Undergraduate Statistics Program
Bruce J. Collings*, Brigham Young University
- 11:45 Where Are We and What Is NIH Doing About It?
Dennis O. Dixon*, NIH, NIAID
- 12:10 Floor Discussion

17. CONTRIBUTED PAPERS: SURVIVAL METHODS WITH RANDOM EFFECTS, NONCOMPLIANCE, OR LONGITUDINAL DATA *Brigade*

Sponsor: IMS

Chair: Xuelin Huang, M.D. Anderson Cancer Center

- 10:30 Semiparametric Linear Transformation Models with Random Effects for Clustered Failure Time Data
Donglin Zeng*, University of North Carolina; Danyu Lin, University of North Carolina; Xihong Lin, University of Michigan
- 10:45 A Corrected Pseudo-Score Approach for Additive Hazards Model with Longitudinal Covariates Measured with Error
Xiao Song*, University of Washington; Yijian Huang, Fred Hutchinson Cancer Research Center
- 11:00 Bivariate Truncated Failure Time Data
Rinku Sutradhar*, University of Waterloo; Richard J. Cook, University of Waterloo

SCIENTIFIC PROGRAM

- 11:15 Nonparametric Estimation of the Bivariate Recurrence Time Distribution
Chiung-Yu Huang*, University of Minnesota; Mei-Cheng Wang, Johns Hopkins University
- 11:30 A Generalized Residual For Multivariate Survival And Frailty Models
Kirby L Jackson*, University of South Carolina
- 11:45 An Improved Small-Sample Variance Estimator for the Analysis of Clustered Survival Data
Rong Zhou*, Medpace, Inc.; Jorge G. Morel, P&G Pharmaceuticals, Inc.
- 12:00 Structural Proportional Hazards Models For Causal Inference In Clinical Trials
Els JT Goetghebeur*, Ghent University; Tom Loeys, Harvard School of Public Health

18. CONTRIBUTED PAPERS: EPIDEMIOLOGIC METHODS

[Rivers](#)

Sponsor: ENAR

Chair: Mitch Gail, National Cancer Institute

- 10:30 Multipoint Linkage Disequilibrium Mapping Using Case-Control Designs
Kung-Yee Liang, National Health Research Institutes; Yen-Feng Chiu*, National Health Research Institutes
- 10:45 Case-Control Studies of Gen-Environment Interactions with Missing Data
Christine Spinka*, Texas A&M University; Nilanjan Chatterjee, National Cancer Institute; Raymond J. Carroll, Texas A&M University
- 11:00 Case-Cohort Designs and Analysis for Clustered Failure Time Data
Shou-En Lu*, University of Medicine and Dentistry of New Jersey; Joanna H. Shih, National Cancer Institute
- 11:15 Semi-Parametric Survival Analysis of Case-Control Family Data
Lu Chen*, University of Washington; Li Hsu, Fred Hutchinson Cancer Research Center
- 11:30 A Comparison of Methods Used to Assess Confounding and Mediation in Logistic Regression
Lorraine G. Ogden*, University of Colorado; Leann Myers, Tulane University
- 11:45 Regression Adjustments to Estimators of Average Causal Effects That Use Stratification or Weighting Via the Propensity Score
Jared K. Lunceford*, Merck Research Laboratories; Marie Davidian, North Carolina State University

- 12:00 Multivariate Graphics for FDA of Hormone Patterns in the Menopausal Transition
Peter M. Meyer*, Rush University Medical Center; Judith L. Luborsky, Rush University Medical Center; Sioban Harlow, University of Michigan; Daniel S. McConnell, University of Michigan; Imke Janssen, Rush University Medical Center

19. CONTRIBUTED PAPERS: LINKAGE DISEQUILIBRIUM AND QTL ANALYSIS

[Heinz](#)

Sponsor: ENAR

Chair: Fei Zou, University of North Carolina at Chapel Hill

- 10:30 High Resolution LD Mapping of QTL by Multiple Markers
Jeesun Jung, Texas A&M University; Ruzong Fan*, Texas A&M University
- 10:45 Tagging Candidate Genes with Spectral Decomposition
Zhaoling Meng*, Merck & Co., Inc.; Dmitri V. Zaykin, GlaxoSmithKline
- 11:00 Fine-Scale Mapping of Disease Genes Based on Linkage Disequilibrium
Na Li*, University of Minnesota; Matthew Stephens, University of Washington
- 11:15 Genetic Association Mapping Under Founder Heterogeneity Via Haplotype Similarity Analysis in Candidate Genes
Kai Yu*, Washington University; Charles Gu, Washington University; Mike Province, Washington University; Chengjie Xiong, Washington University; D.C. Rao, Washington University
- 11:30 A Unifying Model for Fine Mapping of Quantitative Trait Loci Affecting CD4+ T Cell and HIV-1 Dynamics Under Highly Active Antiretroviral Therapy
Zuoheng Wang*, University of Florida; Rongling Wu, University of Florida
- 11:45 A Statistical Genetic Model for Integrating Growth and Development
Min Lin*, University of Florida; Rongling Wu, University of Florida
- 12:00 Floor Discussion

SCIENTIFIC PROGRAM

20. CONTRIBUTED PAPERS: LARGE SCALE MULTIPLE TESTING

Kings Terrace

Sponsors: ENAR/IMS

Chair: Grier P. Page, University of Alabama, Birmingham

- 10:30 A Bayesian Perspective on False Discovery and False Non-Discovery Rates
Jie Chen*, Merck Research Laboratories; Sanat K. Sarkar, Temple University
- 10:45 Improving Estimation of the False Discovery Rate
Stan B. Pounds*, St. Jude Children's Research Hospital; Cheng Cheng, St. Jude Children's Research Hospital
- 11:00 An Empirical Bayes Adjustment to Multiple P-Values
Somnath Datta, University of Georgia; Susmita Datta*, Georgia State University
- 11:15 New Procedures for Controlling False Discovery Rate
Zhongfa Zhang*, Case Western Reserve University; Jiayang Sun, Case Western Reserve University
- 11:30 A Mixture Model for Identifying Differentially Expressed Genes in DNA Microarray
Jiangang Liao*, UMDNJ; Yong Lin, Cancer Institute of New Jersey; Zachariah Selvanayagam, University of Medicine and Dentistry of New Jersey; Joe Shih, University of Medicine and Dentistry of New Jersey
- 11:45 Analysis of Multifactor Designs with Large Number of Factor Levels
Haiyan Wang*, Penn State University; Michael G. Akritas, Penn State University
- 12:00 Sharp Simultaneous Intervals for the Means of Selected Populations with Application to Microarray Data Analysis
Jing Qiu*, Cornell University; J. T. Gene Hwang, Cornell University

21. CONTRIBUTED PAPERS: CAUSAL INFERENCE

Duquesne

Sponsor: ENAR

Chair: Brent Johnson, University of North Carolina at Chapel Hill

- 10:30 Causal Inference in Hybrid Intervention Trials Involving Treatment Choice
Qi Long*, University of Michigan; Roderick J. Little, University of Michigan; Xihong Lin, University of Michigan
- 10:45 Examining Sensitivity to the Exclusion Restriction for Instrumental Variables Estimators of Treatment Effect
Joseph W. Hogan*, Brown University; Tony Lancaster, Brown University; Jason Roy, University of Rochester; Donald Alderson, Brown University

- 11:00 Modifications to Marginal Structural Models and Associated Weighted Estimation
Marshall M. Joffe*, University of Pennsylvania; Jill Knauss, University of Pennsylvania; Bruce Robinson, University of Pennsylvania; Harold I. Feldman, University of Pennsylvania
- 11:15 Causal Random Effects Linear Model Using Nagelkerke's Instrumental Variable Approach in a Longitudinal Setting
Julia Y. Lin*, University of Pennsylvania; Thomas Ten Have, University of Pennsylvania
- 11:30 Exploring Intrinsic Dependence with EI CID
Tailen Hsing, Texas A&M University; Li-yu Liu*, Texas A&M University; Marcel Brun, Texas A&M University; Edward R. Dougherty, Texas A&M University
- 11:45 A Mixed-Effect Propensity Adjustment for Treatment Effectiveness Analyses of Ordinal Doses
Andrew C. Leon*, Cornell University; Don Hedeker, University of Illinois at Chicago; Jedediah Teres, Cornell University
- 12:00 Floor Discussion

22. CONTRIBUTED PAPERS: CLINICAL TRIALS: ADAPTIVE DESIGN AND OPTIMIZATION

Benedum

Sponsors: ENAR/IMS

Chair: Peter Thall, M.D. Anderson Cancer Center

- 10:30 Adaptive Design with Covariates
George Sirbu*, Michigan State University
- 10:45 Power Comparisons of Adaptive Two-Stage Designs
Cyrus R. Mehta*, Cytel Software Corporation; Anastasios Tsiatis, North Carolina State University; Yuliya Lokhnygina, North Carolina State University
- 11:00 An Optimal Biased Coin Adaptive Design for Heteroskedastic Outcomes
Thomas E. Gwise*, University of Virginia; Feifang Hu, University of Virginia
- 11:15 Optimal Two-Stage Adaptive Designs
Yuliya Lokhnygina*, North Carolina State University; Anastasios A. Tsiatis, North Carolina State University
- 11:30 A Flexible Method for Design of Two-Stage Adaptive Procedures
Tatsuki Koyama*, Vanderbilt University Medical Center; Leon J. Gleser, University of Pittsburgh; Allan R. Sampson, University of Pittsburgh
- 11:45 Designing Nonlinear Experiments Using Nonenumerative Strategies
A. John Bailer*, Miami University; Stephen Wright, Miami University
- 12:00 A Note on Comparing Efficacy Proportions in Two-Stage Clinical Trials
Kallappa M. Koti*, U.S. Food and Drug Administration

SCIENTIFIC PROGRAM

Monday, March 29

12:15–1:30 p.m.

ROUNDTABLE LUNCHEONS

[Ballroom 1](#)

Monday, March 29

1:45–3:30 p.m.

23. CHALLENGING ISSUES IN THE ANALYSIS OF LONGITUDINAL/CLUSTERED DATA

[Sterlings 2 & 3](#)

Sponsors: ENAR/ASA Biometrics Section

Organizer: Naisyin Wang, Texas A&M University

Chair: Naisyin Wang, Texas A&M University

- 1:45 Semiparametric Longitudinal/Spatial Modeling of Binary Outcomes, with Application to Aberrant Crypt Foci in Colon Carcinogenesis Experiments
Tatiana V. Apanasovich*, Ph.D., Texas A&M University; Raymond J. Carroll, Texas A&M University
- 2:10 Estimation of Regression Models for the Mean of Repeated Outcomes Under Non-Ignorable Non-Monotone Non-Response
Stijn Vansteelandt*, Ghent University; Andrea Rotnitzky, Harvard School of Public Health; James M. Robins, Harvard School of Public Health
- 2:35 Accounting for Response Correlation in Varying-Coefficient Models
Naisyin Wang, Texas A&M University; Jeng-Min Chiou*, National Health Research Institutes, Taiwan
- 3:00 Dynamic Criteria and Longitudinal Accuracy
Patrick J. Heagerty*, University of Washington
- 3:25 Floor Discussion

24. STATISTICAL ISSUES IN EMERGING DISEASES

[Kings Garden North](#)

Sponsor: ASA Section on Risk Analysis

Organizer: Steve A. Anderson, U.S. Food and Drug Administration

Chair: Susan M. Sereika, University of Pittsburgh

- 1:45 Bse and Prion Diseases in North America - What Are the Risks?
Steven A. Anderson*, U.S. Food and Drug Administration

- 2:15 Estimated Risk of West Nile Virus Transfusion Transmission Through Voluntary Blood Donation in the United States, 2002
Brad J. Biggerstaff*, National Center for Infectious Diseases Centers for Disease Control and Prevention; Lyle R. Petersen, National Center for Infectious Diseases Centers for Disease Control and Prevention
- 2:45 Containing Pandemic or Bioterrorist Influenza with Antivirals
M. Elizabeth Halloran*, Emory University; Ira M. Longini, Emory University; Yang Yang, Emory University; Azhar Nizam, Emory University
- 3:15 Floor Discussion

25. STATISTICAL MODELS FOR COMPLEX DATA IN AGRICULTURAL AND ENVIRONMENTAL STUDIES

[Kings Garden South](#)

Sponsor: ASA Section on Statistics and the Environment

Organizers: Estelle Russek-Cohen, University of Maryland-College Park and Jun Zhu University of Wisconsin-Madison

Chair: Mary Christman, University of Maryland-College Park

- 1:45 A Semiparametric Stochastic Mixed Model for Increment-Averaged Data with Application to Carbon Sequestration in Agricultural Soils
F. Jay Breidt*, Colorado State University; Nan-Jung Hsu, National Tsing Hua University; Stephen Ogle, Colorado State University
- 2:10 Factor Analysis Models for Pollution Source Apportionment
William F. Christensen*, Brigham Young University
- 2:35 A Spatial-Temporal Auto-Logistic Regression Model for the Analysis of the Southern Pine Beetle Outbreaks
Jun Zhu*, University of Wisconsin – Madison; Hsin-Cheng Huang, Institute of Statistical Science, Academia; Sinica Chi-tung Wu, Feng Chia University
- 3:00 Prediction of Riverine Sediment Categories Using a Spatial Mixture Model
Mark S. Kaiser*, Iowa State University; Kyoji Furukawa, Iowa State University
- 3:25 Floor Discussion

SCIENTIFIC PROGRAM

26. THE FUTURE OF BIOSTATISTICAL SCIENCE

[Ballroom 3](#)

Sponsors: ENAR/ASA Biometrics Section
 Organizer: Marvin Zelen, Harvard University.
 Chair: Marvin Zelen, Harvard University

- 1:45 Introductory Remarks
 Panel Discussion
 Donald A. Berry, M.D. Anderson Cancer Center
 David P. Harrington, Dana-Farber Cancer Center
 and Harvard University
 George W. Williams, Amgen, Inc.

27. BAYESIAN METHODS IN SURVIVAL ANALYSIS

[Benedum](#)

Sponsor: IMS
 Organizer: Radu Craiu, University of Toronto
 Chair: Radu Craiu, University of Toronto

- 1:45 Multiresolution Models in Multicenter Survival Data
 Analysis
 Peter Bouman, Northwestern University; Vanja
 Dukic*, University of Chicago; James Dignam,
 University of Chicago; Xiao-Li Meng, Harvard
 University
- 2:10 Bayesian Estimation of Multivariate Frailty Models
 for Exchangeable Survival Data
 Catalina Stefanescu*, London Business School; Bruce
 Turnbull, Cornell University
- 2:35 A Simple Approach to Fitting Bayesian Survival Models
 Paul Gustafson*, University of British Columbia;
 Dana Aeschliman, University of British Columbia;
 Adrian R. Levy, University of British Columbia
- 3:00 A New Model for Longitudinal and Survival Data
 with a Cure Fraction
 Ming-Hui Chen Chen*, University of Connecticut;
 Joseph G. Ibrahim, University of North Carolina;
 Debajyoti Sinha, Medical University of South Carolina
- 3:25 Floor Discussion

28. CONTRIBUTED PAPERS: ANALYSIS OF IMAGING DATA

[Duquesne](#)

Sponsor: ENAR
 Chair: DuBois Bowman, Emory University

- 1:45 Covariate Adjustment in Partial Least Squares for
 the Extraction of the Spatial-Temporal Pattern From
 Positron Emission Tomography Data
 Lei Xu*, Merck & Co., Inc.; Sati Mazumdar,
 University of Pittsburgh; Julie Price, University of
 Pittsburgh
- 2:00 Estimation in Nonlinear Regression with
 Applications to Kinetic Modeling with Pet Imaging Data
 Todd Ogden*, Columbia University; Thaddeus Tarpey,
 Wright State University
- 2:15 Metabolite Curve Fitting in Pet Imaging Studies of
 Neuroreceptor Mapping
 Songmei Wu*, Columbia University; Todd Ogden,
 Columbia University; Ramin V. Parsey, NYSPI;
 John Mann, NYSPI
- 2:30 Exploring Spatial Relationships in Functional
 Neuroimaging Data
 Rajan S. Patel*, Emory University; F. Dubois
 Bowman, Emory University; Chengxing Lu, Emory
 University
- 2:45 A Template Method for Combining Fmri
 Connectivity Across Subjects
 Raymond G. Hoffmann*, Medical College of
 Wisconsin; Natalia S. Lawrence, Institute of
 Psychiatry London, England; Elliot A. Stein, NIDA-IRP
- 3:00 Brain Function, the Movie: Optical Imaging of Brain
 Activity
 Kary L. Myers*, Carnegie Mellon University
- 3:15 Floor Discussion

29. CONTRIBUTED PAPERS: CLINICAL TRIALS: NON-PARAMETRIC METHODS

[Brigade](#)

Sponsor: ENAR
 Chair: Rick Chappell, University of Wisconsin

- 1:45 Nonparametric Prediction of Event Times in
 Randomized Clinical Trials
 Gui-shuang Ying*, University of Pennsylvania; Daniel
 Heitjan, University of Pennsylvania; Tai-Tsang Chen,
 Columbia University
- 2:00 A New Binomial Variance Estimator for the Kaplan-
 Meier Survival Function
 Craig B. Borkowf*, Centers for Disease Control and
 Prevention

SCIENTIFIC PROGRAM

- 2:15 A Nonparametric Test for Comparing Several Survival Curves with Interval Censored Data
Dae Ryong Kang*, Yonsei University, South Korea; Jinheum Kim, University of Suwon, South Korea; Chung Mo Nam, Yonsei University, South Korea
- 2:30 Pair-Wise Comparison Estimation of Clustered Nonparametric Transformation Models
Yan Li*, Rush University Medical Center; Daniel Zelterman, Yale University
- 2:45 Nonparametric Model Selection for Cox's Proportional Hazard Model
Chenlei Leng*, University of Wisconsin; Hao Helen Zhang, North Carolina State University; Yi Lin, University of Wisconsin; Grace Wahba, University of Wisconsin; I. J. Schoenberg, University of Wisconsin
- 3:00 Marginal Estimation in Multistage Models Using Current Status Data
Somnath Datta, University of Georgia; Rajeshwari Sundaram*, University of North Carolina-Charlotte
- 3:15 Floor Discussion

30. CONTRIBUTED PAPERS: LINKAGE ANALYSIS

Sponsors: ENAR/IMS

Chair: Zhaoling Meng, Merck & Co, Inc.

[Heinz](#)

- 1:45 Unification of Variance Components and Haseman-Elston Regression for Quantitative Trait Linkage Analysis
Wei-Min Chen*, Johns Hopkins University; Karl W. Broman, Johns Hopkins University; Kung-Yee Liang, Johns Hopkins University
- 2:00 A Latent Variable Model of Linkage Analysis for Ordinal Traits
Rui Feng*, Yale University; Heping Zhang, Yale University
- 2:15 Asymptotic Equivalence of Several Affected Sibship Statistics in Nonparametric Linkage Analysis
Xiaoling Wu*, Johns Hopkins University; Daniel Q. Naiman, Johns Hopkins University
- 2:30 Testing for Homogeneity in Genetic Linkage Analysis
Yuejiao Fu*, University of Waterloo; Jiahua Chen, University of Waterloo; Jack D. Kalbfleisch, University of Michigan
- 2:45 The IBD Distribution of a Sib Pair and the Limiting Distribution of Test Statistic in Asp Linkage Analysis
Jiahua Chan, University of Waterloo; Mary E. Thompson, University of Waterloo; Zeny Z. Feng*, University of Waterloo

- 3:00 Variance-Component Linkage Methods for Complex Trait Data
Michael P. Epstein*, Emory University; Xihong Lin, University of Michigan; Michael Boehnke, University of Michigan
- 3:15 Quantitative Trait Linkage Analysis Using Gaussian Copulas
Mingyao Li*, University of Michigan; Peter X-K Song, York University-Canada; Goncalo R. Abecasis, University of Michigan; Michael Boehnke, University of Michigan

31. CONTRIBUTED PAPERS: SURVEY RESEARCH DATA

[Rivers](#)

Sponsor: ENAR

Chair: Dawei Xie, University of Michigan

- 1:45 Extensions of Mangat's (1994) Randomized-Response Model
Jong-Min Kim*, University of Minnesota at Morris; Joshua M. Tebbs, Kansas State University; Seung-Won An, Oklahoma State University
- 2:00 Use of Auxiliary Variables in Markov Sampling for Finite Population
G.K. Balasubramani*, University of Pittsburgh; Stephen R. Wisniewski, University of Pittsburgh
- 2:15 A Simple Sample Size Formula for Regression-Based Reference Limits
Carine Bellera*, McGill University; James Hanley, McGill University
- 2:30 Issues in the Analysis of Complex Survey Data: Design Based Vs. Model Based Approaches
Marianne H. Bertolet*, Carnegie Mellon University; Howard Seltman, Carnegie Mellon University; Joel Greenhouse, Carnegie Mellon University; Kelly Kelleher, Ohio State University
- 2:45 Flexible Matching in Sample Surveys Containing Observational Studies
Ben B. Hansen*, University of Michigan
- 3:00 Does Weighting for Nonresponse Increase the Variance of Survey Means?
Roderick J. Little*, University of Michigan; Sonya Vartivarian, University of Michigan
- 3:15 Bias Correction in Rare Events Data
Bin Wang*, University of South Alabama; Yu Xu, University of South Alabama; Candice M Ross, University of South Alabama; Kun Huang, University of South Alabama

SCIENTIFIC PROGRAM

32. CONTRIBUTED PAPERS: GENE SELECTION METHODS

Kings Terrace

Sponsors: ENAR/IMS

Chair: Jane Olson, Case Western Reserve University

- 1:45 Comparison of Gene Selection Methods in Class Prediction
Shibing Deng*, University of North Carolina at Chapel Hill; Young Truong, University of North Carolina at Chapel Hill
- 2:00 Probability Based Class Prediction in Toxicogenomics
Nandini Raghavan*, J&J, PRD L.L.C.; Dhammika Amaratunga, J&J, PRD L.L.C.; James J. Colaianne, J&J, PRD L.L.C.
- 2:15 A Semiparametric Approach for Detecting Prognostic Genes Associated with Survival Outcome in Microarray Studies
Kouros Owzar*, Duke University Medical Center; Sin-Ho Jung, Duke University Medical Center; Pranab Kumar Sen, University of North Carolina at Chapel Hill
- 2:30 Modeling Strategies for Neural Networks Based on AIC Criteria and Genetic Algorithms
Doug Landsittel*, University of Pittsburgh; Harshinder Singh, West Virginia University and NIOSH Biostatistics Branch; Vincent C. Arena, University of Pittsburgh; Stewart J. Anderson, University of Pittsburgh
- 2:45 Some New Results in Multivariate Analysis
Samuel W. Wu*, University of Florida; Hongying Li, University of Florida; George Casella, University of Florida
- 3:00 Floor Discussion

- 2:15 Assessing Poisson-Ness: How Fishy is the Mouse?
Michael A. Newton*, University of Wisconsin-Madison; David Hastie, Bristol University
- 2:30 Quantile Dispersion Graphs for Poisson Regression Models
Siuli Mukhopadhyay*, University of Florida; Andre I. Khuri, University of Florida
- 2:45 Anatomical Location of Active Spinal Neurons Using C-Fos Gene Expression
Xiaofeng Wang*, Case Western Reserve University; Kenneth J. Gustafson, Case Western Reserve University and the Cleveland VA FES Center; Michael A. Moffitt, Case Western Reserve University; Scott Snyder, Case Western Reserve University; Warren M. Grill, Case Western Reserve University and the Cleveland VA FES Center; Jiayang Sun, Case Western Reserve University
- 3:00 Estimation of the Change-Point in Generalized Linear Models
Hongling Zhou*, Johns Hopkins University; Kung-Yee Liang, Johns Hopkins University
- 3:15 Tests for Pairwise Multiple Comparisons of Coefficients of Variation
Yvonne M. Zubovic*, Indiana University Purdue University-Fort Wayne; Chand K. Chauhan, Indiana University Purdue University-Fort Wayne

Monday, March 29

3:30–3:45 p.m.

BREAK

Exhibit Hall-Ballroom 2

33. CONTRIBUTED PAPERS: GENERALIZED LINEAR MODELS

Le Bateau

Sponsor: ENAR

Chair: Boris Iglewicz, Temple University

- 1:45 A Hybrid Method for the Analysis of Poisson and Over-Dispersed Poisson Count Data with Zero-Inflation
Grace X. Liu*, Merck, Inc.; Dirk F. Moore, UMDNJ
- 2:00 Investigation of the Behavior of Poisson Regression with Very Small Counts with Application To Mortality Data Analysis
Yue Cui*, University of Minnesota; Melanie M Wall, University of Minnesota

SCIENTIFIC PROGRAM

Monday, March 29

3:45–5:30 p.m.

34. NONPARAMETRIC REGRESSION TECHNIQUES APPLIED TO ANALYSIS OF SURVIVAL DATA *Kings Garden North*

Sponsors: IMS/ENAR/ASA Biometrics Section
Organizer: Chin-Shang Li, St. Jude Children's Research Hospital
Chair: Chin-Shang Li, St. Jude Children's Research Hospital

- 3:45 Marginal Hazard Models with Varying-Coefficients for Multivariate Failure Time Data
Jianwen Cai*, University of North Carolina at Chapel Hill; Jianqing Fan, Princeton University; Haibo Zhou, University of North Carolina at Chapel Hill; Yong Zhou, University of North Carolina at Chapel Hill
- 4:15 Estimation in Accelerated Failure Models
Robert L. Strawderman*, Cornell University
- 4:45 Analysis of the Additive Hazards Model with Measurement Errors - Nonparametric Correction Method
Jianguo (Tony) Sun*, University of Missouri; Liuquan Sun, University of Missouri
- 5:15 Floor Discussion

35. STATISTICS IN GENETICS *Kings Garden South*

Sponsors: ASA Biopharmaceutical Section/ENAR
Organizer: Jinyu Yuan, Merck Research Laboratories
Chair: Jinyu Yuan, Merck Research Laboratories

- 3:45 Human Quantitative Trait Locus Mapping with Selected Samples
Eleanor Feingold*, University of Pittsburgh; Jin P. Szatkiewicz, University of Pittsburgh
- 4:10 Assessing the Potential Impact of Population Stratification and Inbreeding on Commonly Used Statistical Methods for Association Studies
Zhaohai Li*, George Washington University; Joseph Gastwirth, George Washington University; Mitchell Gail, National Cancer Institute, National Institutes of Health
- 4:35 SNP Selection and Haplotype Analysis in Identifying Disease Genes
Hongyu Zhao*, Yale University
- 5:00 Genetic Dissection of Complex Traits
Kai Yu*, Washington University
- 5:25 Floor Discussion

36. ASSESSING EFFICACY IN A CHANGING BACKGROUND *Sterlings 2 & 3*

Sponsor: ASA Section on Risk Analysis
Organizer: David Banks, Duke University
Chair: Vincent C. Arena, University of Pittsburgh

- 3:45 Heterogeneity of Efficacy Assessment Across Space and Time: Statistical and Healthcare Utilization Implications
Demissie Alemayehu*, Pfizer and Columbia University
- 4:15 Comparing Hypertensive Therapies Over Time and Place
Thomas Lumley*, University of Washington
- 4:45 Lessons in Drug Efficacy Comparisons from Metrology
David Banks*, Duke University
- 5:15 Floor Discussion

37. THE CAUSES AND CONSEQUENCES OF HIPAA: A STATISTICIAN'S PERSPECTIVE (HIPAA: HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT) *Ballroom 3*

Sponsor: ENAR
Organizers: Sumithra J. Mandrekar, Ph.D., Mayo Clinic and Jayawant N. Mandrekar, Ph.D., Mayo Clinic
Chairs: Sumithra J. Mandrekar, Ph.D., Mayo Clinic and Jayawant N. Mandrekar, Ph.D., Mayo Clinic

- 3:45 HIPAA Data Standards and Health Information: The Good, The Bad, and The Ugly
Christopher G. Chute*, Mayo Clinic
- 4:10 Practical Issues of Sharing Data and Biological Specimens Under HIPAA
Gerald J. Beck*, Cleveland Clinic Foundation
- 4:35 Sharing and Publication of Data on Individual Patients: Scientific, Ethical and Regulatory Issues
Andrew J. Vickers*, Memorial Sloan-Kettering Cancer Center
- 5:00 Discussant: David L. DeMets, Department of Biostatistics and Medical Informatics
- 5:25 Floor Discussion

SCIENTIFIC PROGRAM

38. INFORMATION RELATED PROBLEMS IN BIOLOGY

Sponsor: IMS

Organizer: Xiao-Li Meng, Harvard University

Chair: Daniel Nicolae, University of Chicago

[Benedum](#)

- 3:45 Informativeness of Genetic Markers for Inference of Ancestry
Noah A. Rosenberg, University of Southern California; Lei Li, University of Southern California; Ryk Ward, University of Oxford; Jonathan K. Pritchard*, The University of Chicago
- 4:15 Information-Driven Marker Selection for Large Scale Genomic Studies on Complex Traits
Tian Zheng*, Columbia University; Shaw-Hwa Lo, Columbia University
- 4:45 A Bayesian Framework for Quantifying Incomplete Information in Statistical and Genetic Hypothesis Testing
Augustine Kong, deCode Genetics, Iceland; Xiao-Li Meng*, Harvard University; Dan Nicolae, The University of Chicago
- 5:15 Floor Discussion

39. CONTRIBUTED PAPERS: MISSING DATA METHODS

Sponsor: ENAR

Chair: Andrea Troxel, University of Pennsylvania

[Le Bateau](#)

- 3:45 **Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates**
Lan Huang*, University of Connecticut; Ming-Hui Chen, University of Connecticut; Joseph G. Ibrahim, University of North Carolina
- 4:00 Semiparametric Maximum Likelihood for Missing Covariates in Parametric Regression
Zhiwei Zhang*, University of Pittsburgh; Howard E. Rockette, University of Pittsburgh
- 4:15 Semiparametric Models for Missing Covariates in Generalized Linear Mixed Models
Qingxia Chen*, Biostatistics University of North Carolina at Chapel Hill; Joseph Ibrahim, University of North Carolina at Chapel Hill
- 4:30 Inference About the Propensity Penalized Spline Prediction Method for Multivariate Data with Missing Values
Hyonggin An*, University of Michigan; Roderick J. Little, University of Michigan

4:45

Semiparametric Bayesian Analysis of Matched Case Control Studies with Missing Exposure

Samiran Sinha*, University of Florida; Bhramar Mukherjee, University of Florida; Malay Ghosh, University of Florida; Bani K. Mallick, Texas A&M University; Raymond J. Carroll, Texas A&M University
Imputation of Missing Genotypic and Biomarker Data to Improve a Risk Model for Lung Cancer
Carol J Etzel*, M.D. Anderson Cancer Center; Mei Lu, M.D. Anderson Cancer Center; Hui Zhou, M.D. Anderson Cancer Center; Qingy Wei, M.D. Anderson Cancer Center; Xifeng Wu, M.D. Anderson Cancer Center; Margeret Spitz, M.D. Anderson Cancer Center

5:00 Conditional Inference Methods for Poisson Regression Models with Missing Data and Feedback
Jason Roy*, University of Rochester; Don Alderson, Brown University; Joseph W. Hogan, Brown University; Karen Tashima, Brown University

5:00

5:15

40. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS I

Sponsor: ENAR

Chair: Tony Sun, University of Missouri

[Rivers](#)

- 3:45 Analysis of Time-to-Event Data with Incomplete Event Adjudication
Thomas D. Cook*, University of Wisconsin, Madison; Michael R. Kosorok, University of Wisconsin, Madison
- 4:00 Parametric Estimation of Cumulative Incidence Function
Jong-Hyeon Jeong*, University of Pittsburgh
- 4:15 Failure Time Analysis of HIV Vaccine Effects on Viral Load and Treatment Initiation
Peter B. Gilbert, Fred Hutchinson Cancer Research Center and University of Washington; Yanqing Sun*, University of North Carolina at Charlotte
- 4:30 A Sample Size Formula for the Supremum Log-Rank Statistic
Kevin H. Eng, Brown University; Michael R. Kosorok*, University of Wisconsin-Madison
- 4:45 Power Consideration in Survival Analysis with Uncertain Endpoints
Hongwei Wang*, Rutgers University; Cong Chen, Merck Research Laboratories; Steven M. Snapinn, Merck Research Laboratories
- 5:00 Joint Modeling Survival and Recurrent Event Data
Xuelin Huang*, M.D. Anderson Cancer Center
- 5:15 Semiparametric Efficient Estimation of Survival Distribution in Two-Stage Randomization Designs
Abdus S. Wahed*, University of Pittsburgh; Anastasios A. Tsiatis, North Carolina State University

SCIENTIFIC PROGRAM

41. CONTRIBUTED PAPERS: ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

[Brigade](#)

Sponsor: ENAR

Chair: Paul Rathouz, University of Chicago

- 3:45 A Comparative Study of the Use of GAM and GLM in Air Pollution Research
Shui He*, University of Pittsburgh; Sati Mazumdar, University of Pittsburgh; Vincent C. Arena, University of Pittsburgh
- 4:00 Relating Pm2.5 Concentration Levels to Mortality Using an Exposure Simulator
Catherine Calder*, Ohio State University; Christopher Holloman, Battelle Memorial Institute; Steven Bortnick, Battelle Memorial Institute; Warren Strauss, Battelle Memorial Institute; Michele Morara, Battelle Memorial Institute
- 4:15 A Simple Statistical Air Dispersion Model and Its Prediction
Tae-Young Heo*, North Carolina State University; Jacqueline M. Hughes-Oliver, North Carolina State University
- 4:30 Biologically Based Model for Estimating the Formation Rate of Biomarker from Butadiene Air Exposure: Frequentist and Bayesian Approaches
Sohee Park*, Harvard School of Public Health; Louise M. Ryan, Harvard School of Public Health
- 4:45 Linear Regression with Type I Interval- and Left-Censored Data
Mary-Lou Thompson, University of Washington; Kerrie P. Nelson*, University of South Carolina
- 5:00 Floor Discussion

42. CONTRIBUTED PAPERS: LONGITUDINAL DATA ANALYSIS I

[Duquesne](#)

Sponsor: ENAR

Chair: Wensheng Guo, University of Pennsylvania

- 3:45 A Restricted Multivariate Random-Effects Model with Applications to Assessing Radiation Therapy for Brain Tumors
Hongbin Fang*, University of Maryland; Guoliang Tian, University of Maryland; Xiaoping Xiong, St. Jude Children's Research Hospital; Ming Tan, University of Maryland
- 4:00 Extending Historical Functional Linear Model Using Penalized Regression Splines
Jaroslaw Harezlak*, Harvard School of Public Health; Nan M. Laird, Harvard School of Public Health

- 4:15 Modeling Unequally Spaced Bivariate Growth Curve with a Kalman Filter Approach
Qianyu Dang*, University of Pittsburgh; Stewart J. Anderson, University of Pittsburgh; Sati Mazumdar, University of Pittsburgh
- 4:30 A Latent Class Regression Model for Longitudinal Ordinal Outcomes
Don Hedeker*, University of Illinois at Chicago
- 4:45 Latent Predictors of Repeated Clinical Measures – Application to Behavioral Risk Factors and Lung Function Among Youth with Cystic Fibrosis
Melanie M. Wall*, University of Minnesota; Joan Patterson, University of Minnesota
- 5:00 A Latent Change-point Model Using a Generalized Estimating Equations Approach
Ji-Hyun Lee*, University of South Florida; Bahjat F. Qaqish, University of North Carolina at Chapel Hill
- 5:15 Floor Discussion

43. CONTRIBUTED PAPERS: CLUSTERING AND CLASSIFICATION METHODS

[Heinz](#)

Sponsors: ENAR/IMS

Chair: Hao Helen Zhang, North Carolina State University

- 3:45 Correcting The Loss of Cell-Cycle Synchrony in Clustering Analysis of Microarray Data Using Weights
Fenghai Duan*, Yale University; Heping Zhang, Yale University
- 4:00 An Up and Down Regulation Directed Clustering of Temporal Gene Expressions
Lang Li*, Indiana University; Stephen J. Metthew, Indiana University; Howard Edenberg, Indiana University
- 4:15 Building Mixture Trees from Binary Sequence Data
Shu-Chuan Chen*, National Health Research Institutes Taiwan, ROC; Bruce G. Lindsay, Penn State University
- 4:30 CATS: Clustering After Transformation and Smoothing
Nicoleta Serban*, Carnegie Mellon University; Larry Wasserman, Carnegie Mellon University
- 4:45 Sets of Complementary Lack of Fit Tests Based on Edge Clusters
James W. Neill*, Kansas State University; Forrest R. Miller, Kansas State University
- 5:00 Floor Discussion

SCIENTIFIC PROGRAM

44. CONTRIBUTED PAPERS: MIXTURE MODELS AND QT CORRECTIONS

Kings Terrace

Sponsors: ENAR/IMS

Chair: Paul Albert, National Institutes of Health

- 3:45 A Bayesian Mixture Model Relating Dose to Critical Organs and Functional Complication in 3d Conformal Radiation Therapy
Timothy D. Johnson*, University of Michigan; Jeremy M.G. Taylor, University of Michigan
- 4:00 Hypothesis Testing in Mixture Regression Models
Hongtu Zhu*, Columbia University; Heping Zhang, Yale University
- 4:15 A Non-Linear Mixed Model of Parkinson's Disease Progression
Paulo Guimaraes*, Medical University of South Carolina; Jordan Elm, Medical University of South Carolina; Peng Huang, Medical University of South Carolina; Yuko Palesch, Medical University of South Carolina; Karl Kiebertz, University of Rochester; Barbara Tilley, Medical University of South Carolina
- 4:30 Risk-Analysis Based Mixture Model Selection
Surajit Ray*, University of North Carolina; Bruce G. Lindsay, Pennsylvania State University
- 4:45 Asymptotic Properties of Minimum Distance Estimator in Beta Mixture Distributions
Xuejun Peng*, Cleveland Clinic Foundation; Arnold Stromberg, University of Kentucky; Constance Wood, University of Kentucky
- 5:00 Optimal Data-Based Heart Rate Correction for QTL
Rene Kubiak*, Boehringer Ingelheim Pharmaceuticals, Inc.; Steve Weldon, Boehringer Ingelheim Pharmaceuticals, Inc.
- 5:15 Historical-Based Formula for Heart-Rate Corrected QT Intervals in Beagle Dogs
Kevin Guo*, Eli Lilly and Company; Wendell C. Smith, Eli Lilly and Company

Tuesday, March 30

8:30–10:15 a.m.

45. WHAT MORE NEEDS TO BE DONE IN THE GRADUATE TRAINING OF STATISTICIANS?

Sterlings 2 & 3

Sponsors: ASA Section on Statistical Education/
ASA Biometrics Section

Organizer: Linda Young, University of Florida

Chair: Linda Young, University of Florida

- 8:30 From Graduate School to Industry: Softening the Blow
Janet Wittes*, Statistics Collaborative, Inc.
- 9:00 Preparing Statisticians for Collaborative and Independent Research at the National Institutes of Health
Mitchell H. Gail*, National Cancer Institute
- 9:30 The Exemplary Biostatistician in Academic Medical Centers
Ralph G. O'Brien*, Cleveland Clinic Foundation
- 10:00 Floor Discussion

46. METHODS FOR COMBINING GENE EXPRESSION DATASETS FROM DIFFERENT STUDIES

Kings Garden North

Sponsor: ENAR

Organizer: Elizabeth Garrett, Johns Hopkins University, Sidney Kimmel Comprehensive Cancer Center

Chair: Elizabeth Garrett, Johns Hopkins University, Sidney Kimmel Comprehensive Cancer Center

- 8:30 Comparing Mouse Tissue-Specific Gene Expression Datasets
Timothy R. Hughes*, University of Toronto; Wen Zhang, University of Toronto; Quaid Morris, University of Toronto
- 8:55 Large-Scale Analysis of Gene Coexpression
Paul Pavlidis*, Columbia University
- 9:20 Multi-Study Genomic Data Analysis
Giovanni Parmigiani*, Johns Hopkins University; Elizabeth S. Garrett, Johns Hopkins University; Edward Gabrielson, Johns Hopkins University
- 9:45 Statistical Methods for Combining Data from Multiple Cancer Studies
Debashis Ghosh*, University of Michigan
- 10:10 Floor Discussion

SCIENTIFIC PROGRAM

47. RECENT DEVELOPMENTS IN SURVIVAL ANALYSIS

[Ballroom 3](#)

Sponsors: ENAR/IMS/ASA Biometrics Section

Organizer: Danyu Lin, University of North Carolina

Chair: Danyu Lin, University of North Carolina

- 8:30 Aspects of the Analysis of Multivariate Failure Time Data
Ross Prentice*, University of Washington
- 9:00 An Overview of Joint Modeling of Survival and Longitudinal Covariates Measured with Error
Anastasios A. Tsiatis*, North Carolina State University; Marie Davidian, North Carolina State University
- 9:30 Recent Developments for the Linear Regression Model with Censored Event Time Data
L.J. Wei*, Harvard University School of Public Health
- 10:00 Discussant: David Oakes, University of Rochester

48. NEW ADVANCES IN NONPARAMETRIC STATISTICS

[Kings Garden South](#)

Sponsor: IMS

Organizer: Wei Biao Wu, The University of Chicago

Chair: David Shera, Children's Hospital of Philadelphia/ University of Pennsylvania

- 8:30 Comparing Nonparametric Procedures
Yang Yuhong*, Iowa State University
- 8:50 Spline Estimation in Partially Linear Models for Longitudinal Data
Jianhua Huang*, The University of Pennsylvania Wharton School
- 9:15 Image Segmentation for cDNA Microarray Data Based on Local Linear Kernel Smoothing
Peihua Qiu*, University of Minnesota; Jingran Sun, University of Minnesota
- 9:35 State of the Art in Nonparametric Lack-of-Fit Tests
Jeffrey D. Hart*, Texas A&M University
- 9:55 Kernel Estimation for Stationary Processes
Wei Biao Wu*, The University of Chicago

49. SPECIAL CONTRIBUTED SESSION: COMBINING DATA FROM DISPARATE SURVEYS TO IMPROVE SMALL AREA ESTIMATION

[Benedum](#)

Sponsors: ENAR/ASA Section on Survey Research Methods

Organizers: Trivellore E. Raghunathan, University of Michigan and Michael R.. Elliott, University of Pennsylvania School of Medicine

Chair: Trivellore E. Raghunathan, University of Michigan

- 8:30 Combining Information from Multiple Surveys for Small-Area Estimation: A Bayesian Approach
Trivellore E. Raghunathan, University of Michigan; Dawei Xie, University of Michigan; Nathaniel Schenker*, National Center for Health Statistics; Van L. Parsons, National Center for Health Statistics; William W. Davis, National Cancer Institute; Eric J. Feuer, National Cancer Institute; Kevin W. Dodd, National Cancer Institute
- 8:55 Modeling the County-Level Response Rate for the Behavioral Risk Factors Surveillance System (BRFSS)
William W. Davis*, National Cancer Institute; Ali H. Mokdad, Centers for Disease Control and Prevention; Machell Town, Centers for Disease Control and Prevention; Eric J. Feuer, National Cancer Institute
- 9:20 A Bayesian Approach for Combining Information from Multiple Surveys in Small Area Estimation Using Public-Use Data
Dawei Xie*, University of Michigan; Trivellore E. Raghunathan, University of Michigan
- 9:45 Obtaining Cancer Risk Factor Prevalence Estimates in Small Areas: Combining Data from the BRFSS and the NHIS
Michael R. Elliott*, University of Pennsylvania
- 10:10 Floor Discussion

50. CONTRIBUTED PAPERS: DISEASE DETECTION AND SURVEILLANCE

[Duquesne](#)

Sponsor: ENAR

Chair: Ken Kleinman, Harvard Pilgrim Healthcare

- 8:30 Estimating Incidence Rates Using Pedigree Sampling
Steven N. MacEachern, Ohio State University; Amy K. Ferketich*, Ohio State University; Stanley Lemeshow, Ohio State University; Lei Shen, Ohio State University; Judith A. Westman, Ohio State University; Albert de la Chapelle, Ohio State University; Clara D. Bloomfield, Ohio State University
- 8:45 HIV/AIDS Projection- A Back Calculation Approach
Anbupalam Thalamuthu*, University of Pittsburgh; Ramanan Ramanujam, Presidency College, Chennai, India
- 9:00 A Dynamic Graphical Model for Influenza Surveillance
Paola Sebastiani*, Boston University School of Public Health; Ken D. Mandy, Harvard Medical School; Peter Szolovits, Massachusetts Institute of Technology; Isaac S. Kohane, Harvard Medical School; Marco F. Ramoni, Harvard Medical School

SCIENTIFIC PROGRAM

- 9:15 Constrained Bayes Estimates of Random Effects When Data are Subject to a Detection Limit
Renee H. Moore*, Emory University; Robert H. Lyles, Emory University; Amita K. Manatunga, Emory University
- 9:30 Validity of Short-Term Beta-Agonists Prescriptions as a Marker for Asthma
Xiaoming Bao*, University of Chicago; Paul J. Rathouz, University of Chicago; Vanja M. Dukic, University of Chicago
- 9:45 Multiple Cutpoint Selection for Categorizing a Continuous Predictor
Sean M. O'Brien*, National Institute of Environmental Health Sciences
- 10:00 Infectious Disease Surveillance Scheme C: A Multivariate Approach
Bo Hong*, University of Alabama; Michael J. Hardin, University of Alabama

51. CONTRIBUTED PAPERS: BAYESIAN SURVIVAL ANALYSIS

[Rivers](#)

Sponsor: ENAR

Chair: Xiao-Li Meng, Harvard University

- 8:30 **Joint Longitudinal-Survival-Cure Models and Their Application to Prostate Cancer**
Menggang Yu*, University of Michigan; Ngayee J. Law, Roche Molecular Systems, Inc.; Jeremy M.G. Taylor, University of Michigan; Sandler M. Howard, University of Michigan
- 8:45 Bayesian Analysis of Constrained Hazard Functions
Gregg E. Dinse*, National Institute of Environmental Health Sciences; David B. Dunson, National Institute of Environmental Health Sciences
- 9:00 A Class of Bayesian Box-Cox Transformation Hazard Regression Models
Guosheng Yin*, M. D. Anderson Cancer Center; Joseph G. Ibrahim, University of North Carolina at Chapel Hill
- 9:15 Bivariate Positive Stable Frailty Models
Madhuja Mallick*, University of Connecticut; Nalini Ravishanker, University of Connecticut
- 9:30 Proportional Mean Regression Models for Censored Data
Sujit K. Ghosh*, North Carolina State University; Subhashis Ghosal, North Carolina State University
- 9:45 Selecting Factors Predictive of Heterogeneity in Multivariate Event Time Data
David B. Dunson, National Institute of Environmental Health Sciences; Zhen Chen*, University of Pennsylvania
- 10:00 Floor Discussion

52. CONTRIBUTED PAPERS: DIAGNOSTIC AND SCREENING TESTS

[Brigade](#)

Sponsor: ENAR

Chair: Dennis Boos, North Carolina State University

- 8:30 Evaluation of Methods for Assessing the Performance of a New Test When There is Imperfect Gold Standard
H.Refik Burgut*, Cukurova University; Macit U. Sandikci, Cukurova University; Fatin Kocak, Cukurova University
- 8:45 Comparison of QT Prolongation Using Rodent Models
Kanaka D. Tatikola*, Johnson & Johnson Pharmaceutical Research and Development, L.L.C.; Jaya Natarajan, Johnson & Johnson Pharmaceutical Research and Development, L.L.C.; James J. Colaianne, Johnson & Johnson Pharmaceutical Research and Development, L.L.C.; Quan Li, Johnson & Johnson Pharmaceutical Research and Development, L.L.C
- 9:00 Use of Receiver Operating Characteristic Curve in Medical Decision Making
Jiping Wang*, University of Pittsburgh; Howard E. Rockette, University of Pittsburgh
- 9:15 Combining Correlated Diagnostic Tests Application to Neuropathologic Diagnosis of Alzheimer's Disease
Chengjie Xiong*, Washington University in St. Louis; Daniel W. McKeel, Jr., Washington University in St. Louis; J. Phillip Miller, Washington University in St. Louis; John C. Morris, Washington University in St. Louis
- 9:30 Multi-Reader, Multi-Case ROC Analysis: An Empirical Comparison of Five Methods
Nancy A. Obuchowski*, Cleveland Clinic Foundation; Sergey V. Beiden, U.S. Food and Drug Administration; Kevin S. Berbaum, University of Iowa; Stephen Hillis, Iowa City VA Medical Center; Hemant Ishwaran, Cleveland Clinic Foundation; Hae Hiang Song, Catholic University Medical College, Seoul, Korea; Robert F. Wagner, U.S. Food and Drug Administration
- 9:45 Equivalent Variations of the Dorfman-Berbaum-Metz and Obuchowski-Rockette Methods for ROC Data
Stephen L. Hillis*, Iowa City VA Medical Center; Nancy A. Obuchowski, Cleveland Clinic Foundation; Kevin M. Schartz, University of Iowa; Kevin S Berbaum, University of Iowa
- 10:00 Simultaneous Inference For Prevalence Using Pooled Assessments
Joshua M. Tebbs*, Kansas State University; Melinda H. McCann, Oklahoma State University

SCIENTIFIC PROGRAM

53. CONTRIBUTED PAPERS: SEQUENCE ANALYSIS AND PROTEOMICS

[Kings Terrace](#)

Sponsor: ENAR

Chair: Michael Newton, University of Wisconsin

- 8:30 Some Statistical Aspects of the Analysis of Genomic Sequences
Lily Wang*, University of North Carolina at Chapel Hill; Kumar Sen, University of North Carolina at Chapel Hill
- 8:45 Modeling Nucleotide Sequence Variation in a Viral Quasispecies
Philip M. Dixon*, Iowa State University; Ozlem Ilk, Iowa State University
- 9:00 Comparison and Validation of Protein Identification Searching Tools Using Mass Spectrometry Data
JungBok Lee*, Korea University; Jae Won Lee, Korea University; Mira Park, Eulji University
- 9:15 A Functional Data Approach to MALDI-TOF MS Protein Analysis
Dean Billheimer*, Vanderbilt University
- 9:30 Model-Based Assignment and Inference of Protein Backbone Nuclear Magnetic Resonance
Olga Vitek*, Purdue University; Jan Vitek, Purdue University; Bruce A. Craig, Purdue University; Chris Bailey-Kellogg, Purdue University
- 9:45 Systematic Variation in Difference Gel Electrophoresis Data
Kimberly F. Sellers*, Carnegie Mellon University; Jeffrey Miecznikowski, Carnegie Mellon University; Surya Viswanathan, Carnegie Mellon University; William F. Eddy, Carnegie Mellon University; Jonathan S. Minden, Carnegie Mellon University
- 10:00 Wavelet Based Analysis of Homologous DNA Sequences for Four Mammalian Species
Jean Roayaei*, U.S. Food and Drug Administration, CDRH; Juan-Miguel Marin, Universidad Rey Juan Carlos Madrid, Spain

54. CONTRIBUTED PAPERS: SEQUENTIAL TRIALS, INTERIM DECISION MAKING, AND MULTIPLICITY

[Le Bateau](#)

Sponsor: ENAR

Chair: Janet Wittes, Statistics Collaborative Inc.

- 8:30 The Design and Implementation of Group Sequential Trials for Longitudinal Endpoints
John M. Kittelson*, University of Colorado; Katrina Sharples, University of Otago Dunedin, New Zealand; Scott S. Emerson, University of Washington Seattle
- 8:45 Confidence Interval Estimation of the Success Rate in Exact Group Sequential Designs
Jihneeh Yu*, Roswell Park Cancer Institute; James L. Kepner, SUNY at Buffalo
- 9:00 Software for Design and Analysis of Group Sequential Clinical Trials with Multiple Primary Endpoints
Shuangge Ma*, University of Wisconsin; Michael R. Kosorok, University of Wisconsin; Thomas D. Cook, University of Wisconsin
- 9:15 Programming Strategies for Moderately Complicated Clinical Research Studies
Martin C. Weinrich*, University of Louisville
- 9:30 Simultaneous Group Sequential Analysis of Rank-Based and Weighted Kaplan-Meier Tests for Paired Censored Survival Data
Adin-Cristian Andrei*, University of Michigan; Susan Murray, University of Michigan
- 9:45 Timing of Futility Analyses for 'Proof of Concept' Trials
A. Lawrence Gould*, Merck Research Laboratories
- 10:00 Performance of Some Multiple Testing Procedures to Compare Three Doses of a Test Drug and Placebo
Lan Kong*, University of Pittsburgh; Gary G Koch, University of North Carolina

Tuesday, March 30

10:15–10:30 a.m.

BREAK

[Exhibit Hall-Ballroom 2](#)

SCIENTIFIC PROGRAM

Tuesday, March 30

10:30 a.m.–12:15 p.m.

PRESIDENTIAL INVITED ADDRESS

[Ballroom 1](#)

Sponsor: ENAR

Organizer/Chair: Marie Davidian, North Carolina State University

- 10:30 Introduction: Marie Davidian, North Carolina State University
- 10:35 Distinguished Student Paper Awards: Carol Redmond, University of Pittsburgh
- 10:45 Statisticians: Uncertainty Experts? Louise Ryan, Harvard University School of Public Health

Tuesday, March 30

1:45–3:30 p.m.

55. STATISTICAL METHODS IN COMPUTER VISION

[Kings Garden North](#)

Sponsor: ENAR

Organizer: Ralph Gross, Carnegie Mellon University

Chair: Ralph Gross, Carnegie Mellon University

- 1:45 Extended Multi-Linear Models for Computer Vision Fernando De la Torre*, Carnegie Mellon University; Takeo Kanade, Carnegie Mellon University; Michael Black, Brown University
- 2:10 Vision Systems Engineering and Analysis – A Unified Perspective Visvanathan Ramesh*, Siemens Corporate Research
- 2:35 Empirical Statistical Models for Object Recognition in Photographs Henry Schneidermann*, Carnegie Mellon University
- 3:00 Detecting Unusual Human Behavior in Video Jianbo Shi*, University of Pennsylvania
- 3:25 Floor Discussion

56. NEW DEVELOPMENTS ON ESTIMATING EQUATIONS

[Kings Garden South](#)

Sponsors: ENAR/IMS

Organizer: Jianhua Huang, University of Pennsylvania

Chair: Jianhua Huang, University of Pennsylvania

- 1:45 Distribution-Based Marginal Regression Models for Longitudinal Data Jianhua Z. Huang, University of Pennsylvania; Colin O. Wu*, National Heart, Lung and Blood Institute; Lan Zhou, University of Pennsylvania

- 2:10 Estimating Equations for Longitudinal Data with Time-Dependent Covariates Dylan S. Small*, University of Pennsylvania; Tze L. Lai, Stanford University
- 2:35 Implementation of a Complex Correlation Structure Appropriate for Repeated Bouts of Measurements in a Longitudinal Study Justine N. Shults*, University of Pennsylvania; Carissa A. Mazurick, Merck & Co, Inc; Richard Landis, University of Pennsylvania
- 3:00 Consistency and Inconsistency of Maximum Quasi Likelihood Estimators Bing Li*, Pennsylvania State University
- 3:25 Floor Discussion

57. BAYESIAN METHODS IN GENOMICS

[Ballroom 3](#)

Sponsors: ENAR/IMS

Organizer: Marina Vannucci, Texas A&M University

Chair: Marina Vannucci, Texas A&M University

- 1:45 Variable Selection for Gene Expression Data with Time-to-Event Responses Najjun Sha*, The University of Texas at El Paso; Marina Vannucci, Texas A&M University; Mahlet Tadesse, Texas A&M University
- 2:10 New Classes of Mixture Models for Analyzing DNA Microarray Data Joseph G. Ibrahim*, University of North Carolina; Ming-Hui Chen, University of Connecticut
- 2:35 A Bayesian Mixture Model for Differential Gene Expression Kim-Anh Do, M.D. Anderson Cancer Center; Peter Mueller*, M.D. Anderson Cancer Center; Feng Tang, M.D. Anderson Cancer Center
- 3:00 A Bayesian Method for Simultaneous Class Discovery and Gene Selection Mahlet G. Tadesse*, Texas A&M University
- 3:25 Floor Discussion

SCIENTIFIC PROGRAM

58. NONPARAMETRIC FUNCTION ESTIMATION

Sponsors: ENAR/IMS

Organizers: Nicoleta Serban, Carnegie Mellon University; Larry Wasserman, Carnegie Mellon University and Christopher Genovese Carnegie Mellon University

Chair: Nicoleta Serban, Carnegie Mellon University

- 1:45 Functional Adaptive Model Estimation
Gareth M. James*, University of Southern California;
Bernard W. Silverman, Oxford University
- 2:15 High Dimension Low Sample Size Survival Analysis
Brent Johnson, University of North Carolina at
Chapel Hill; Danyu Lin, University of North Carolina
at Chapel Hill; J. S. Marron*, University of North
Carolina at Chapel Hill
- 2:45 Reconstruction, Alignment, and Comparison of
Continuous Time Course Expression Profiles
Ziv Bar-Joseph, Carnegie Mellon University; Georg
Gerber, Massachusetts Institute of Technology;
Itamar Simon, Hebrew University; David Gifford,
Massachusetts Institute of Technology; Tommi Jaakkola*,
Massachusetts Institute of Technology
- 3:15 Floor Discussion

59. STOCHASTIC MODELING AND INFERENCE IN SCIENCE

Sponsor: IMS

Organizer: Samuel Kou, Harvard University

Chair: Chengcheng Hu, Harvard University

- 1:45 Model Selection in Irregular Problems: Application
to Mapping QTL
David O. Siegmund*, Stanford University
- 2:10 Spatial and Spatio-Temporal Hierarchical Modeling
for the Analysis of Spatial Data
Alan E. Gelfand*, Duke University
- 2:35 Normalization and Significant Analysis of cDNA
Micro-Arrays
Jianqing Fan*, Princeton University
- 3:00 Bayesian Analysis of Single Molecule Experiments
Samuel Kou*, Harvard University; Sunney Xie,
Harvard University; Jun Liu, Harvard University
- 3:25 Floor Discussion

60. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS II

Sponsors: ENAR/IMS

Chair: Jack Kalbfleisch, University of Michigan

- 1:45 Empirical Likelihood Ratio with Arbitrary Censored/
Truncated Data by Constrained EM Algorithm
Min Chen*, University of Kentucky; Jingyu Luan,
University of Kentucky; Mai Zhou, University of
Kentucky
- 2:00 On the Analysis of Informatively Interval-Censored
Discrete Failure-Time Data: A Sensitivity Analysis
Approach
Michelle D. Shardell*, Johns Hopkins University;
Daniel O. Scharfstein, Johns Hopkins University;
Noya Galai, Johns Hopkins University; David Vlahov,
Johns Hopkins University
- 2:15 Parametric Estimation for Additive Rate Regression
Models
M. Brent McHenry*, University of Pittsburgh; Stuart
R. Lipsitz, Medical University of South Carolina
- 2:30 Empirical Likelihood Ratio with Right-Censoring and
Left-Truncation Data
Jingyu Luan*, University of Kentucky; Min Chen,
University of Kentucky; Mai Zhou, University of
Kentucky
- 2:45 **Survival Analysis Using Auxiliary Variables Via
Nonparametric Multiple Imputation**
Chiu-Hsieh Hsu*, University of Arizona; Jeremy M.G.
Taylor, University of Michigan; Susan Murray,
University of Michigan
- 3:00 Reducing Bias in Parameter Estimates from Stepwise
Regression in Proportional Hazards Regression with
Right-Censored Data
Chang-Heok Soh*, Dana-Farber Cancer Institute and
Harvard School of Public Health; David P.
Harrington, Dana-Farber Cancer Institute and
Harvard School of Public Health
- 3:15 On Semiparametric Transformation Cure Models
Wenbin Lu*, North Carolina State University;
Zhiliang Ying, Columbia University

SCIENTIFIC PROGRAM

61. CONTRIBUTED PAPERS: EXPERIMENTAL DESIGN

[Le Bateau](#)

Sponsor: ENAR

Chair: Ralph O'Brien, Cleveland Clinic

- 1:45 Two At A Time? Or All At Once?
Donald J. Schuirmann*, Food and Drug Administration
- 2:00 Estimating the Variance Components in a Replicated Crossover Study
Guoyong Jiang*, Cephalon, Inc.; Jianrong Wu, St. Jude Children's Research Hospital
- 2:15 On the Use of a Crossover Design in Vaccine Research
Jonathan B. Skinner*, Wyeth Vaccines Research
- 2:30 Carryover Effects Examined in Collections of Crossover Studies
Mary E. Putt*, University of Pennsylvania
- 2:45 A Beta-Binomial Model for Cross-Over Studies of Correlated Bernoulli Outcomes, with an Application to Condom Failure Data
Douglas J. Taylor*, Family Health International; Natalie W. Cheung, Family Health International; Mark A. Weaver, Family Health International
- 3:00 Optimal Designs for Ordinal Outcomes Using Beta Regression Models
Yuehui Wu*, University of Pennsylvania; Valeri V. Fedorov, GlaxoSmithKline; Kathleen J. Propert, University of Pennsylvania
- 3:15 Cross Trial Statistical Inference: Bridging Vs. Non-Inferiority Scenarios
Sue-Jane Wang*, Food and Drug Administration

62. CONTRIBUTED PAPERS: ADVANCES IN QTL ANALYSIS

[Kings Terrace](#)

Sponsor: ENAR

Chair: Ruzong Fan, Texas A&M University

- 1:45 **Mapping Quantitative Trait Loci with Censored Observations**
Guoqing Diao*, University of North Carolina at Chapel Hill; Danyu Lin, University of North Carolina at Chapel Hill; Fei Zou, University of North Carolina at Chapel Hill
- 2:00 Bayesian Model and Variable Selection Approaches to Mapping Multiple Quantitative Trait Loci
Kyoungmi Kim*, University of Alabama at Birmingham; Nengjun Yi, University of Alabama at Birmingham; Alfred Bartolucci, University of Alabama at Birmingham; David B. Allison, University of Alabama at Birmingham

- 2:15 An Empirical Bayes Approach for Expression Trait Loci (ETL) Mapping
Christina Kendzioriski*, University of Wisconsin
- 2:30 A Nonparametric Test Against Stochastic Ordering with Applications in QTL Mapping
Fei Zou*, University of North Carolina at Chapel Hill; Lei Nie, UMBC
- 2:45 **A Model Selection Based Interval Mapping Method for Autopolyploids**
Dachuang Cao*, Purdue University; Bruce A. Craig, Purdue University; Rebecca W. Doerge, Purdue University
- 3:00 A Unified Semiparametric Framework for QTL Analyses, with Application to 'Spike' Phenotypes
Chunfang Jin*, University of Wisconsin-Madison
- 3:15 A Two-Stage Hierarchical Model for Detecting the Epistatic Control of Maternal-Offspring QTL Over Endosperm Development
Yuehua Cui*, University of Florida; George Casella, University of Florida; Rongling Wu, University of Florida

63. CONTRIBUTED PAPERS: HEALTH SERVICE AND EPIDEMIOLOGIC METHODS

[Brigade](#)

Sponsor: ENAR

Chair: Mari Palta, University of Wisconsin

- 1:45 Assessing Cost-Effectiveness of Treatments Using Propensity Scores in the Presence of Missing Data
Alka Indurkha*, Harvard University, School of Public Health; Nandita Mitra, Memorial Sloan-Kettering Cancer Center
- 2:00 Inferences on the Means for Lognormal Data Containing Zeros Based on Generalized Variable Method
Lili Tian*, University of Florida
- 2:15 Applying Markov Mixed-Effects Regression Models in Analysis of Risk Factors for Transitions of ICU Patients' Cognitive Status
Ayumi K. Shintani*, Vanderbilt University; E. Wesley Ely, Vanderbilt University; Theodore Speroff, Vanderbilt University

SCIENTIFIC PROGRAM

- 2:30 A New Method of Predicting U.S. and State-Level Cancer Mortality Counts for the Current Calendar Year
Ram C Tiwari*, National Cancer Institute; Kaushik Ghosh, George Washington University and National Cancer Institute; Ahmedin Jemal, American Cancer Society; Mark Hachey, Information Management Services, Inc.; Elizabeth Ward, American Cancer Society; Michael J. Thun, American Cancer Society; Eric J. Feuer, National Cancer Institute
- 2:45 A Bias-Corrected Weighted Average Method for Measurements from Heterogeneous Instruments with Systematic Biases
Sharon X. Xie*, University of Pennsylvania; Vernon M. Chinchilli, Pennsylvania State University
- 3:00 Size Does Matter: A Case for the .04 Significance Level
Adrian E. Levin*, Apex College, Nepal; Michael Atwatter, Apex College, Nepal
- 3:15 Floor Discussion

64. CONTRIBUTED PAPERS: APPLICATIONS OF RANDOM EFFECTS MODELS

[Rivers](#)

Sponsor: ENAR

Chair: Melanie Wall, University of Minnesota

- 1:45 An Efficient MCEM Algorithm for Fitting Generalized Linear Mixed Models for Correlated Binary Data
Ming Tan, University of Maryland Greenebaum Cancer Center; Guoliang Tian*, University of Maryland Greenebaum Cancer Center; Hongbin Fang, University of Maryland Greenebaum Cancer Center
- 2:00 Challenges in the Analysis of Observational Longitudinal Aging Studies
Geert Verbeke*, Katholieke Universiteit Leuven, Belgium; Steffen Fieuws, Katholieke Universiteit Leuven, Belgium; Bart Spiessens, Katholieke Universiteit Leuven, Belgium; Emmanuel Lesaffre, Katholieke Universiteit Leuven, Belgium
- 2:15 **A Mixed-Effects Regression Model for Longitudinal Multivariate Ordinal Data**
Li C Liu*, University of Illinois at Chicago; Donald Hedeker, University of Illinois at Chicago
- 2:30 A Multivariate Random Effects Model with Closed-Form Likelihood for Clustered Binary Data
Brent A. Coull*, Harvard School of Public Health; Rebecca A. Betensky, Harvard School of Public Health

- 2:45 GLMM's with Correlated Random Effects to Model Discrete Time Series
Bernhard Klingenberg*, University of Florida; Alan Agresti, University of Florida; James Booth, University of Florida
- 3:00 An Adaptive Method for Checking Convergence of the Monte Carlo Em Algorithm
Jens C. Eickhoff*, University of Wisconsin; Jun Zhu, University of Wisconsin; Yasuo Amemiya, IBM Thomas J. Watson Research Center
- 3:15 Modeling and Estimation in Two-Stage Models with Applications
John P. Buonaccorsi*, University of Massachusetts at Amherst

65. CONTRIBUTED PAPERS: BIOASSAYS AND CELL KINETICS

[Heinz](#)

Sponsor: ENAR

Chair: Kevin Gross, North Carolina State University

- 1:45 Sample Size Calculation and Concordance Assessment for an Agreement Study
Jason Liao*, Merck Research Laboratories
- 2:00 Statistical Issues in Calibration of Serum Creatinine Testing in the Chronic Renal Insufficiency Cohort Study
Christina Gaughan*, University of Pennsylvania; Marshall M. Joffe, University of Pennsylvania; J. Richard Landis, University of Pennsylvania; Harold I. Feldman, University of Pennsylvania
- 2:15 A Threshold Dose-Response Model with Random Effects in Teratological Experiments
Daniel Hunt, St. Jude Children's Research Hospital; Shesh N. Rai*, St. Jude Children's Research Hospital
- 2:30 Some State Space Models of Carcinogenesis and Estimation of Parameters
Wai-Yuan Tan, University of Memphis; Lijun Zhang*, University of Memphis
- 2:45 Summarizing Flare Assay Images in Colon Carcinogenesis
Malgorzata Leyk Williams*, Texas A&M University; Raymond J. Carroll, Texas A&M University
- 3:00 Correlation Coefficient Inference on Censored Bioassay Data
Liang Li*, Cleveland Clinic Foundation; William Wang, Merck Research Laboratories; Ivan Chan, Merck Research Laboratories
- 3:15 A Stochastic Model to Analyze Clonal Data on Multi-Type Cell Populations
Ollivier Hyrien*, University of Rochester Medical Center

SCIENTIFIC PROGRAM

Tuesday, March 30

3:30–3:45 p.m.

BREAK

[Exhibit Hall-Ballroom 2](#)

Tuesday, March 30

3:45–5:30 p.m.

66. RECENT ADVANCES IN TUMOR GROWTH MODELING

[Sterlings 2 & 3](#)

Sponsor: ENAR

Organizer: Paul Albert, National Institutes of Health

Chair: Shou-En Lu, UMDNJ-SPH

- 3:45 Longitudinal Models with Constrained Parameters for Tumor Xenograft Experiments
Ming T. Tan*, University of Maryland; Hongbin Fang, University of Maryland; Guo Tian, University of Maryland
- 4:15 Modeling Tumor Growth with Random Onset
Paul. S Albert*, National Cancer Institute; Joanna H. Shih, National Cancer Institute
- 4:45 Quantal Response Models of Cancer Detection in the Presence of Screening
Andrei Y. Yakovlev*, University of Rochester
- 5:20 Discussant: Daniel Heitjan, University of Pennsylvania School of Medicine

67. FUNCTIONAL AND LONGITUDINAL DATA ANALYSIS

[Ballroom 3](#)

Sponsors: ENAR/IMS

Organizer: Peter Mueller, University of Texas, M.D. Anderson Cancer Center

Chair: Peter Mueller, University of Texas, M.D. Anderson Cancer Center

- 3:45 Nonparametric and Semiparametric Regression for Clustered/Longitudinal Data Using Kernel and Spline Methods
Xihong Lin*, University of Michigan; Naisyin Wang, Texas A&M University; Alan Welsh, Southampton University; Raymond J. Carroll, Texas A&M University
- 4:15 Nonparametric Mixed-Effect Models
Chong Gu*, Purdue University
- 4:45 Bayesian Inference for Wavelet-Based Modeling Of Functional Data
Marina Vannucci*, Texas A&M University
- 5:20 Discussant: Gary Rosner, M.D. Anderson Cancer Center

68. BIOINFORMATICS IN GENOMICS STUDIES

[Kings Garden North](#)

Sponsor: ASA Biopharmaceutical Section

Organizer: Sue Jane Wang, Food and Drug Administration

Chair: Sue Jane Wang, Food and Drug Administration

- 3:45 Stochastic Models for Short Oligonucleotide Microarrays Based on the Theory of Chemical Physics
Rafael A. Irizarry*
- 4:05 Clinical Genomics - Principles And Applications
Chuanbo Xu*, Genaissance Pharmaceuticals, Inc.; Krishnan Nandabalan, Genaissance Pharmaceuticals, Inc.; Madan Kumar, Genaissance Pharmaceuticals, Inc.; Ruhong Jiang, Genaissance Pharmaceuticals, Inc.; Richard Judson, Genaissance Pharmaceuticals, Inc.; Brad Dain, Genaissance Pharmaceuticals, Inc.; Carol Reed, Genaissance Pharmaceuticals, Inc
- 4:30 An Innovative Approach That Synthesizes Expression Profiles, Genotypes And Phenotypes
Stephanie A. Monks*, University of Washington, Seattle; Rosetta Inpharmatics LLC; Haoyuan Zhu, Rosetta Inpharmatics LLC; Amy Leonardson, Rosetta Inpharmatics LLC; Eric Schadt, Rosetta Inpharmatics LLC
- 4:50 Genomics Studies in AIDS - Statistical Model Integrating Longitudinal Cohort Studies and Transversal Observations
Catherine L. Huber-Carol*, Université de Paris
- 5:10 Pooling Information Across Different Studies and Affymetrix Chip Types to Identify Prognostic Genes For Lung Cancer
Jeffrey S. Morris*, M.D. Anderson Cancer Center; Li Zhang, M.D. Anderson Cancer Center; Guosheng Yin, M.D. Anderson Cancer Center; Keith Baggerly, M.D. Anderson Cancer Center; Chunlei Wu, M.D. Anderson Cancer Center

69. MODELING AND SURVEILLANCE OF SEQUENCES OF HEALTH EVENTS

[Kings Garden South](#)

Sponsor: ENAR

Organizer: Andrew Lawson, University of South Carolina

Chair: Lance Waller, Emory University

- 3:45 Optimal Surveillance of Health Events
Marianne E. Frisén*, Statistical Research Unit Göteborg University

SCIENTIFIC PROGRAM

- 4:05 Simulating Anthrax Attacks and Evaluating Detection Techniques
Ken Kleinman*, Harvard Medical School; Allyson Abrams, Harvard Medical School; Richard Platt, Harvard Medical School; Martin Kulldorff, Harvard Medical School
- 4:30 Surveillance of Spatial Maxima
Peter A. Rogerson*, University at Buffalo
- 4:50 Modeling Spatially Correlated Survival Data with Multiple Cancers
Sudipto Banerjee*, University of Minnesota
- 5:10 The Analysis of Spatially-Referenced Recurrent Event Health Data with Comorbidities
Andrew B. Lawson*, University of South Carolina; Shae Sutton, University of South Carolina; Charity Moore, University of South Carolina

70. ANALYSIS OF INTENSIVELY COLLECTED DATA

Sponsor: IMS

[Benedum](#)

Organizers: Runze Li, Penn State University and Joe Schaffer, Penn State University

Chair: Lei Nie, University of Maryland at Baltimore County

- 3:45 Scientific Issues Driving Model Development for Intensive Longitudinal Data
Theodore A. Walls*, Penn State University
- 4:10 Backfitting for Random Varying-Coefficient Models
Hua Liang*, St. Jude Children's Research Hospital; Wu Hulin, University of Rochester
- 4:35 Semi-Nonparametric Models and Inference Functions for High Dimensional Data
Annie Qu*, Oregon State University
- 5:00 Models And Methods For Analysis Of Intensively Collected Data
Runze Li*, Penn State University; Tammy L. Root, Penn State University
- 5:25 Floor Discussion

71. CONTRIBUTED PAPERS: METHODS FOR MISSING OR UNBALANCED DATA

Sponsor: ENAR

[Duquesne](#)

Chair: Jason Roy, University of Rochester

- 3:45 Effects of Truncation on Evaluation of Risk Factors, with Case-Study in Anthrax
Sarah J.E. Barry*, Johns Hopkins School of Public Health; Ron Brookmeyer, Johns Hopkins School of Public Health

- 4:00 An Empirical Comparison of Imputation Methods for Use with Public Health Data
Stephanie T. Broyles*, Tulane University
- 4:15 **A Simple Sensitivity Analysis Tool for Nonignorable Coarsening: Application to Competing Risk**
Jiameng Zhang*, University of Pennsylvania; Daniel F. Heitjan, University of Pennsylvania
- 4:30 Comparing the Analyses Of Unbalanced Split-Plot Experiments Using Mixed Models with REML Estimates of the Variance Components to Other Inference Procedures Previously Recommended
Christina Smith*, Kansas State University; Dallas Johnson, Kansas State University
- 4:45 Analysis of Longitudinal Data with Dropouts and Intermittently Missing Values
Gong Tang*, University of Pittsburgh
- 5:00 **Analysis of Incomplete HRQoL Data in the Rematch Trial**
Huilin Li*, Columbia University; Daniel F. Heitjan, University of Pennsylvania
- 5:15 Large Sample Bias in the Last Observation Carried Forward Approach Under Informative Dropout
Chandan Saha*, Indiana University; Michael P. Jones, University of Iowa

72. CONTRIBUTED PAPERS: GENETICS AND DRUG SCREENING

[Heinz](#)

Sponsor: ENAR

Chair: Christina Kendzioriski, University of Wisconsin

- 3:45 Analysis Of Allelic Loss Data Using Marginal Models
Yan Li*, Ohio State University; Lei Shen, Ohio State University
- 4:00 Inference on the Location of a Tumor Suppressor Gene by Modeling Frequency of Allelic Loss
Andrew Sterrett*, University of North Carolina at Chapel Hill; Fred A. Wright, University of North Carolina at Chapel Hill
- 4:15 Location-Spread Displays of Many Point Estimates and Their Associated Measures of Dispersion
Mario Peruggia*, Ohio State University; Jason C. Hsu, Ohio State University
- 4:30 A Graphical Approach for Quality Control of Oligonucleotide Array Data
Dung-Tsa Chen*, University of Alabama at Birmingham

SCIENTIFIC PROGRAM

- 4:45 Statistical Design and Analysis of Pools Using Optimal Coverage and Minimal Collision
Katja S. Remlinger*, North Carolina State University; Jacqueline M. Hughes-Oliver, North Carolina State University; S. Stanley Young, National Institute of Statistical Sciences; Raymond L. Lam, GlaxoSmithKline
- 5:00 Probabilistic High-Dimensional Discrimination for Drug Screening
Alexander G Gray*, Carnegie Mellon University
- 5:15 Floor Discussion

73. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS: METHODS FOR MULTIPLE ENDPOINTS

Sponsor: ENAR

Chair: Jianwen Cai, University of North Carolina at Chapel Hill

[*Kings Terrace*](#)

- 3:45 **Shared Frailty Models for Recurrent Events and a Terminal Event**
Lei Liu*, University of Michigan; Robert A. Wolfe, University of Michigan; Xuelin Huang, M. D. Anderson Cancer Center
- 4:00 Modeling of Recurrent Event Survival Data Based on Extensions of the Andersen-Gill Model
Zekarias Berhane*, University of Pittsburgh; Lisa A. Weissfeld, University of Pittsburgh
- 4:15 Controlling Bias in Internal Pilots with the Univariate Approach to Repeated Measures
Christopher S. Coffey*, University of Alabama at Birmingham; Keith E. Muller, University of North Carolina at Chapel Hill; David Sawrie, University of Alabama at Birmingham
- 4:30 Models and Analysis of Recurrent Event Data
Edsel A. Pena*, University of South Carolina; Elizabeth H. Slate, Medical University of South Carolina; Juan R. Gonzalez, Catalan Institute of Oncology, Spain
- 4:45 Regression Analysis with Multivariate Grouped and Continuous Survival Data
Andrea B. Troxel*, University of Pennsylvania School of Medicine; Denise A. Esserman, Columbia University
- 5:00 Floor Discussion

74. CONTRIBUTED PAPERS: CATEGORICAL DATA

Sponsor: ENAR

Chair: John Williamson, Center for Disease Control and Prevention

[*Brigade*](#)

- 3:45 Marginal Models for Binary Data: Existence, Uniqueness and Implications
Bahjat F. Qaqish*, University of North Carolina at Chapel Hill; Anastasia Ivanova, University of North Carolina at Chapel Hill
- 4:00 Optimal Nonrandomized Tests for Comparing Two Binomial Probabilities
Weizhen Wang*, Wright State University
- 4:15 The Effect of Observational Unit Size on Poisson Goodness-of-Fit Tests
John W. Ely, University of Iowa; Jeffrey D. Dawson*, University of Iowa
- 4:30 Exact Bootstrap Confidence Intervals for the Difference of Two Probabilities Using Data from Two Independent Binomial Samples with Small Sample Sizes
Yan Lin*, Medical University of South Carolina; Stuart R. Lipsitz, Medical University of South Carolina
- 4:45 Estimation Of Parameters For Binary Outcome Data in the Presence of Missing Covariate Information
Gina M. D'Angelo*, University of Pittsburgh; Lisa Weissfeld, University of Pittsburgh
- 5:00 A Bias Analysis of the Logistic Normal Likelihood Ratio Test
Rafael E. Diaz*, Tulane University
- 5:15 Modeling the On-Off Process to Analyze the Postural Instability of PD Patients
Debajyoti Sinha*, Medical University of South Carolina; Ming-Hui Chen, University of Connecticut; Peng Huang, Medical University of South Carolina

75. CONTRIBUTED PAPERS: NONPARAMETRIC METHODS FOR LONGITUDINAL DATA AND BAYESIAN ANALYSIS

Sponsors: ENAR/IMS

Chair: Sujit Ghosh, North Carolina State University

[*Le Bateau*](#)

- 3:45 Testing Similarity Between Two Groups of Curves
Yolanda Munoz Maldonado*, Texas A&M University
- 4:00 Bootstrap Confidence Bands
Laura L. Johnson*, National Cancer Institute; Paula Diehr, University of Washington

SCIENTIFIC PROGRAM

- 4:15 Dirichlet Process Mixture Models for Markov Processes
Yongqiang Tang*, North Carolina State University;
Subhashis Ghosal, North Carolina State University
- 4:30 **Functional Mixed Effects Models with Prior Information**
Li Qin*, University of Pennsylvania; Wensheng Guo,
University of Pennsylvania
- 4:45 Reference Prior in Bayesian Penalized Spline Analysis
Xin Zhao*, Cornell University; Martin T. Wells,
Cornell University
- 5:00 Floor Discussion

76. CONTRIBUTED PAPERS: SURROGATE ENDPOINTS, SCREENING AND ASSESSING AGREEMENT

[Rivers](#)

Sponsor: ENAR

Chair: Scarlett Bellamy, University of Pennsylvania

- 3:45 A Critical Review of Statistical Methods for Assessing Agreement Between Continuous Measurements
Joseph L. Hagan*, University of Louisville; Stephen W. Looney, Louisiana State University
- 4:00 Statistical Methods for the Analysis of Distortion Product Otoacoustic Emissions
Peter F. Craigmile*, The Ohio State University; Wayne M. King, The Ohio State University
- 4:15 Statistical Models for Vaccine Correlates of Protection
Andrew J. Dunning*, Wyeth Vaccines Research
- 4:30 Left, Right and Interval Bivariate Censored Data: Evaluating Screening Mammography in the Presence of Lead-Time Bias, Length Bias and Over-Detection
Jonathan D. Mahnken*, University of Texas; Wenyaw Chan, University of Texas; Jean L. Freeman, University of Texas
- 4:45 The Analysis of Clustered Matched-Pair Data
Valerie L. Durkalski*, The Clinical Innovation Group; Yuko Y. Palesch, The Clinical Innovation Group and The Medical University of South Carolina; Stuart Lipsitz, The Medical University of South Carolina; Phil Rust, The Medical University of South Carolina
- 5:00 **Isotonic Logistic Discrimination**
Sungyoung Auh*, University of Pittsburgh; Allan R. Sampson, University of Pittsburgh
- 5:15 Modeling the Agreement of Discrete Bivariate Survival Times Using Kappa Coefficient
Ying Guo*; Emory University; Amita K. Manatunga, Emory University

5:30-6:30 ENAR Business Meeting (open to all members)

[Brigade](#)

Wednesday, March 31

8:30–10:15 a.m.

77. CURRENT DEVELOPMENTS AND STATISTICAL CHALLENGES IN DEVELOPING SYNDROMIC SURVEILLANCE SYSTEMS FOR ANTI-BIOTERRORISM

[Sterlings 2 & 3](#)

Sponsor: ENAR

Organizers: Yiliang Zhu, University of South Florida and Wei Wang, University of South Florida

Chair: Yiliang Zhu, University of South Florida

- 8:30 Current Developments and Statistical Challenges in Developing Syndromic Surveillance Systems for Anti-Bioterrorism
Lori Hutwagner*, Centers for Disease Control
- 9:00 Syndrome Surveillance System for Anti-Bioterrorism: Statistical Issues in Trend Analysis and Outbreak Detection
Wei Wang*, University of South Florida; Yiliang Zhu, University of South Florida
- 9:30 Missing Data in Daily Syndromic Surveillance Systems
Martin Kulldorff*, Harvard Medical School; Farzad Mostashari, New York City Department of Health and Mental Hygiene; Rick Heffernan, New York City Department of Health and Mental Hygiene; Jessica Hartman, New York Academy of Medicine
- 10:00 Floor Discussion

78. NOVEL STATISTICAL APPROACHES TO HIGH DIMENSIONAL ANALYSES

[Ballroom 3](#)

Sponsor: ENAR

Organizer: David B. Allison, University of Alabama at Birmingham

Chair: David B. Allison, University of Alabama at Birmingham

- 8:30 Analysis of Genome-Wide Epistasis via Dimensionality Reduction Procedures
Marylyn D. Ritchie*, Vanderbilt University
- 8:55 Analysis of Genome-Wide Epistasis Via Bayesian Model Selection
Nengjun Yi*, The University of Alabama at Birmingham
- 9:20 Use and Estimation of False Discovery Rate in Genetics
Chiara Sabatti*, UCLA

SCIENTIFIC PROGRAM

9:45 Power & Sample Size Estimation When Many Simultaneous Hypothesis Tests are Planned
Grier P. Page*, University of Alabama at Birmingham;
Gary Gadbury, University of Missouri – Rolla; Jode W. Edwards, University of Alabama at Birmingham
David B. Allison, University of Alabama at Birmingham

10:10 Floor Discussion

79. NON-INFERIORITY/EQUIVALENCE TRIALS

[Ballroom 4](#)

Sponsor: ASA Biopharmaceutical Section

Organizers: Paul Stober, GlaxoSmithKline and Darryl Downing, GlaxoSmithKline

Chairs: Paul Stober, GlaxoSmithKline and Darryl Downing, GlaxoSmithKline

8:30 Delta What? Choice of Outcome Scale in Non-Inferiority Trials
Rick Chappell*, University of Wisconsin - Madison

9:00 Choice of Non-Inferiority Margin
H.M. James Hung*, Food and Drug Administration;
Sue-Jane Wang, Food and Drug Administration

9:30 Non-Inferiority and Equivalence Margins in Clinical Pharmacology
Scott D. Patterson*, GlaxoSmithKline Pharmaceuticals

10:00 Discussant: Liang Yuh, AstraZeneca

80. CONTRIBUTED PAPERS: STATISTICAL ISSUES IN CLINICAL TRIALS

[Kings Garden North](#)

Sponsor: ENAR

Chair: Cyrus Mehta, Cytel

8:30 Continuous Toxicity Monitoring in Phase II Trials in Oncology
Anastasia Ivanova*, University of North Carolina at Chapel Hill; Bahjat F. Qaqish, University of North Carolina at Chapel Hill; Michael Schell, University of North Carolina at Chapel Hill

8:45 Effects of AIDS Vaccine on Sub-Populations of CD4 T Cells, CD8 T Cells and B Cells Under HIV Infection
Wai-Yuan Tan, The University of Memphis; Ping Zhang*, The University of Memphis; Xiaoping Xiong, St. Jude Children's Research Hospital

9:00 A Critical History of Individual and Collective Ethics
Charles M. Heilig*, Centers for Disease Control and Prevention; Charles Weijer, Dalhousie University

9:15 Evaluation of Power Change Based on Correctness of Model, Sample Size, and Detectable Dosage in First Human Dose Studies

Adeline Yeo*, Eli Lilly and Company

9:30 Estimating Correct-Use Condom Failure Probabilities

Mark A. Weaver*, Family Health International; Douglas J. Taylor, Family Health International; Rosalie C. Dominik, Family Health International

9:45 Floor Discussion

81. CONTRIBUTED PAPERS: GENE EXPRESSION ANALYSIS

[Brigade](#)

Sponsor: ENAR

Chair: Jung-Ying Tzeng, North Carolina State University

8:30 Statistical Methods for Analyzing Tissue Microarray Data
Xueli Liu*, UCLA; Vladimir Minin, UCLA; Yunda Huang, UCLA; David B. Seligson, UCLA; Steve Horvath, UCLA

8:45 Associations Between Gene Expressions and Survival Times in the Presence of Covariates
Taesung Park*, Seoul National University, Korea; Sin-Ho Jung, Duke University; Seung Yeoun Lee, Sejong University, Korea

9:00 Inference in High Dimension Low Sample Size Settings with Censored Outcomes
Brent A. Johnson*, University of North Carolina

9:15 **Modeling the Relationship Between LVAD Support Time and Gene Expression Changes in the Human Heart by Penalized Partial Least Squares**
Xiaohong Huang*, University of Minnesota; Wei Pan, University of Minnesota

9:30 Hidden Markov Modeling of Genomic Interactions
Ernst C. Wit, University of Glasgow; Nial Friel, University of Glasgow; John D. McClure*, University of Glasgow

9:45 Identification of Genomic Aberrations From Microarray Data
Chiang-Ching Huang*, Northwestern University; Jeremy Taylor, University of Michigan

SCIENTIFIC PROGRAM

10:00 Semiparametric Regression for Microarray Gene Expression Data: Support Vector Machines and Linear Mixed Models
Dawei Liu*, University of Michigan; Xihong Lin, University of Michigan; Debashis Ghosh, University of Michigan

82. CONTRIBUTED PAPERS: NON- AND SEMI-PARAMETRIC MODELS

[Duquesne](#)

Sponsor: ENAR

Chair: Calvin L. Williams, National Science Foundation and Clemson University

8:30 A Goodness-Of-Fit Test for Nonlinear Models Via Local Linear Smoothers
Chin-Shang Li*, St. Jude Children's Research Hospital

8:45 Conditionally Generalized Linear Models
Joan G. Staniswalis*, University of Texas at El Paso

9:00 A Unified Approach to Testing and Measuring Association Through a Semiparametric Odds Ratio Model
Hua Yun Chen*, University of Illinois at Chicago

9:15 Semiparametric Inferences in Stability Studies
Young Kyoung Min*, University of Florida; Annpey Pong, Forest Laboratories, Inc.; Ying Qing Chen, University of Florida and University of California at Berkeley

9:30 A Semiparametric Bayesian Model For Time Series Data
Kaushik Ghosh*, National Cancer Institute; Ram C. Tiwari, National Cancer Institute

10:00 Floor Discussion

83. CONTRIBUTED PAPERS: SPATIAL MODELING II

[Benedum](#)

Sponsor: ENAR

Chair: Justine Shults, University of Pennsylvania

8:30 Spatial Analysis Of Alcohol Availability And Violent Crime
Dennis Gorman, Texas A&M University; Li Zhu*, Texas A&M University; Scott Horel, Texas A&M University

8:45 Second-Order Spatial Analysis of Epidermal Nerve Fibers
Lance A. Waller*, Emory University; Leong L. Traci, Emory University William R. Kennedy, University of Minnesota; Gwen Wendelschafer-Crabb, University of Minnesota; Aila Sarkka, Chalmers University of Technology

9:00 Bayesian Spatial Prediction Using MCMC
Arin Chaudhuri*, North Carolina State University; Montserrat Fuentes, North Carolina State University; David Holland, U.S. Environmental Protection Agency

9:15 On the Statistical Modeling and Analysis of Adverse Health Outcomes Due to Chronic Arsenic Exposure
Munni Begum*, University of North Carolina at Chapel Hill; Pranab K. Sen, University of North Carolina at Chapel Hill Chapel Hill

9:30 **Multivariate Parametric Spatiotemporal Model For County Level Breast Cancer Survival Data**
Xiaoping Jin*, University of Minnesota; Bradley P. Carlin, University of Minnesota

9:45 Floor Discussion

84. CONTRIBUTED PAPERS: RANDOM EFFECTS MODELS

Sponsor: ENAR

[Kings Garden South](#)

Chair: Oliver Schabenberger, SAS

8:30 A Residual Information Criterion for the General Linear Mixed Model
Lloyd J Edwards*, University of North Carolina at Chapel Hill; Pranab K. Sen, University of North Carolina at Chapel Hill; Bahjat F. Qaqish, University of North Carolina at Chapel Hill; Matthew J. Gurka, University of North Carolina at Chapel Hill

8:45 Inference Techniques for Data Transformations in Linear Mixed Effects Models for Longitudinal Data
Matthew J. Gurka*, University of North Carolina at Chapel Hill; Lloyd J. Edwards, University of North Carolina at Chapel Hill

9:00 Synthesizing Conditional and Marginal Inferences in Covariate-Dependent Nonlinear Mixed-Effects Models
Zengri Wang*, 3M Company; Thomas A. Louis, Johns Hopkins University

9:15 Robust R-Estimation for Unbalanced Heteroscedastic One-Way Random Effects Models
Inkyung Jung*, University of North Carolina at Chapel Hill; Pranab K. Sen, University of North Carolina at Chapel Hill

9:30 When BLUP's Are Bad
Ed Stanek*, University of Massachusetts – Amherst; Julio M. Singer, University of Sao Paulo, Brazil; George Reed, University of Massachusetts – Worcester; Wenjun Li, Tufts-NEMC

SCIENTIFIC PROGRAM

- 9:45 Design and data analysis options for comparisons of assisted reproductive technologies
James A. Hanley*, McGill University; Nandini Dendukuri, McGill University; Robert Platt, McGill University; Marie-Helene Mayrand, McGill University
- 10:00 Measuring the Complexity of Generalized Linear Hierarchical Models
Haolan Lu, University of Minnesota; James Hodges, University of Minnesota; Brad Carlin, University of Minnesota

85. CONTRIBUTED PAPERS: MARKOV CHAIN MODELS

[Rivers](#)

Sponsor: ENAR

Chair: Zhen Chen, University of Pennsylvania

- 8:30 **Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions**
Ming Yuan*, University of Wisconsin at Madison; Christina M. Kendzierski, University of Wisconsin at Madison
- 8:45 Estimating Sojourn Time Distributions with Periodic Observations Under a Semi-Markov Model
Pai-Lien Chen*, Family Health International; Hsiao-Chuan Tien, University of North Carolina at Chapel Hill
- 9:00 Modeling the Natural Progression of a Chronic Condition From Observational Data with Time-Dependent Covariates
Yali Liu*, Purdue University; Bruce A. Craig, Purdue University
- 9:15 **Bayesian Searching for Markov Models of HIV Mutation Process**
Jing Zhao*, University of Pennsylvania; Edward I. George, University of Pennsylvania Wharton School; Andrea S. Foulkes, University of Pennsylvania
- 9:30 Fitting a Mixed Effects Markov Model for Repeated Binary Outcomes with Nonignorable Dropout
Robert J. Gallop*, West Chester University; Thomas R. Ten Have, University of Pennsylvania
- 9:45 Characterizing The Progression of Genomic Mutations in HIV: A Markov Process Approach
David Loeckel*, Harvard School Of Public Health; Victor DeGruttola, Harvard School of Public Health
- 10:00 Floor Discussion

Wednesday, March 31

10:15–10:30 a.m.

BREAK

[Exhibit Hall-Ballroom 2](#)

Wednesday, March 31

10:30 a.m.–12:15 p.m.

86. ASSESSING AGREEMENT MEASUREMENTS: CURRENT ISSUES AND NEW DEVELOPMENTS

[Ballroom 3](#)

Sponsors: ENAR/ASA Biopharmaceutical Section

Organizer: Lawrence Lin, Baxter Healthcare Corporation

Chair: Vern Chinchilli, Penn State College of Medicine

- 10:30 A Repeated Measures Concordance Correlation Coefficient for Clustered Paired Data
Tonya S. King*, Penn State University; Vernon M Chinchilli, Penn State University
- 11:00 Assay Validation with Censored Data
Huiman X. Barnhart*, Duke University; Jingli Song, Eli Lilly and Company; Robert Lyles, Emory University
- 11:30 On The Robustness of the Concordance Correlation Coefficient Estimated by Variance Components
Josep L. Carrasco*, Universitat de Barcelona
- 12:00 Discussant: Lawrence Lin, Baxter Healthcare Corporation

87. APPLICATIONS OF RESAMPLING METHODS TO ESTIMATING FUNCTIONS

[Ballroom 4](#)

Sponsors: ENAR/IMS

Organizer: Thomas Braun, University of Michigan

Chair: Thomas Braun, University of Michigan

- 10:30 Marginal Analysis of Clustered Data When Cluster Size is Informative
John M. Williamson*, Centers for Disease Control and Prevention; Glen A. Satten, Centers for Disease Control and Prevention; Somnath Datta, University of Georgia
- 11:00 Making Inference by Perturbing a Non-Smooth Estimating Function
Lu Tian*, Harvard University; L.J. Wei, Harvard University
- 11:30 Inference Based on Estimating Functions with U Statistic Structure
Jack Kalbfleisch*, University of Michigan; Wenyu Jiang, University of Waterloo
- 12:00 Floor Discussion

SCIENTIFIC PROGRAM

88. STATISTICAL GENETICS - MODELING INTERACTION AND MULTILOCUS ANALYSES

[*Sterlings 2 & 3*](#)

Sponsors: IMS

Organizer: Lei Sun, University of Toronto

Chair: Shelley Bull, Samuel Lunenfeld Research Institute, University of Toronto

- 10:30 Definition, Detection and Interpretation of Gene-Gene Interaction (Epistasis) in Complex Disease
Heather J. Cordell*, University of Cambridge
- 10:55 Testing for Locus-Locus Interactions Using Affected Relative Pairs: Multipoint Analysis and Incomplete IBD
Peter A. Holmans*, University of Wales
- 11:20 Joint Analyses of Linked Disease Genes: Location Estimation and Hypothesis Testing Methods
Joanna M. Biernacka*, University of Toronto; Shelley B. Bull, University of Toronto; Samuel Lunenfeld, Research Institute, Toronto; Lei Sun, University of Toronto
- 11:45 Multilocus Models for Linkage Analysis with Covariates
Jane M. Olson, Case Western Reserve University
- 12:10 Floor Discussion

89. CONTRIBUTED PAPERS: BAYESIAN COMPUTING AND ANALYSIS

[*Kings Garden North*](#)

Sponsor: ENAR

Chair: Valerie Ventura, Carnegie-Mellon University

- 10:30 **Loss Function Based Ranking in Two-Stages Hierarchical Models**
Rongheng Lin*, Johns Hopkins University; Thomas A. Louis, Johns Hopkins University, Susan Paddock, Rand Statistics Group; Greg Ridgeway, Rand Statistics Group
- 10:45 Bayesian Factor Analysis Models for Mental Disorders: A Case Study of the Whiteley Index For Hypochondriasis
David M. Shera*, The Children's Hospital of Philadelphia / University of Pennsylvania; Joseph G. Ibrahim, University of North Carolina
- 11:00 Benefits Of Benchmark Estimation
Steven N. MacEachern*, The Ohio State University; Subharup Guha, The Ohio State University
- 11:15 Bayesian Machine Learning for Large P Small N Regression
Sounak Chakraborty*, University of Florida; Malay Ghosh, University of Florida; Bani K. Mallick, Texas A&M University

- 11:30 Evaluation of Bayesian Model Diagnostic Techniques
Guofen Yan*, Cleveland Clinic Foundation; J. Sedransk, Case Western Reserve University and Energy Information Administration
- 11:45 Combining Phylogenetic Information and Environmental Factors in Hierarchical Modeling to Explain Species Diversity
Shanshan Wu*, University of Connecticut; Paul O. Lewis, University of Connecticut; Mark Holder, University of Connecticut; John A. Silander, University of Connecticut; Andrew Latimer, University of Connecticut; Anthony G. Rebelo, National Botanic Institute, Kirstenbosch, South Africa; Alan E. Gelfand, Duke University
- 12:00 Floor Discussion

90. CONTRIBUTED PAPERS: NONPARAMETRIC METHODS

[*Rivers*](#)

Sponsor: ENAR

Chair: Joan Staniswalis, University of Texas El Paso

- 10:30 Nonparametric Confidence Sets for Densities and Clusters
Woncheol Jang*, Carnegie Mellon University; Christopher Genovese, Carnegie Mellon University; Larry Wasserman, Carnegie Mellon University
- 10:45 On Nonparametric Regression for Current Status Data
Deborah Burr*, Ohio State University; Shanti Gomatam, Center for Devices and Radiological Health, U.S. Food and Drug Administration
- 11:00 Rank-Sum Tests for Clustered Data
Somnath Datta*, University of Georgia; Glen A. Satten, Centers for Disease Control and Prevention
- 11:15 Nonparametric Estimation of an Intensity Function and Its Breakpoint with Applications to Spike Train Data in Neuroscience
Matt Gregas*, University of Minnesota
- 11:30 Local Isotonic Regression
Derick R. Peterson*, University of Rochester
- 11:45 **A Penalized Nonparametric Maximum Likelihood Approach to Species Richness Estimation**
Ji-Ping Z. Wang*, Northwestern University; Bruce G. Lindsay, Penn State University
- 12:00 Nonparametric Variable Selection via Component Shrinkage
Yi Lin, University of Wisconsin, Madison; Hao H. Zhang*, North Carolina State University

SCIENTIFIC PROGRAM

91. CONTRIBUTED PAPERS: MEASUREMENT ERROR METHODS

[Brigade](#)

Sponsor: ENAR

Chair: Abdus Wahed, University of Pittsburgh

- 10:30 A Weighted Estimated Likelihood Method for Two-Stage Outcome-Dependent Study Using Kernel Smoother
Xiaofei Wang*, Duke University; Haibo Zhou, University of North Carolina at Chapel Hill
- 10:45 A Logistic Regression Model with Misclassification in the Outcome and Categorical Covariate
Michelle A. Detry*, The University of Texas; Wenyaw Chan, The University of Texas
- 11:00 Impact of Measurement Accuracy on Sample Size Estimation Using Pilot Pain Assessment: A Simulation Study
Yen-Hong Kuo*, Jersey Shore University Medical Center; Yen-Liang Kuo, Kinmen County Hospital
- 11:15 Measurement Error in Time Series Models
John W. Staudenmayer*, University of Massachusetts at Amherst; John P. Buonaccorsi, University of Massachusetts at Amherst
- 11:30 Closed Form Semiparametric Estimators for Measurement Error Models
Yanyuan Ma*, North Carolina State University; Anastasios A. Tsiatis, North Carolina State University
- 11:45 Covariate Measurement Error in Dual Systems Models
Gaitri Gunasekara*, Michigan Technologies University; Thomas D. Drummer, Michigan Technological University
- 12:00 A Model for a Latent Factor Predicting a Categorical Latent Outcome with Application to Eating Disorders Risk and Body Satisfaction
Jia Guo*, University of Minnesota; Melanie M. Wall, University of Minnesota

92. CONTRIBUTED PAPERS: LONGITUDINAL DATA II

[Kings Garden South](#)

Sponsor: ENAR

Chair: Gong Tang, University of Pittsburgh

- 10:30 A Joint Latent Autoregressive Model for Longitudinal Ordinal Data and Patient Dropout with Informative Missingness
George A. Capuano*, University of North Carolina University at Chapel Hill; Pranab K. Sen, North Carolina University at Chapel Hill

- 10:45 Choosing the Optimal Poct Device Via a Longitudinal Study
Jason T. Connor*, Carnegie Mellon University; Kenneth M. Shermock, The Johns Hopkins Hospital; Emily Stewart, Brigham Young University; Jodie Fink, Cleveland Clinic Foundation; Lee Bragg, Kaiser Permanente
- 11:00 Bayesian Analysis of Mixed-Effects Time Varying-Coefficient Models for Longitudinal Data
Yi-Liang Tung*, National Taiwan University, Taiwan, R.O.C
- 11:15 Multivariate Transition Models for Longitudinal Binary Data
Leilei Zeng*, University of Waterloo; Richard R.J. Cook, University of Waterloo
- 11:30 The Efficient Estimation of Nonlinear Effects in Both Cross-Sectional and Longitudinal SEM
Andreas G. Klein*, University of Illinois at Urbana-Champaign; Bengt O. Muthen, University of California Los Angeles
- 11:45 Sample Size Calculation for Studies with Correlated Ordinal Outcomes
Hae-Young Kim*, Centers for Disease Control and Prevention; John M. Williamson, Centers for Disease Control and Prevention; Cynthia M. Lyles, Centers for Disease Control and Prevention
- 12:00 Floor Discussion

93. CONTRIBUTED PAPERS: GENETIC MAPS, LINKAGE DISEQUILIBRIUM AND HAPLOTYPE ANALYSIS

[Duquesne](#)

Sponsor: ENAR

Chair: Eleanor Feingold, University of Pittsburgh

- 10:30 A Closer Look at Human Recombination
John Barnard*, Cleveland Clinic Foundation
- 10:45 On the Use of Prior Information to Develop Genetic Maps
Chris A. Andrews*, Oberlin College
- 11:00 A Uniform Measure of Linkage Disequilibrium for Markers with Two or More Alleles
Chun Li*, Center for Human Genetics Research, Vanderbilt University
- 11:15 Estimating the Haplotype Specific Disease Risk in a Case-Control Study
Nandita Mitra*, Memorial Sloan-Kettering Cancer Center; E.S. Venkatraman, Memorial Sloan-Kettering Cancer Center; Colin B Begg, Memorial Sloan-Kettering Cancer Center

SCIENTIFIC PROGRAM

- 11:30 Pruning The Reconstructed Haplotype Frequencies in Association Analysis
Jung-Ying Tzeng*, North Carolina State University
- 11:45 Integration of Association Statistics Over Genomic Regions Using Bayesian Adaptive Regression Splines
Xiaohua Zhang*, Merck Research Laboratories; Kathryn Roeder, Carnegie Mellon University; Garrick Wallstrom, Carnegie Mellon University; Bernie Devlin, University of Pittsburgh School of Medicine
- 12:00 Floor Discussion

- 12:00 Statistical Issues in the Analysis of Experimental Psychopharmacology Studies
Ralitza V. Gueorguieva*, Yale University; Yu-Te Wu, Yale University; John Krystal, Yale University

94. CONTRIBUTED PAPERS: METHODS FOR MULTIPLE ENDPOINTS [Benedum](#)

Sponsor: ENAR

Chair: Janelle S. Erickson, Eli Lilly and Company

- 10:30 **Conditional Estimation for Generalized Linear Models When Covariates are Subject-Specific Parameters in a Mixed Model for Longitudinal Measurements**
Ering Li*, North Carolina State University; Daowen Zhang, North Carolina State University; Marie Davidian, North Carolina State University
- 10:45 Joint Modeling of Adherence and Repeated Measures
Wei Lang*, Wake Forest University; Lisa A. Weissfeld, University of Pittsburgh; Ralph B. D'Agostino, Jr., Wake Forest University
- 11:00 K-Means Cluster Analysis of Longitudinal Trajectories with Smoothing
David R. Gagnon*, Boston University School of Public Health; Margaret Stedman, Boston University School of Public Health
- 11:15 Application of the Time-Varying Autoregressive Model with Covariates
Tulay Koru-Sengul*, University of Pittsburgh; David S. Stoffer, University of Pittsburgh; Patricia R. Houck, University of Pittsburgh; Jean M. Miewald, University of Pittsburgh; Daniel J. Buysse, University of Pittsburgh; David J. Kupfer, University of Pittsburgh
- 11:30 Point Estimate After Group Sequential Tests
Keyue Ding*, National Cancer Institute of Canada, Clinical Trials Group Queen's University
- 11:45 Investigation Of Two Kinds Of Decision Rules For Bivariate Sequential Tests Based On Marginal Criteria
Yanli Zhao*, University of Minnesota; Patricia M Grambsch, University of Minnesota; James D. Neaton, University of Minnesota

95. CONTRIBUTED PAPERS: BIostatistical Theory AND ESTIMATION [Le Bateau](#)

Sponsors: ENAR/IMS

Chair: Anastasios Tsiatis, North Carolina State University

- 10:30 Applications Of Convolution Parameter Estimation
Calvin L. Williams*, National Science Foundation
- 10:45 Bounds on the Biasing Parameter and the Moments of the Stochastic Shrinkage Parameters of the Liu Type Estimator
Fikri Akdeniz*, Cukurova University
- 11:00 Estimation of Attributable Fraction Using Data From Partial Questionnaire
E Andres Houseman*, Harvard School of Public Health; Donald K Milton, Harvard School of Public Health
- 11:15 Estimate the Mean Treatment Effect in the Presence of Measurement Errors
Zhe Shang*, Wyeth Pharmaceuticals
- 11:30 Properties of Wald and Score Statistics — The Two Sample Case for Generalized Linear Models, and Implications for More General Models
David I Warton*, University of New South Wales
- 11:45 What is the Degree of Deviation From Balance in Comparing Two Independent Groups Means? A Simulation Study
Handan Çamdeviren*, Mersin Universitesi; Mehmet Mendeş, Mersin Universitesi; Refik Burgut, Mersin Universitesi, Mehmet A. Sungur, Mersin Universitesi Seval Kul, Mersin Universitesi
- 12:00 Adjusting O'Brien's Test to Control Type I Error for the Generalized Nonparametric Behrens-Fisher Problem
Peng Huang*, Medical University of South Carolina; Barbara Tilley, Medical University of South Carolina; Stuart Lipsitz, Medical University of South Carolina; Robert Woolson, Medical University of South Carolina

POSTER SESSION SUMMARY

1. A Comparison of Estimators of Population Slope under Informative Drop-Out in Longitudinal Studies
Hai Lin and Elizabeth H. Slate, Medical University of South Carolina
2. A Bayesian Nonparametric Repeated Fractional Data Model
Ying Yang and Peter Mueller, The University of Texas M.D. Anderson Cancer Center
3. Modeling Two or More Categorical Variables that Allow for Multiple Category Choices
Christopher R. Bilder, University of Nebraska-Lincoln; Thomas M. Loughin, Kansas State University
4. Model Selection for Clustered Recurrent Event Data Using Nested Gamma Frailty Models
Xin Zhi, University of Minnesota School of Public Health; Lynn E. Eberly, University of Minnesota School of Public Health
5. Reliability and Validity of Daily Collection of Symptom Data in Midlife Women
Imke Janssen, Rush University; Peter M. Meyer, Rush University; Kelly Karavolos, Rush University; Lynda H. Powell, Rush University
6. Schematic: A Program for Classifying and Diagramming Mediators, Moderators, and Independent, Proxy and Overlapping Risk Factors
John Scott, University of Pittsburgh School of Medicine; Helena C. Kraemer, Stanford University; Marilyn J. Essex, University of Wisconsin – Madison; Ellen Frank, University of Pittsburgh School of Medicine; David J. Kupfer, University of Pittsburgh School of Medicine
7. Sample Size Determination in Studies where Health State Utility Assessments are Compared across Groups and Time
Barbara H. Hanusa, University of Pittsburgh; Christopher R.H. Hanusa, University of Washington; Chung-Chou H. Hanusa, University of Pittsburgh; Kevin L. Kraemer, University of Pittsburgh.
8. Spatial Cluster Detection using Bayes Factors from Overparameterized Models
Ronald E. Gangnon, University of Wisconsin - Madison; Murray K. Clayton, University of Wisconsin - Madison
9. Extending the Two-Part Model to Accommodate Multiple Response Opportunities
Leann Myers, Tulane University School of Public Health & Tropical Medicine; Yeonjoo Yi, Tulane University School of Public Health & Tropical Medicine
10. Effects of Misclassification Error in Hierarchical Poisson Models
Leslie H. Morgan, Tulane University; Leann Myers, Tulane University; Frances J. Mather, Tulane University
11. A Comparison of Goodness-of-fit Tests for Binomial Generalized Estimating Equations Models
Huiyi Lin, Tulane University; Leann Myers, Tulane University
12. QTL Analysis in Genotypically Selected Experimental Populations
Zongli Xu, University of North Carolina at Chapel Hill; Fei Zou, University of North Carolina at Chapel Hill; Todd J. Vision, University of North Carolina at Chapel Hill
13. A Cost Effectiveness Analysis of Mean Time in Moderate – to-Severe Alzheimer’s Disease
Xiaoning (Joanie) Zhu, Medical University of South Carolina; Kit N. Simpson, University of South Carolina
14. A Full Pedigree based Method for the Statistical Assessment of Genetic Anticipation
Tracy J. Costello, M.D. Anderson Cancer Center and The University of Texas - Houston; Christopher I. Amos, M.D. Anderson Cancer Center
15. Latent Trait Rater Agreement Models with Applications in Oral Health
Elizabeth G. Hill, Medical University of South Carolina; Elizabeth H. Slate, Medical University of South Carolina
16. Understanding Systematic Variation in Gel Electrophoresis
Jeffrey C. Miecznikowski, Carnegie Mellon University, Kim Sellers, Carnegie Mellon University; William F. Eddy, Carnegie Mellon University; Jon Minden, Carnegie Mellon University
17. Some New Multivariate Cusum Quality Monitoring Procedures
Richard A. Johnson, University of Wisconsin-Madison; Ruoqia Li, University of Wisconsin-Madison

POSTER SESSION SUMMARY

18. Factors Affecting the Performance of a Syndromic Surveillance Outbreak Detection Algorithm
Ling Wang, Boston University School of Public Health;
Paola Sebastiani, Boston University School of Public Health;
Kenneth D. Mandl, Harvard Medical School
19. Bayesian Methods for Individual Bioequivalence Studies
Janelle S. Erickson, Eli Lilly and Company; John W. Seaman, Jr., Baylor University
20. Receiver Operating Characteristic Curve Analysis in Follow-Up Studies With Censoring
James B. Kampert, The Cooper Institute
21. Missing Data in Palliative Care
J. Lynn Palmer, M.D. Anderson Cancer Center
22. Comparison for Stochastic Compartment Models of a Digestive System in Forage-Feed Animals
Yoonsung Jung, Kansas State University
23. Shrinkage Correction for Prediction – Future Random X Cases
Xue Xin, Tulane University School of Public Health and Tropical Medicine; Leann Myers, Tulane University School of Public Health and Tropical Medicine
24. EM Algorithm Based Genetic Tests of Association and/or Linkage with Missing Parental Data
Chao-Yu Guo, Boston University School of Public Health; Kathryn Lunetta, Genome Therapeutics, Corp.; Anita DeStefano, Boston University School of Public Health; Josée Dupuis, Boston University School of Public Health; L. Adrienne Cupples, Boston University School of Public Health
25. Systematic Assessment of Affymetrix Microarray Data Quality
Chaohui Wang, Cleveland Clinic Foundation; Xuejun Peng, Cleveland Clinic Foundation; Eric Blalock, University of Kentucky
26. The Extended Clinical Trial
Robert A. Parker, Center for Biostatistics in AIDS Research, Harvard School of Public Health
27. A Sensitivity Analysis of Longitudinal Dropout Data Using OSWALD
Amy E. Begley, University of Pittsburgh; Gong Tang, University of Pittsburgh; Patricia R. Houck, University of Pittsburgh; Sati Mazumdar, University of Pittsburgh; Benoit H. Mulsant, University of Pittsburgh; Charles F. Reynolds, University of Pittsburgh

NOTES

ENAR WEBSITE
WWW.ENAR.ORG



1. Advances in Latent Variable Methodology: Bridging the Gap Between Social Science and Biomedical

LATENT VARIABLES IN THE SOCIAL SCIENCES

Kenneth A. Bollen*, University of North Carolina, Chapel Hill

This paper is an overview of the meaning, properties, and uses of latent variables in the social sciences. More specifically, the purposes of the paper are: (1) to present definitions and uses of latent variables, (2) to contrast the uses of latent variables in different statistical models, and (3) to discuss common issues that emerge when using latent variables. The paper reviews several common ways of defining latent variables and introduces a sample realization definition. According to the sample realization definition, a latent random variable for a case is a random variable for which there is no sample realization for that case in a given sample. An analogous definition is given for latent nonrandom variables. Properties attributed to latent variables are reviewed. Following this are discussions of latent variables in multiple regression, limited dependent variable models, classical test theory, factor analysis, item response theory, latent structure analysis, general structural equation models, and random coefficients as latent variables. The paper concludes that the sample realization definition of latent variables offers some advantages in highlighting common problems and properties of these type of variables in typical statistical models.

e-mail: bollen@unc.edu

BAYESIAN LATENT VARIABLE METHODS IN BIOMEDICAL APPLICATIONS

David B. Dunson*, National Institute of Environmental Health Sciences

In recent years, biomedical studies have become increasingly complex, often involving highly multidimensional measurements having mixed discrete and continuous scales. In such settings, latent variable models are very useful for characterizing the complex dependency structure, while reducing dimensionality and simplify joint inferences (e.g, on the overall effect of a treatment or exposure variable). This talk briefly reviews recent Bayesian work on underlying normal and generalized linear models with latent variables. A new class of underlying Poisson-gamma latent variable models is then proposed for mixed discrete data, motivated by tumor studies that collect event time, count, and binary responses. Properties of the model are discussed, an efficient auxiliary variables MCMC algorithm is proposed for computation, and the approach is applied to data from a transgenic mouse bioassay study.

e-mail: dunson1@niehs.nih.gov

INTEGRATING BIOMETRIC AND PSYCHOMETRIC LATENT VARIABLE MODELING

Bengt O. Muthen*, UCLA

This talk discusses analysis with latent variables in a general modeling framework. The combined use of both categorical and continuous latent variables allows representation of common statistical concepts such as finite mixtures, random effects, and frailties as well as psychometric concepts such as latent classes and factors measured by multiple indicators. Observed outcomes may be continuous, categorical, counts, and limit-inflated. Examples include latent class regression models, growth mixture modeling with latent trajectory classes, discrete-time mixture survival modeling, latent transition modeling, hidden Markov mixture models, factor mixture analysis, structural equation models with non-linearities among the latent variables, modeling of non-ignorable missing data, and modeling of multilevel data. The general modeling framework integrates all these concepts and model components, leading to a rich set of extensions of existing models. Examples are drawn from biometrics and psychometrics and are estimated using the new version of the Mplus program, which implements this general modeling framework.

e-mail: bmuthen@ucla.edu

LATENT VARIABLE MODELS; CAUSAL ESTIMATION, EXTRAPOLATION AND SPECULATION

Scott L Zeger*, The Johns Hopkins University
Kathryn A Ziegler, The Johns Hopkins University

Latent variables models are widely used in biomedical and public health research to address causal questions. Examples include generalized linear mixed models, latent class and structural equation models. The parameters in these models are attractive because they often have causal interpretations. For example, a random effects logistic regression parameter approximates the ratio of an individual's risk given exposure to what that same person's risk would be if unexposed. This talk will discuss the evidence relevant to parameter estimation in latent class models and the role of model assumptions in quantifying that evidence. We will begin by reviewing the concepts of estimation and extrapolation in the simple problem of predicting a response Y at a new covariate value $X=c$ given a training set of X,Y values for X s in the interval $[a,b]$. We will then extend these ideas to the case of mixed models and propose graphical displays and quantitative measures that help us better understand where the inference about a particular parameter resides on the estimation – extrapolation – speculation continuum. The role for model averaging will be discussed. The ideas will be illustrated with an analysis of a recent public health data set.

e-mail: kziegler@jhsph.edu

2. Evaluating National Security Surveillance Tools

NONPARAMETRIC MEASURES OF DEPENDENCE, AND AGGREGATED ALGORITHMS, IN BIOMETRIC RECOGNITION PROBLEMS

Andrew L. Rukhin*, NIST and UMBC

Automated biometric systems to detect or verify personal identity using physiological and/or behavioral characteristics are widely used in homeland security. Various such systems are now commercially available, and recent technological progress makes it possible to evaluate their performance consistently and comprehensively. In identification systems, a biometric signature of an unknown person is presented to a system which compares this ‘probe’ with a database of biometric signatures of known individuals, called the gallery. From this comparison, the system reports similarity scores relating the probe to the signatures in the gallery. The gallery items are then ranked according to these similarity scores. The matches with highest similarity scores are expected to contain the true identity. This talk addresses two related issues: how to compare, and how to combine, different face recognition algorithms on the basis of their similarity scores. An example from the Face Recognition Technology program, with four face recognition algorithms, is examined.

e-mail: rukhin@cam.nist.gov

SYNDROMIC SURVEILLANCE: IS IT WORTH THE EFFORT?

Michael A. Stoto*, RAND
Matt Schonlau, RAND
Louis T. Mariano, RAND

Syndromic surveillance systems are intended to give early warnings of bioterrorist attacks. However, even with access to the requisite data and perfect organizational coordination and cooperation, the statistical challenges in detecting an incident are formidable, involving a tradeoff among sensitivity, the false positive rate, and timeliness. We analyzed four statistical detection algorithms using daily counts of patients with influenza-like illness from a hospital emergency department. Under these conditions outside of the flu season, the detection algorithms had a roughly 50 percent chance of detecting a “fast” outbreak on the second day. The “slow” outbreak was more difficult to detect; detection algorithms that integrate data from more than one day were roughly equivalent to each other but have only a 50 percent chance at day 9. The benefits of syndromic surveillance also depend on how well these systems are integrated into public health. Even in the best of circumstances syndromic surveillance sets off an alarm that must be investigated before any action can be taken, and deciding what to do can be difficult. Given these results, city and state health departments should be cautious in investing in costly syndromic surveillance systems.

e-mail: mstoto@rand.org

ASSESSING PERFORMANCE AND WEIGHING EVIDENCE FROM NATIONAL SECURITY SCREENING TOOLS

Peter B. Imrey*, Ph.D., Cleveland Clinic Foundation
David L Faigman, J.D., Hastings College of the Law

The controversial Defense Advance Projects Research Agency Program titled “Total Information Awareness,” subsequently renamed “Terrorism Information Awareness (TIA),” researched technologies with potential counterterrorist applications, including database aggregation initiatives and development of tools oriented in some sense “biometrically.” The latter included Human-ID, integrating automated facial recognition, human physical interaction patterns, and other information; and the Bio-Event Advanced Leading Indicator Recognition Technology (Bio-ALIRT) Program for disease outbreak detection. Although Congress eliminated 2004 DARPA funding for TIA, some components (e.g., Bio-ALIRT) were excepted and other efforts have likely been transferred, with deployment restrictions, to the classified National Foreign Intelligence Program (NFIP). Informed by previous work on national security polygraph screening, this paper will consider statistical issues common to evaluations of high-throughput “biometric” screening technologies, and suggest how statistical and other perspectives might influence how courts deal with their results. Implications and inadequacies of possible evaluation models, such as the conventional signal detection paradigm for disease diagnosis and screening, and of how the legal system processes scientific testimony, will be discussed.

e-mail: pimrey@bio.ri.ccf.org

3. Recent Developments in Phase I Trial Design

DETERMINING A MAXIMUM TOLERATED SCHEDULE OF A CYTOXIC AGENT

Thomas M. Braun*, University of Michigan
Peter F. Thall, M.D. Anderson Cancer Center
Zheng Yuan, University of Michigan

Most phase I clinical trials are designed to determine a maximum tolerated dose (MTD) for a single initial administration or treatment course of an experimental cytotoxic agent. In actual clinical practice, however, physicians often administer the agent to the patient repeatedly and monitor long-term toxicity due to cumulative effects. We propose a new dose-finding method for such settings, based on the time to toxicity, rather than on a binary outcome used in traditional Phase I designs. Our model and method account for the patient’s entire sequence of administrations, with the overall hazard of toxicity modeled as the sum of a sequence of hazards, each associated with one administration. The goal is to determine a maximum tolerated schedule, rather than a conventional MTD for only the first course. Data monitoring and decision making are done continuously throughout the trial. We illustrate the method in an allogeneic bone marrow transplantation trial seeking to determine how long a recombinant human growth factor can be administered as prophylaxis for acute graft-versus-host disease. We also present a simulation study in the context of this trial.

e-mail: tombraun@umich.edu

MODELING DOSE-TOXICITY RELATIONSHIP IN PHASE I CLINICAL TRIALS

Ken Cheung*, M.D. Columbia University

The continual reassessment method (CRM) is a sequential design used in phase I clinical trials to determine the maximal dose with acceptable toxicity. It has been established that the CRM is consistent under certain conditions but not generally. When the method does not converge to the target dose, some dose-response models will be more sensitive than others in terms of how close the converged dose is to the target. This talk gives a quick survey on the choice of dose-response models and provides a simple technique to evaluate the sensitivity of one-parameter models when used with the CRM. The technique presented is found to be a useful tool in planning phase I trials that use the CRM.

e-mail: cheung@biostat.columbia.edu

DOSE-FINDING WITH TWO AGENTS IN PHASE I ONCOLOGY TRIALS

Peter F. Thall*, M.D. Anderson Cancer Center
Randall E. Millikan, M.D. Anderson Cancer Center
Peter Mueller, M.D. Anderson Cancer Center
Sang-Joon Lee, M.D. Anderson Cancer Center

An adaptive two-stage Bayesian design is proposed for finding one or more acceptable dose combinations of two cytotoxic agents used together in a phase I clinical trial. The method requires dose-toxicity information on each of the two agents used alone as a single agent, either elicited from the physicians planning the trial or obtained from historical data. A parametric model is assumed for the probability of toxicity as a function of the two doses. Informative priors are constructed for parameters characterizing the single-agent toxicity probability curves, and vague priors are assumed for parameters characterizing two-agent interactions. A practical method for eliciting the single-agent parameter priors is described. The design is applied to a trial of gemcitabine and cyclophosphamide, and a simulation study of the design is presented.

e-mail: rex@mdanderson.org

4. Statistics and Neuroscience

MAXIMUM LIKELIHOOD ESTIMATION FOR MULTIVARIATE POINT PROCESS MODELS OF NEURAL SPIKING ACTIVITY

Emery Brown*, Harvard University

Multivariate point process models are a much needed tool of analyzing neural spiking activity. While these models are straight forward to specify for simulation purposes, their use in estimation has seen limited application. We describe an efficient implementation of an algorithm originally developed by Chornoboy for fitting multivariate point process models by maximum likelihood. We describe the theoretical properties of the algorithm and illustrate its application by constructing a multivariate point process model to a collection of 32 hippocampal neurons recorded simultaneously during 13 minutes. The algorithm suggests a computationally tractable approach to assessing network structure in an ensemble of simultaneously recorded neurons.

e-mail:

FINDING CONSISTENT ACTIVATION ACROSS FMRI TASKS: USING PFDR TO TEST A UNION OF NULLS

Thomas E Nichols*, University of Michigan

Psychologists studying memory use Functional Magnetic Resonance Imaging (fMRI) to understand how information is encoded, stored and retrieved in the brain. Short term, or working memory, can be ‘probed’ in a number of ways, say by asking a subject to remember a list of words, or a list of digits, or even even a collection of shapes (triangles, squares, etc). The goal is to identify regions of the brain that support working memory generically, but which are not specialized to words or numbers or shapes. This requires testing a union of nulls (no effect in one or more of the three tasks) versus an intersection of alternatives (effects in all three tasks). Worsley and Friston proposed using the maximum P-value to test for an intersection of effects, but their inference is based on the intersection of nulls. We propose using Storey’s Positive False Discovery Rate (pFDR) to make inference on the union of nulls. Storey shows that pFDR can be interpreted as the posterior probability of the null given that a statistic lies in a rejection region. Our method can approximately be seen as making inference on the sum of q-values. We demonstrate the method on simulated and real data.

e-mail: nichols@umich.edu

STATISTICAL MODELS AND ISSUES FOR NEURAL PROSTHESES

Valerie Ventura, Carnegie Mellon University
Cari Kaufman*, Carnegie Mellon University

Neurons in motor cortex areas encode information about movement variables in their firing rate, so that, reversing the problem, it is possible to predict movements given the observed firing rates of a population of neurons. Now that it is possible to record the activity of many neurons simultaneously, this paradigm can be used to build prosthetic devices. In this presentation, we propose a statistical model that allows a monkey to use a prosthetic arm for reaching tasks. Issues involve determining what movement or feed-back variables cortical neurons encode, building an efficient model that is robust to unpredictable variations in the brain, and producing smooth dynamic real-time arm movements. Key words: particule filter, importance sampling spherical Fisher distribution, point process models.

e-mail: vventura@stat.cmu.edu

5. Statistical Review in the Medical Literature

REVIEWING CLINICAL RESEARCH

James H. Ware*, Harvard School of Public Health

The role of biostatistical thinking and methods in the design, analysis, and interpretation of clinical studies has grown dramatically over the past two decades. Twenty- five years ago, most reports of clinical research employed rudimentary statistical methods. Today, most clinical research requires strong statistical collaboration. In response, leading medical journals have strenthened the statistical review of submitted manuscripts. Many leading journals appoint biostatisticians to their editorial staff. Recently, growing interest in meta-analysis and research synthesis has generated a movement toward greater standardization of reporting, as reflected in the CONSORT guidelines for reporting of clinical trials and similar guidelines for other types of research. Such guidelines, however, can address only the most basic elements of the quality of statistical reporting. This talk will build on more than a decade of experience as a statistical consultant to the New England Journal of Medicine to reflect on the best practices in the statistical review of clinical research, challenges for biostatistical reviewers, and the potential for setting useful standards for reporting to guide authors and reviewers. Reflections are offered on the some of the common methodologic disagreements that lead to disagreements between authors and reviewers.

e-mail: ware@hsph.harvard.edu

STATISTICAL REVIEW IN THE MEDICAL LITERATURE

Russell Localio*, University of Pennsylvania

Although statistical review in the medical literature has clearly improved the quality of both the design and reporting of medical studies, the frequency and type of statistical problem in submitted and published manuscripts reflect a series of continuing structural and process barriers to the best use of statistical methods. The presentation will outline some of the persistent statistical problems and then discuss the possible influence of these factors: (1) Medical schools have not figured out how to pay for biostatistics. (2) Some clinicians view statistics as a static science that has been well-distilled into commercial software that they can use without statistical support. (3) Large studies with many authors, all looking out for promotion, lead to thin slices from the data and analysis deli. (4) Inertia and bad precedent continue to generate suboptimal use of statistics (What was good enough for JAMA (or NIH or the FDA) is good enough for us). Options for medical journal editors, statisticians, educators, and clinical investigators will be discussed.

e-mail: rlocalio@cceb.upenn.edu

WHY ALL BIOTATISTICIANS SHOULD REVIEW FOR MEDICAL JOURNALS, OR, HOW TO RETAIN YOUR LIMBS WHILE SWIMMING WITH SHARKS

Steven Goodman*, Johns Hopkins University

The author will present insights and examples from his experience of serving as a statistical editor for *Annals of Internal Medicine* for the past 17 years. Research done over the last decade on the effects of statistical review and the statistical review policies of other medical journals will also be presented and discussed. Many statisticians complain about the quality of methods employed in published medical research without helping to solve the problem. The case for all biostatisticians playing a role in the review of medical research will be presented, along with a number of cautions.

e-mail: sgoodman@jhmi.edu

6. Microarray Normalization and Expression Indices

OPTIMAL USE OF PROBE LEVEL INFORMATION IN AFFYMETRIX GENECHIP ANALYSIS

Jianbo Li*, Cleveland Clinic Foundation

Microarray technologies revolutionize life sciences through simultaneous assessment of expression of thousands of genes. Challenges posed by the analysis of oligonucleotide microarray data lie in part in the uncertainty of the association between signal intensity and the level of gene expression. We examine the signal variability at the probe level, a major source of variation, with the consideration of possible non-specific hybridization and weak hybridization of the perfect match and the mismatch probes, as well as the number of probe pairs for each gene. We develop a method for optimal use of such information in analysis of Affymetrix GeneChip data.

e-mail: lij1@ccf.org

NORMALIZATION OF MICROARRAY DATA USING MIXTURE OF SPLINES

Yongsung Joo, University of Florida
George Casella, University of Florida
James Booth, University of Florida
Keunbaik Lee*, University of Florida

The normalization has been one of big issues in analyzing microarray data. This paper proposes to detrend the dye bias in cDNA microarray analysis using the Bayesian mixture of splines model. Also we developed a test for between group comparisons of each gene, using the average of cumulative probabilities.

e-mail: lee@stat.ufl.edu

A TWO-WAY SEMI-LINEAR MODEL FOR NORMALIZATION AND SIGNIFICANT ANALYSIS

Deli Wang*, The University of Iowa
Jian Huang, The University of Iowa
Cuihui Zhang, Rutgers University

Microarray analysis is a technology for monitoring gene expression levels on a large scale and has been widely used in functional genomics. A basic problem in the analysis of cDNA microarray data is normalization. A proper normalization procedure ensures that the intensity ratios provide meaningful measures of relative expression values. We propose a two-way semi-linear model (TW-SLM) for normalization and significant analysis of microarray data. This method does not make the usual assumptions underlying some of the existing methods. For example, it does not assume that: (i) the percentage of differentially expressed genes is small; or (ii) the numbers of up- and down-regulated genes are about the same, as required in the lowest normalization method. The TW-SLM also naturally incorporates uncertainties due to normalization into significant analysis of microarrays. We use a robust semiparametric approach to estimate the normalization curves and the normalized expression values. We also conduct simulation studies to evaluate the TW-SLM method and apply it to two microarray data sets as illustrations.

e-mail: deli-wang@uiowa.edu

EVALUATION OF AFFYMETRIX PROBE LEVEL SUMMARIZATION ALGORITHMS USING QUALITY CONTROL GENES

Yaomin Xu*, Cleveland Clinic Foundation
Xuejun Peng, Cleveland Clinic Foundation

Multiple probe pairs are designed for an individual gene with Affymetrix oligonucleotide microarray technology. It has been a big challenge to develop an algorithm to summarize over these multiple probe pairs to get a good measure that truly reveals the relative abundance of mRNA copies in the sample. Many algorithms have been proposed, such as MAS5.0, Li-Wong and RMA, etc. Comparisons of these algorithms have been made using the Spike-in data and the dilution data. Here we compare them using the quality control probe sets designed on every Affymetrix commercial Genechips®. The consistency and discrepancy between the above algorithms are evaluated. Using the permutations of the arrays, we also evaluated the false positive rates of each algorithm. We found that MAS5.0 has the lowest false positive rate.

e-mail: yxu@bio.ri.ccf.org

SOME ISSUES ON CROSS-VALIDATION IN MICROARRAY DATA ANALYSIS

Wenjiang M. Fu*, Texas A & M University

Cross-validation (CV) has been used widely in model fitting and data analysis to prevent overfitting or to select tuning parameters. In recent works, it has been used to estimate misclassification error in micro-array data analysis. Although it provides unbiased estimation for error rate in general, its large variation for small samples makes it less competitive. In this talk, we will consider some issues on the CV and propose some methods aiming to improve the CV in the special setting, i.e. $n \ll p$, for micro array data.

e-mail: wfu@stat.tamu.edu

A CASE STUDY ON CHOOSING NORMALIZATION METHODS AND TEST STATISTICS FOR MICROARRAY DATA

Yang Xie*, University of Minnesota
Wei Pan, University of Minnesota
Keyong S. Jeong, University of Minnesota
Arkady Khodursky, University of Minnesota

DNA microarray is a biological technology which permits the whole genome to be monitored simultaneously on a single chip. Some very common questions for microarray data analysis are: Should we normalize arrays to remove the potential systematic biases? What normalization method to use? How should we implement statistical tests? There may not be uniform and easy answers to these questions. In this paper, we use a real data example to illustrate practical approaches to addressing these questions. The example is taken from a DNA-protein binding microarray experiment aimed at furthering our understanding of transcription regulation mechanisms, one of the most important issues in biology. For the purpose of preprocessing data, we suggest looking at descriptive plots first to decide whether we need do normalization or not, and if yes, what normalization method to use. For one-sample comparison, we recommend using an empirical Bayes method (B statistic) since it performs much better than traditional methods, such as the sample mean (M statistic) and Student's t statistic, and it is also relatively easy to compute and explain when compared to others. False discovery rate is used to evaluate the different methods, and our comparative results lend support to our above suggestions.

e-mail: yangxie@biostat.umn.edu

ADAPTIVE SVD APPROACHES OF ESTIMATING EXPRESSION INDEXES FOR OLIGONUCLEOTIDE ARRAYS

Jianhua Hu*, University of North Carolina at Chapel Hill
Fred Wright, University of North Carolina at Chapel Hill
Fei Zou, University of North Carolina at Chapel Hill

A widely used type of microarray is the multiprobe oligonucleotide array, with the attractive feature of the probe redundancy. The term “expression index” describes a statistic used to represent expression level for a particular gene estimated from raw hybridization intensities on the array. The careful estimation of expression indexes becomes important, because the resulting statistical inferences in an experiment are all based on it. We discuss two approaches for constructing expression indexes using singular value decomposition (SVD). One approach can adaptively estimate expression indexes from raw intensities of oligonucleotide array probes, and requires a data structure often employed in array experiment. We show that a popular model-based expression index proposed by Li and Wong (2001a) is a special case of our estimates. The other approach uses a data transformation guided by an entropy measure, which itself is computed from the SVD-based estimates. The entropy-based approach can be viewed as a means of improving model fit, thereby improving the estimated expression index. The methods are demonstrated with simulation and applications to two real data sets.

e-mail: jhu@bios.unc.edu

7. Missing Data for Longitudinal Data Trials

A PATTERN-SELECTION APPROACH FOR THE ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE MISSINGNESS

Soomin Park*, Eli Lilly and Company
Lei Shen, The Ohio State University

The phenomenon of missing data is common in longitudinal studies. If data are not missing at random and mixed effect models are used in the analysis, then the missingness is non-ignorable, in the sense that ignoring the missing mechanism leads to biased estimators. Random-effect-based non-ignorable missingness, also termed informative missingness, is reasonable when the missing probabilities are related to unobserved individual characteristics. In this talk we illustrate the relationship between selection and pattern-mixture models. We then propose a pattern-selection approach to combine the strengths of these two types of models. Some components of this approach, such as the computation of summary measures for missingness and the choice of identifying restrictions, are guided by a selection model, while the estimation is in principle done using a pattern mixture model to achieve robustness. Efficient estimators are obtained using an empirical Bayes method.

e-mail: park_soomin@lilly.com

ANALYSIS OF THE BIAS CAUSED BY USING THE AVAILABLE CASE METHOD WITH LONGITUDINAL DATA

Lin Wang*, University of Wisconsin Madison
Mari Palta, University of Wisconsin Madison
Jun Shao, University of Wisconsin Madison

When correlated observations are made repeatedly over time on the same individuals, there are often observations missing at one or several time points. Available case analysis is the simplest method to apply in the analysis of such data, but depending on the model structure and missing mechanism, biased estimators may result. We derive an upper bound for the asymptotic bias in regression coefficients resulting from available case analysis for Gaussian data where the probability of observing the outcome at a given data point follows a probit model conditionally on either some preceding or the present outcomes. This upper bound can be calculated directly from the data. Also a GEE method with separate weight for each observation is considered to adjust for bias.

e-mail: wangl@stat.wisc.edu

NONPARAMETRIC MULTIPLE IMPUTATION OF INCOMPLETE LONGITUDINAL DATA VIA FUNCTIONAL MIXED MODEL

Trivellore E. Raghunathan, University of Michigan
Yulei He*, University of Michigan

Unbalanced data, where not all the individuals are observed at the same time points, is a common feature in many longitudinal studies. The primary reason is due to nonparticipation even though the original design may have called for measuring all the individuals at the same time. Alternatively, by design the data may not be collected from all the individuals at every wave of data collection. A case study that motivated this research involved relating the wealth of the parents during the critical developmental age on their childrens development. Similar problems occur when the longitudinal data are used to characterize the exposure information over a life course or during a certain critical period. Multiple imputation approach provides a framework for handling missing data in such instances. This paper discusses a semiparametric approach using spline models, within a Bayesian framework, to create imputations. Gibbs sampling is used to obtain the draws. We illustrate the proposed method by applying it to the data set of Panel Study of Income Dynamics (PSID). A simulation study is carried out to evaluate the repeated sampling properties of the inferences obtained using this approach.

e-mail: yuleih@umich.edu

ESTIMATING TREATMENT EFFECTS FROM LONGITUDINAL CLINICAL TRIAL DATA WITH MISSING VALUES: COMPARATIVE ANALYSES UNDER DIFFERENT MODEL ASSUMPTIONS

Patricia R. Houck*, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh
Tulay Koru-Sengul, University of Pittsburgh
Gong Tang, University of Pittsburgh
Benoit Mulsant, University of Pittsburgh
Bruce Pollock, University of Pittsburgh
Charles F. Reynolds, University of Pittsburgh

Different analytic approaches for handling missing data in clinical trials can yield conflicting inference. The selection of a method to analyze clinical trial data often depends on the field of practice and the availability of software. Traditional methods such as endpoint ANCOVA in completers and intent-to-treat ANCOVA with LOCF, assumes MCAR. Current methods such as ignorable maximum likelihood (ML) and multiple imputation (MI) assume the data are missing at random (MAR). We have applied these methods along with some unconventional methods such as piecewise modeling and rescue (treatment adjustment) approach to a clinical trial comparing the antidepressant effects in an elderly depressed sample. The different methods give a broad range of results, which affects the inference from the trial. We conclude that characterization of the missingness mechanism should be an integral part of any clinical trial data analysis. Research supported in part by the Mental Health Intervention Research Center for the Study of Late Life Mood Disorders (MH52247) and the Mental Health Intervention Research Center for Mood and Anxiety Disorder (MH30915).

e-mail: houckpr@upmc.edu

MULTIPLE IMPUTATION FOR LONGITUDINAL DATA WITH INFORMATIVE MISSINGNESS

Wei Deng*, The Ohio State University
Lei Shen, The Ohio State University

The method of multiple imputation by Rubin (1978) is widely used to handle missing data. It calls for imputing draws from a predictive distribution and incorporates the sampling variability due to the missing values. If data are missing at random in the sense of Rubin (1976), multiple imputation based on the correct missing data model, when used along with maximum likelihood, yields consistent estimators and valid inference. However, multiple imputation for longitudinal data, especially when missingness is not at random, has not been well studied. We consider longitudinal data with informative missingness, where the missing data process depends on the individual random effects, propose natural extensions of the conventional multiple imputation method, and study their performance in this situation. It is shown, both analytically and using simulations, that imputing draws from appropriate predictive distributions can correct the bias in the naive estimators due to ignoring the missing data mechanism.

e-mail: deng.32@osu.edu

SEMIPARAMETRIC MODELS FOR LONGITUDINAL BIVARIATE ORDINAL OUTCOMES SUBJECT TO INFORMATIVE DROPOUTS

David Todem*, Michigan State University
Kyung Mann Kim, University of Wisconsin-Madison

Analysis of clinical trials on antidepressant drugs is typically based on a bivariate ordinal vector comprising information on efficacy and safety parameters. Additionally, it is often the case that subjects drop out prematurely from study, yielding unbalanced data. A statistical analysis of these data raises a number of challenging issues. The clustering due to repeated observations from the same subject and the multiplicity of outcomes necessitate the use of methods for correlated data. The dropout process is another important issue. If the censoring process is related to unobserved outcomes, estimates of the models that do not incorporate this mechanism may be biased. We assume a likelihood-based approach that simultaneously models bivariate ordinal outcomes and the attrition process within a unified model-based framework. Estimating this nonignorable model is, however, not our priority because the numerical optimization can be difficult and the model can be unidentifiable. Instead, this class of models is used to assess perturbations of the estimates near the ignorable model. Data from a psychiatric trial, the Fluvoxamine (an antidepressant drug) study are used to illustrate the method.

e-mail: todem@epi.msu.edu

APPLICATION OF PATTERN MIXTURE MODELS TO NON-IGNORABLE MISSING DATA USING PSEUDO MAXIMUM LIKELIHOOD ESTIMATION

Changyu Shen, University of Pittsburgh
Lisa Weissfeld, University of Pittsburgh

In this work, we fit pattern mixture models to data with a non-ignorable missing response. In estimating the regression parameters that are identifiable, we use the pseudo maximum likelihood method introduced by Gourieroux et al (1984). This procedure provides a consistent estimator when the mean structure is correctly specified for each pattern, with further information on the variance structure giving an efficient estimator. The estimator is much easier to compute than the current estimators for this problem. Simulations are carried out to compare this procedure with other methods. A Hausman type test (Hausman, 1978) of model misspecification is also developed. Finally we apply the proposed method to an epidemiologic cohort study to examine cognition decline among elderly.

e-mail: chsst52@pitt.edu

8 Applications in Statistical Genetics

EFFECT OF MEASUREMENT ERROR IN REPORTED FAMILY HISTORY ON PREDICTIONS FROM MENDELIAN MODELS OF DISEASE GENES

Hormuzd A. Katki*, National Cancer Institute, NIH, DHHS
Giovanni Parmigiani, John Hopkins University

Patients with familial history of disease often consult with genetic counsellors about their chance of carrying mutations that increase disease risk. To aid them, many models have been proposed that combine the patient's reported family history with knowledge of each gene's transmission, penetrance and prevalence to predict whether the patient carries deleterious mutations. Such models rely on accurate information about whether and when each family member developed the disease. Unfortunately, accurate information is often unavailable, so it is important to evaluate the sensitivity of the predictions to mistakes in reported family history. General results about the sensitivity of the predictions to different types of errors in reported family history are derived. If a patient is unsure of some family history information, these results lend guidance as to what types of family history information demand verification. We apply the results to the BRCA1/2 model, which predicts the risk of carrying a mutation in the breast and ovarian cancer genes BRCA1 and BRCA2.

e-mail: katkih@mail.nih.gov

EFFICIENT ALGORITHMS FOR ALLELE SHARING METHODS TO DETECT LINKAGE USING EXTENDED PEDIGREES

Saonli Basu*, University of Washington, Seattle

Allele sharing methods provide a robust approach for linkage detection by studying the association in the joint inheritance pattern of the trait and a set of markers in pedigrees. Existing allele sharing methods like sib pair or relative pair studies are less informative about the joint segregation pattern of trait and marker alleles than studies using moderate or large size pedigrees with a large set of affected individuals and individuals typed for multiple markers. Moreover, affected individuals in extended pedigrees tend to be more homogeneous with respect to the genetic basis of the trait and environmental background. Hence, allele sharing methods using extended pedigrees can have enormous potential for genetic mapping of traits, especially complex traits, where it is difficult to model the trait. I propose methodologies for efficient implementation of allele sharing methods using extended pedigrees for linkage detection that includes construction of a robust IBD measure, designing a test statistic that makes more efficient use of the available data and suggesting a computationally efficient testing procedure to test for linkage. Since exact computation of the test statistic is infeasible on large pedigrees, I use MCMC methods to provide a consistent estimate of the statistic.

e-mail: saonli@stat.washington.edu

EXTREME SIBLING PAIR DESIGNS FOR QUANTITATIVE TRAIT LOCUS MAPPING

Jin P. Szatkiewicz*, University of Pittsburgh
Eleanor Feingold, University of Pittsburgh

Gu et al. (1996) proposed the so-called “extreme discordant and concordant” (EDAC) sampling design for quantitative trait locus mapping using sibling pairs. Since then, the design has been used a number of times, and there has been a fair amount of literature further developing the theory. There has not, however, been a systematic study of the interplay between sampling plan and choice of test statistic for such designs. This is especially important given that there are a number of new statistics in the literature that are appropriate for EDAC data. In this work, we discuss several of the most important EDAC-type designs in the literature, and use simulation studies to show what statistics are most powerful for each design. We then compare the power of the different designs, using the most powerful statistic for each. We conclude that the most powerful statistics for optimal EDAC samples are to be found among the new statistics; the statistic proposed by Gu et al. (1996) is no longer the best choice. We also suggest that certain EDAC designs are clearly more powerful and cost-effective than others - in particular it is probably best to sample more discordant pairs and fewer concordant pairs than most previous studies have done.

e-mail: jip5@pitt.edu

A HIDDEN MARKOV MODELING APPROACH FOR ADMIXTURE MAPPING OF COMPLEX DISEASES

Chun Zhang*, University of California, Davis
Kun Chen, University of California, Davis
Michael F. Seldin, University of California, Davis
Hongzhe Li, University of California, Davis

Admixture mapping is potentially a powerful method for mapping genes for complex human diseases, when the disease frequency due to a particular disease susceptible gene is different between founding populations of different ethnicity. The method tests for genetic linkage by detecting association of the allele ancestry with the disease. Since the markers used to define ancestral populations are not fully informative for the ancestry status, direct test of such association is not possible. In this paper, we develop a hidden Markov model (HMM) framework for estimating the unobserved ancestry haplotypes across a chromosomal region based on marker haplotypes. The HMM efficiently utilizes all the marker data to infer the latent ancestry states at the putative disease locus. In this modelling framework, we consider a likelihood based approach for detecting genetic linkage based on case-control data. We evaluate by simulations how several factors affect the power of admixture mapping, including sample size, ethnicity relative risk, marker density and the different admixture dynamics. Our simulation results indicate correct type 1 error rates of the proposed likelihood ratio test and great impact of marker density on the power. In addition, simulation results indicate that the methods work well for the admixed populations derived from both hybrid-isolation and continuous gene-flowing models.

e-mail: usamelie@yahoo.com

A METHOD FOR USING COMPLETE AND INCOMPLETE TRIOS TO IDENTIFY GENES RELATED TO A QUANTITATIVE TRAIT

Emily O. Kistner*, National Institute of Environmental Health Sciences
Clarice R. Weinberg, National Institute of Environmental Health Sciences

A number of tests for linkage and association with qualitative traits have been developed, with the most well known being the Transmission/Disequilibrium Test (TDT). For quantitative traits, varying extensions of the TDT have been suggested. The quantitative trait approach we propose is based on extending the log-linear model for case-parent trio data (Weinberg et al. 1998). Like the log-linear approach for qualitative traits, our proposed polytomous logistic approach for quantitative traits allows for population admixture by conditioning on parental genotypes. Compared to other methods, simulations demonstrate good power and robustness of the proposed test under various scenarios of the genotype effect, distribution of the quantitative trait, and population stratification. In addition, missing parental genotype data can be accommodated through an Expectation-Maximization (EM) algorithm approach. The EM approach allows recovery of most of the lost power due to incomplete trios.

e-mail: ekistner@bios.unc.edu

STOCHASTIC SEARCH GENE SUGGESTION: A BAYESIAN HIERARCHICAL MODEL FOR GENE MAPPING

Michael D. Swartz*, Rice University, M.D. Anderson Cancer Center
Marek Kimmel, Rice University

Mapping the genes for a complex disease, such as Diabetes or Rheumatoid Arthritis, involves finding multiple genetic loci that may contribute to the onset of the disease. Pairwise testing of the loci leads to the problem of multiple testing. To avoid multiple tests, we can look at haplotypes, or linear sets of loci; but this results in a contingency table with sparse counts, especially when using marker loci with multiple alleles. Using case-parent triad data, we construct a hierarchical Bayesian model using a conditional logistic regression likelihood to model the probability of transmission to a diseased child. We extend the Bayesian model developed by Thomas, et al. [(1995) Genetic Epidemiology 12:455-466] by defining prior distributions on the allele main effects that model the genetic dependencies present in the HLA region of Chromosome 6. We also added a hierarchical level for model selection that accounts for both locus and allele selection. Thus we cast the problem of identifying genetic loci relevant to the disease into a problem of Bayesian model selection. We evaluate the performance of the procedure with some simulated examples and then apply our procedure to identifying genetic effects influencing risk for Rheumatoid Arthritis.

e-mail: mswartz@stat.rice.edu

9. Clinical Trials and Epidemiologic Methods

RECURRENT EPISODES ANALYSIS

Jun Yan*, University of Iowa

Jason P. Fine, University of Wisconsin-Madison

In a clinical trial to assess the efficacy of a treatment in treating cystic fibrosis patients, many of the patients experienced multiple episodes of pulmonary exacerbations. When a patient has an exacerbation, the exacerbation lasts a random length of time, with longer episodes being more severe. A naive recurrent event analysis ignores the length of the episode which may be an important measure of the severity of the disease, and may be reflected in the quality of life of the patient and their associated medical cost. This paper constructs new temporal processes to reflect both the number and the length of the recurrences. Time-varying coefficient models for these new quantities will be developed using the temporal process regression. The time-varying coefficients are estimated nonparametrically from functional estimating equations, and their properties are developed using empirical process techniques. The practical utility of the model will be illustrated by simulations and a real data analysis.

e-mail: jyan@stat.uiowa.edu

A PHASE II TRIAL DESIGN BASED ON MEDIAN TIME TO EVENT

Theodore G. Karrison*, University of Chicago

Dezheng Huo, University of Chicago

In phase II cancer trials the primary outcome variable is usually the proportion of patients who achieve a 50% or greater reduction in the size of their tumor(s). However some have argued that the endpoint of a phase II trial should be comparable to that for phase III trials, i.e., survival-based. We therefore propose a two-stage design in which time to progression (or death) is the primary outcome variable. The investigator chooses an “uninteresting” value, m_0 , equal to the median time to progression achievable with standard therapy. The objective is then to determine whether the lower limit of the, say, 90% one-sided confidence interval for the median exceeds m_0 . A two-stage procedure can be implemented by analyzing the data after a specified fraction of patients, f , have had an event and terminating the trial if the observed median is no greater than m_0 . Simulations were carried out to evaluate type I and type II error rates, expected sample size and study duration. For $f=1/2$, the power loss is minimal ($< 2\%$). Under the null hypothesis, the savings in terms of study duration can be up to 25%, whereas the savings in terms of the number of patients entered depends upon the rate of accrual relative to the survival duration. Results for other choices of f are presented and discussed.

e-mail: tkarrison@health.bsd.uchicago.edu

PIECEWISE CONSTANT CROSS-RATIO ESTIMATES FOR BIVARIATE SURVIVAL DATA

Bin Nan*, University of Michigan
Xihong Lin, University of Michigan
Lynda D. Lisabeth, University of Michigan
Sioban D. Harlow, University of Michigan

We propose a sequential two-stage method of cross-ratio estimation for bivariate survival data assuming cross-ratio is a piecewise constant function of one of the two failure times. The estimation procedure estimates the marginal survival functions sequentially for the estimation of each constant cross-ratio using a selective data set with left truncation and right censoring, which handles covariates easily. The method is then applied to the analysis of Tremin Trust data for the association of women's age at a marker event and age at menopause.

e-mail: bnan@umich.edu

FIRST-YEAR SURVIVAL OF INFANTS BORN WITH CONGENITAL HEART DEFECTS IN ARKANSAS (1993-1998): A SURVIVAL ANALYSIS USING REGISTRY DATA

Mario A. Cleves, University of Arkansas for Medical Sciences
Sadia Ghaffar, University of Arkansas for Medical Sciences
Weizhi Zhao*, University of Arkansas for Medical Sciences
Bridget S. Mosley, University of Arkansas for Medical Sciences
Charlotte A. Hoobs, University of Arkansas for Medical Sciences

Background: In the United States and other developed nations, congenital heart defects (CHDs) are among the most prevalent and fatal of all birth defects. Here we report the survival probability of infants born with CHDs in Arkansas and examine the impact of multiple malformations on survival. **Methods:** Birth and death certificate records were linked to birth defects registry data for infants born with CHDs (1993-1998) in Arkansas. Both neonatal and first-year survival probabilities were estimated non-parametrically using Kaplan-Meier's product limit method. A Cox model was used to evaluate the relative importance of additional malformations on survival. **Results:** A total of 1,983 infants with CHDs were included. The neonatal survival probability for this cohort was 94.0% (95% CI: 93.0%, 95.1%), and the first- year survival probability was 88.2% (95% CI: 86.8%, 89.6%). The presence of hypoplastic left heart syndrome conferred the greatest reduction in survival. Infants with multiple CHDs had decreased survival compared to those with isolated heart defects. Survival was also adversely affected by the presence of congenital abnormalities in other body systems. **Conclusions:** Neonatal and first-year survival of infants with CHDs varies by both type of cardiac malformation and the presence of additional cardiac and non-cardiac malformations.

e-mail: zhaoweizhi@uams.edu

SAMPLE SIZE AND POWER OF RANDOMIZED DESIGNS

Feifang Hu*, University of Virginia

In the literature of randomized designs, the power and sample size are usually obtained by ignoring the randomness of the allocation. But the power is a random variable for a fixed sample size n , when a randomized design is used. In this talk, we focus on the power function (random variable) and the sample of two-arm (drug versus control) randomized clinical trials. We first derive an approximated distribution of the power function for a fixed sample size. Then a formula of sample size is derived for randomized designs. Applications and simulations are also reported.

e-mail: fh6e@virginia.edu

A GENERAL IMPUTATION METHOD FOR CAUSAL INFERENCE AND G-COMPUTATION FORMULA

Zhuo Yu*, Bristol-Myers Squibb

Mark van der Laan, University of California at Berkeley

Robins' causal inference theory assumes existence of treatment specific counterfactual variables so that the observed data structure augmented with these counterfactual variables satisfies a consistency and a randomization assumption. In this paper we provide a function which maps the observed data structure into such treatment specific counterfactual variables which satisfy the consistency and randomization assumption. This provides a general imputation method and leads thereby to a new statistical method for estimating causal parameters such as treatment specific counterfactual distributions. Robins [1987] shows that the counterfactual distribution can be identified from the observed data distribution by a G-computation formula for the discrete case. Gill & Robins [2001] prove the G-computation formula for continuous variables under some continuity assumptions and reformulation of the consistency and the randomization assumptions. We prove that if treatment is discrete, but the covariates and outcome are possibly continuous, then Robins' G-computation formula holds under the original consistency, randomization, and a generalized version of the identifiability assumption.

e-mail: zhuo.yu@bms.com

COMPARISON OF ESTIMATION TECHNIQUES FOR MEDICAL COSTS WITH CENSORED DATA

Corina M. Sirbu*, Michigan State University
Joseph C. Gardiner, Michigan State University

In medical follow up studies incomplete observation due to censoring would preclude ascertainment of outcomes in some subjects. Standard assumptions used in survival analysis do not apply to medical costs because the cumulative cost at the endpoint of interest will generally be correlated with the cumulative cost at the time of censoring. Several methods have been proposed for analysis of medical costs from censored data where selected sub-samples are used in estimation. We describe a general method based on a multi- state non-homogeneous Markov model that describes the probabilistic mechanisms that govern transition between states with costs incurred at transitions between states and during sojourn in states. Selection probabilities are estimated from the censored sample of event times using parametric, semi-parametric and nonparametric (Kaplan- Meier) estimators of the censoring distribution. We estimate net present value for costs incurred over a finite time horizon. When specialized to the two-state case, our model gives estimates similar to those previously described in the literature. Properties of these estimators are described and compared.

e-mail: sirbucor@msu.edu

10. Bayesian Clinical Trial Design and Analysis

BAYESIAN OPTIMAL DESIGN FOR PHASE II SCREENING TRIALS

Meichun Ding*, Rice University
Gary L. Rosner, M.D. Anderson Cancer Center
Peter Mueller, M.D. Anderson Cancer Center

Rapid progress in biomedical research necessitates clinical evaluation that identifies promising innovations quickly and efficiently. Rapid evaluation is especially important if the number of innovations is large compared to the supply of suitable study patients. Most phase II screening designs available in the literature consider one treatment at a time. Each study is considered in isolation. We propose a more systematic decision- making approach to the phase II screening process. The sequential study design allows for more efficiency and greater learning about treatments. Randomly many treatments are considered concurrently, allowing researchers to screen and find active treatments in less time than they would with serial evaluation. The approach incorporates a Bayesian hierarchical model that allows combining information across several related studies in a formal way and improves estimation in small data sets by borrowing strength from other treatments. The underlying probability model is a hierarchical probit regression model that also allows for covariates. The design criterion is to maximize the utility for the new treatment, which incorporates the total number of treatment successes over the standard treatment, a sampling cost per patient, and the possible demonstration of a significant treatment benefit in a future randomized clinical trial.

e-mail: glr@odin.mdacc.tmc.edu

ADJUSTABLE BAYESIAN TWO-STAGE DESIGN

Ming Li*, Vanderbilt University Medical Center
Yu Shyr, Vanderbilt University Medical Center

The nature of the research goal on phase II clinical trial makes Bayesian methodology appealing since it seems to fit the setting of decision theory naturally. However, the requirement for the utility functions and relatively complicate computation procedure may discourage many from adopting this approach. Most recently, an user friendly Bayesian two-stage design has been proposed by Tan and Machin (2002). Given all the strength of their methods, we argue that appropriate modification on the sample size reassessment could still be done. To keep the basic framework of their approach yet allow more flexibility on further sample size reconsideration at stage II, we update by adding more sample size choices for stage II. We show that the way of calculating new sample size for stage II is consistent with the original design and such reassessment can still guarantee to meet the constraints for overall response rate required by the original design. The decision criteria for comparisons of the results from stage I with pre-calculated critical values are provided. To keep it as user friendly as possible, our method emphasizes the practical aspect. Especially, for those trials showing promising at stage I, the overall sample size is reduced.

e-mail: ming.li@vanderbilt.edu

A BAYESIAN APPROACH FOR DESIGNING FUTILITY RULES IN LARGE CLINICAL TRIALS

Ben L. Trzaskoma*, Eli Lilly & Company
Andreas Sashegyi, Eli Lilly & Company

We consider a class of futility rules based on a Bayesian approach for computing the predictive probability of success (PPS) for large clinical trials, given a certain amount of observed data. This paper focuses on binary outcomes in particular, but with the choice of a suitable prior distribution the ideas extend to other types of response variables. We contrast the PPS under various assumptions about the prior distribution of future event rates, using the beta distribution as a natural and mathematically convenient choice. Even small absolute changes in the standard deviation of the beta can produce vast differences in the PPS, suggesting that the prior variance should not be arbitrarily assigned. The proposed method updates the prior distribution based on the observed data, assigning its mean and variance to a function of the data. This results in reasonable estimates of the PPS under a broad range of scenarios. Interpretation of the PPS and the impact of the prior distribution will be discussed in some detail. Guidance for the practical use of the suggested method will be given, based on the results of a simulation study.

e-mail: trzaskomabl@lilly.com

SAMPLE SIZE RE-ESTIMATION USING POSTERIOR PREDICTIVE POWER IN RANDOMIZED CLINICAL TRIALS WITH BINARY OUTCOMES

Hyung Woo Kim*, M.D. Anderson Cancer Center
Donald Berry, M.D. Anderson Cancer Center

It is important for clinical trials to be adequately powered. Sample size determination largely depends on initial estimates of the treatment effect. However, these estimates are usually uncertain. Sample size re-estimation using interim data can ensure that a study's objectives are achieved with adequate power. We explore a design in which sample size is re-estimated using posterior predictive power. Outcomes are binary. Using computer simulations we explore various design setups and present their operating characteristics, including type I error rate, power, and average sample size.

e-mail: hyungwooda@hotmail.com

PREDICTIVE PROBABILITY IN PHASE II CANCER CLINICAL TRIALS

Diane D. Liu*, M.D. Anderson Cancer Center
J. Jack Lee, M.D. Anderson Cancer Center

As an alternative to the commonly used two- or three-stage designs in phase II cancer clinical trials, a more flexible design is proposed based on computing the predictive probability in a Bayesian setting. The new design allows more frequent monitoring of the trial outcomes such that early termination of the trial can be reached when the interim data indicate that the experimental regimen is not promising. In the mean time, the predictive probability approach still possesses good frequentist's properties by controlling type I and type II errors. A bi-dimensional search algorithm is implemented to determine the design parameters. Exact computation and simulation studies demonstrate that the predictive probability approach is more efficient than the traditional multi-stage designs. The predictive probability design not only is more flexible in evaluating the study outcome but also remains robust even when deviations occur in monitoring cohorts of patients as specified in the trial design. Examples will be given to illustrate the statistical properties of various designs.

e-mail: jjlee@mdanderson.org

BAYESIAN APPROACH TO INDIVIDUAL BIOEQUIVALENCE

Pulak Ghosh*, Georgia State University
Ravindra Khattree, Oakland University

The concept of drug ‘interchangeability’ has been studied extensively in the literature since the seminal article by Anderson and Hauck (1990). Drug interchangeability can be classified as drug prescribability and drug switchability. While drug prescribability refers to Population bioequivalence (PBE), drug switchability refers to Individual bioequivalence (IBE). Based on these concepts, the FDA (1999) has recently recommended the adoption of IBE and PBE procedure for declaring two drugs to be bioequivalent. In this talk I will develop a new testing procedure for IBE and PBE based on a Bayesian Approach. It is applicable to both balanced and unbalanced data in a broad class of crossover design. I will also discuss several advantages of using Bayesian approach. I will compare our method with the existing FDA test procedure using real data from FDA.

e-mail: pghosh@mathstat.gsu.edu

TWO-SAMPLE CONFIDENCE INTERVALS FOR THRESHOLD VIOLATION PROBABILITIES: APPLICATION TO NONINFERIORITY TESTING

Alan H. Feiveson*, Johnson Space Center
Linda C. Shackelford, Johnson Space Center

The standard noninferiority test is predicated on a definition of “noninferiority” in terms of means only; i.e. that the mean of an experimental treatment response is no “worse” than that of the control. More generally, we define the experimental treatment to be noninferior if the probability of violating an industrial or clinical acceptability threshold is no higher with the experimental treatment than with the control. In this setting, we define the threshold as an extreme percentage point of the distribution of the response variable under the control regimen. For the case of a single sample from a normal distribution, confidence limits for PV , the probability of violating a known threshold, can be obtained by an inverse application of tolerance limits (confidence intervals for percentage points); however in the two-sample scenario, the threshold is not known and must be estimated from the control sample. A method of obtaining confidence limits for PV in this two-sample case applicable to all control and experimental sample sizes is presented. If desired, these confidence limits can then be used to define a critical region for a threshold-based noninferiority test.

e-mail: alan.h.feiveson@nasa.gov

11. Spatial Modeling I

GENERALIZED MULTIPLE SPATIAL FACTOR MODEL

Xuan Liu*, University of Minnesota
Melanie M. Wall, University of Minnesota

Multivariate spatial data are very common in public health and environmental research. In this type of data, outcome variables are commonly correlated with each other within the same location and also correlated across locations. A generalized common spatial model proposed by Wang and Wall (2003) is very useful to model this type of data. This current paper extends the generalized common spatial factor model to allow for multiple underlying factors. We applied this model to two simulated data. Model identifiability and rotations issues for the multiple factor model are discussed.

e-mail: xuanliu@biostat.umn.edu

CONFIDENCE INTERVALS FOR THE RATIO OF INTENSITIES

Traci L. Leong*, Emory University
Andrew N. Hill, Emory University
Lance A. Waller, Emory University

Intensity estimation is widely used in spatial data analysis. Intensity refers to the average number of events expected per unit area and therefore, computing the ratio of intensities is useful for relative risk analyses. The existing methodology for confidence intervals for the ratio of intensities is pointwise. Unfortunately, we have spatially correlated data and the current methodology does not preserve the overall type I error. We explore different methods of confidence intervals for the ratio of intensities. What is presented in this work are different confidence interval methods of the ratio of intensities (both pointwise and simultaneous) and their coverage probabilities.

e-mail: tleong@sph.emory.edu

WEIGHTED U-STATISTICS FROM NON-IDENTICALLY DISTRIBUTED DATA WITH APPLICATION TO THE REGRESSION OF INTERPOINT DISTANCES

Dionne A. Graham*, Harvard School of Public Health
Marcello Pagano, Harvard School of Public Health

The modeling of distance matrices has applications in many fields. Linear regression can play an important role in the modeling; but its use, when applied to interpoint distances, is hampered by the highly correlated and heteroscedastic nature of these distances. Here we show that the estimated parameters from linear models can be written in the form of weighted U-statistics drawn from non-identically distributed data. As such, they are consistent and asymptotically normal under mild assumptions. The results are demonstrated through simulation and then used to model the relationship between human immunodeficiency virus genotypic- and phenotypic distances.

e-mail: dgraham@hsph.harvard.edu

EXPLORING GOODNESS-OF-FIT AND SPATIAL CORRELATION USING COMPONENTS OF TANGO'S INDEX OF SPATIAL CLUSTERING

Monica C. Jackson*, Emory University
Lance A. Waller, Emory University

One purpose of spatial analysis in public health is to detect local clusters or anomalies in patterns of disease. Historically, the spatial analysis literature tends to fall into one of two groups regarding spatial clustering. The statistical literature often assumes independent regional counts and seeks to identify local areas inconsistent with global patterns of disease risk. In contrast, the geography and spatial econometric literature often builds inference based on global and local indices of spatial autocorrelation. Tango (1995) developed a statistic to detect spatial clustering. Rogerson (1999) noticed that Tango's statistic could be broken into two components, the first measuring goodness-of-fit and the second measuring spatial autocorrelation. We explore the use of Rogerson's (1999) expression of Tango's (1995) index of spatial clustering as a hybrid between these two types of approaches, and provide examples of patterns driving the goodness-of-fit and spatial autocorrelation components of the statistic.

e-mail: monica@drlady.com

DECOUPLING LOCAL AND GLOBAL BEHAVIOR IN SPATIAL MODELS

Tilmann J. Gneiting*, University of Washington
Martin Schlather, Universite du Luxembourg

Fractal behavior and long-range dependence have been described in an astonishing number of physical, biological, geological, and socio-economic systems. Time series and spatial data have been characterized by their fractal dimension, a local measure of roughness, and by the Hurst coefficient, a global measure of long-memory dependence. Either phenomenon has been modeled and explained by self-similar random functions, such as fractional Gaussian noise and fractional Brownian motion. The assumption of statistical self-similarity implies a linear relationship between fractal dimension and Hurst coefficient and thereby links the two phenomena. In this talk, I will introduce stochastic models that allow for any combination of fractal dimension and Hurst coefficient, thereby demonstrating that the two phenomena are independent of each other. In particular, local and global behavior are decoupled. I will review the historical development of the aforementioned notions, and I will point to research opportunities for statisticians.

e-mail: tilmann@stat.washington.edu

MAXIMUM EFFICIENCY ROBUST TESTS FOR FOCUSED CLUSTERING

Pablo E. Bonangelino*, The George Washington University
Thomas A. Louis, Johns Hopkins University

Score tests have been demonstrated to be among the most powerful of tests for whether disease events cluster near possible sources of a harmful exposure, called focused clustering. However, score tests are dependent on a specific parameterization of the exposure, and will lose power if the assumed exposure is far from the truth. In this paper, we examine how to optimally specify the exposure when it is unknown. This is done by developing maximum efficiency robust tests (MERTs) for candidate sets of possible exposures as well as by examining individual score tests. The possible tests are compared by simulation and by application to leukemia incidence data from upstate New York. It is found that the exponential parameterization is preferred with respect to the shape of exposure, and that the MERT may be useful over a range of scale.

e-mail: pablo@gwu.edu

TEST FOR ISOTROPY FOR SPATIAL POINT PROCESSES

Yongtao Guan*, University of Miami
Michael Sherman, Texas A&M University
James A. Calvin, Texas A&M University

The development of second-order properties is essential for understanding and modeling spatial point processes. A common practice is to assume isotropy. While a convenient assumption, this is not always appropriate. A conventional approach to check for isotropy is to informally assess plots of either the direction-specific sample second-order intensity functions or K-functions. These graphical techniques, however, are often difficult to assess and open to interpretation. In this article, we propose a formal nonparametric approach to test for isotropy based on the asymptotic joint normality of the sample second-order intensity function. We derive an L_2 consistent subsampling estimator for the asymptotic covariance matrix of the sample second-order intensity functions and use this to construct a test statistic with a χ^2 limiting distribution. Applications to two real data sets, on crime occurrences and on locations of leukemia patients both in the Houston area, and simulations demonstrate the efficacy of the approach.

e-mail: yguan@miami.edu

12. Estimation and Inference in Mixed-Effects Models

LAPLACE APPROXIMATION BASED METHODS IN NONLINEAR MIXED EFFECTS MODELS

Lei Nie*, University of Maryland at Baltimore County
Edward F. Vonesh, Baxter Healthcare Corporation
Jinglin Zhong, Merck & Co.
Dibyen Majumdar, University of Illinois at Chicago

For nonlinear mixed effects models for clustered data, with fixed or flexible random effects distribution, the Maximum likelihood estimation in the general setup is difficult due to the fact that the marginal likelihood can be expressed only as an integral over the random effects. By applying the Laplace approximation to the marginal likelihood, we obtain an approximated likelihood. The estimator that maximizes the approximated likelihood may be viewed as ‘approximate’ maximum likelihood estimators. Consistency of these estimators is established and their performances are compared through a simulation study.

e-mail: nie@math.umbc.edu

HYPOTHESIS TESTING IN MIXED-EFFECTS MODELS

Edward F. Vonesh*, Applied Statistics Center, Baxter Healthcare Corporation

Except for very some limited applications (e.g., balanced random-coefficient growth curve models, split-plot designs, etc.), valid inference under the mixed-effects setting is based chiefly on asymptotic theory and large-sample tests rather than exact tests valid for large or small samples (e.g., F-tests, t-tests). In this talk, we examine various small-sample approximations used to test hypotheses in the presence of sparse data. Based on some limited simulations and an appeal to methods based on balanced random-coefficient growth curve models (linear or nonlinear), we propose the use of several small-sample tests appropriate for the analysis of a two-stage random-effects model. Extensions to two-stage mixed-effects models with both random and fixed effects are discussed and illustrated using simulations.

e-mail: voneshe@baxter.com

INFLUENCE AND RESIDUAL DIAGNOSTICS IN MIXED MODELS

Oliver Schabenberger*, SAS Institute Inc.

Measures to gauge the influence of one or more observations on the analysis are well established in the general linear model for uncorrelated data. Computationally, these measures present no difficulty because closed-form update expressions allow their evaluation without refitting the model. When applying notions of statistical influence to mixed models, certain complications arise. For example, data points that exhibit influence are likely to impact fixed effects and covariance parameter estimates. Closed-form update formulas are hard to find or require untenable assumptions. In repeated measures or longitudinal studies one is often interested in multivariate influence, rather than the impact of isolated points. The presentation examines influence and residual measures in mixed models and their utility in discerning influential cases and sets of observations.

e-mail: oliver.schabenberger@sas.com

SCORE TESTS FOR VARIANCE COMPONENTS

Geert Molenberghs*, Limburgs Universitair Centrum, Diepenbeek, Belgium
Geert Verbeke, Katholieke Universiteit Leuven, Leuven, Belgium

Whenever inferences for variance components are required, the choice between one-sided and two-sided tests is crucial. This choice is usually driven by whether or not negative variance components are permitted. For two-sided tests, classical inferential procedures can be followed, based on likelihood ratios, score statistics, or Wald statistics. For one-sided tests, however, one-sided test statistics need to be developed, and their null distribution derived. While this has received considerable attention in the context of the likelihood ratio test, there appears to be much confusion about the related problem for the score test. The aim of this paper is to illustrate that classical (two-sided) score test statistics, frequently advocated in practice, cannot be used in this context, but that well-chosen one-sided counterparts could be used instead. A tight relationship with likelihood ratio tests will be established, and all results are illustrated in an analysis of continuous longitudinal data using linear mixed models.

e-mail: geert.molenberghs@luc.ac.be

13. Methodological Issues in Preserving Confidentiality in Public Use Datasets

INFERENTIALLY VALID SYNTHETIC DATA SETS FOR DISCLOSURE LIMITATION

Trivellore E. Raghunathan*, University of Michigan

Demands for microdata from surveys, the availability of commercial databases with identifying information and advances in computer technology has increased concerns about protecting the confidentiality of respondents. Thus, the fundamental tension faced by the data collection and disseminating agencies is the trade-off between meeting the demands for microdata and the protection of identity of respondents, a serious pledge made by the data collector to every respondent. Several approaches have been developed that alter the data before releasing it to the public that can make it difficult for an intruder to identify the respondent. Such techniques include data swapping, post-randomization, transformation to mention only a few. The alterations may succeed in limiting disclosure, but may also introduce bias in statistical inferences. Multiple imputation approach can be used to create synthetic data sets for public release, with a dual goal of protecting confidentiality and providing valid statistical inferences. This article reviews this methodology, presents extensions and evaluates it using simulated data sets. A general purpose semi-parametric approach for creating multiple synthetic data sets is described and evaluated.

e-mail: teraghu@umich.edu

DISCLOSURE LIMITATION VIA PARTIALLY SYNTHETIC MULTIPLY IMPUTED DATASETS

Jerome P. Reiter*, ISDS, Duke University

To limit disclosure risks, one approach is to release partially synthetic, multiply-imputed datasets. These comprise the units originally surveyed, but some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, are replaced with multiple imputations. This differs from the fully synthetic approach of Rubin (1993) and Raghunathan, Reiter, and Rubin (2003), as not all released values are synthetic. In this talk, I discuss methods of generating partially synthetic data. I also discuss how users can obtain valid inferences using the concepts of multiple imputation for missing data (Rubin, 1987). The rules for combining point and variance estimates differ from those of Rubin (1987) and from those of Raghunathan et al. (2003). The approach is illustrated in simulation studies.

e-mail: jerry@stat.duke.edu

14. Modeling Treatment Use & Effectiveness in Mental Illness

A NATURALISTIC STUDY OF SERIOUS ADVERSE EFFECTS OF ATYPICAL ANTIPSYCHOTICS IN A LARGE MULTI-ETHNIC PUBLICLY INSURED US POPULATION

Marcela Horvitz-Lennon*, Harvard Medical School

Rusty Tchernis, Harvard Medical School

Larry Zaborsky, Harvard Medical School

Sharon-Lise Normand, Harvard Medical School and Harvard School of Public Health

Evidence on the relative risk of metabolic and nutritional adverse effects for atypical (AAs) versus conventional antipsychotics (CAs) remain conflicting. We undertook a naturalistic study of schizophrenia treatment episodes built with 1994-2001 Florida Medicaid claims data. Each episode had a 6-month pre-treatment, minimal exposure to antipsychotics, and a 12-month follow-up. A new diagnosis of diabetes, hyperlipidemia, or obesity, or initiation of treatment defined incidence. We assumed ignorability of treatment assignment based on observed covariates, used a matching estimator based on propensity scores, and evaluated significance with a Cochrane Mantel-Haenszel statistic. Eighty-eight percent of patients contributed 1 episode, and 41% of the 10,392 episodes involved AAs. Demographic and clinical characteristics predicted agent used. Unadjusted incidences (%), AAs vs. CAs, were: 4.6 vs. 4.7 (diabetes); 11.7 vs. 9.6 (hyperlipidemia); and 3.8 vs. 2.5 (obesity). Relative Risks were significant for hyperlipidemia (1.22, 95% CI: 1.06-1.41) and obesity (1.66, 95% CI: 1.27 - 2.16). We present evidence of higher hyperlipidemia and obesity risk for AAs vs. CAs based on principled analyses of a large and diverse sample. Lack of randomization is a limitation. Drug analyses and comparison with non-schizophrenic controls are forthcoming.

e-mail: horvitz@hcp.med.harvard.edu

EVALUATING THE USE OF REFRESHMENT SAMPLES FOR MAKING INFERENCES IN THE PRESENCE OF NON-IGNORABLE MISSING DATA

Douglas E. Levy*, Harvard Medical School
Sharon-Lise Normand, Harvard Medical School; and Harvard School of Public Health

Missing data are always a concern in longitudinal studies, potentially biasing estimation if data are not missing at random. However, knowledge of the missingness mechanism may allow unbiased estimation. Unfortunately, specifying the missingness mechanism often relies on assumptions that are unverifiable. Hirano et al. recently proposed using a 'refreshment sample' to test the assumptions of several common missingness models. Furthermore, Hirano et al.'s method allows for estimating more flexible, intermediate models for missing data called 'additive non-ignorable' models. While establishing the theory behind such models, Hirano et al. do not formally evaluate the robustness of their procedure. We evaluate the circumstances under which refreshment samples determine the correct missingness mechanism and assess the nominal coverage of 95% confidence intervals and the bias of estimates for a simple two period difference analyses using the proposed additive non-ignorable missingness model as well as established missingness models.

e-mail: delevy@fas.harvard.edu

ACCOUNTING FOR TREATMENT-NONCOMPLIANCE AND MISSING OUTCOMES IN RANDOMIZED TRIALS

A. James O'Malley*, Harvard Medical School
Sharon-Lise Normand, Harvard Medical School

Randomization is the medium through which causal inferences are drawn about the effect of treatment. When randomized trials are broken via noncompliance and non-response, analyses must account for these mechanisms to ensure that reliable results are obtained. Frangakis and Rubin (1999) recently developed a methods-of-moments (MOM) estimator to account for non-compliance and non-response in trials involving cross-sectional outcomes. The Frangakis and Rubin methodology assumes that a patient has an underlying latent trait that defines their compliance to the treatment regime under each possible treatment assignment. If outcomes are missing-at-random within compliance groups, the MOM estimator yields consistent estimates of the causal effect of treatment assignment. In this talk, we develop a parametric counterpart based on the EM-algorithm to the MOM estimator. We demonstrate that the EM estimator yields more efficient estimates on both normal and non-normal data, but is slightly more biased in the latter case. If time permits, the EM estimator will be extended to trials with longitudinal outcomes. The methods will be applied to clinical trial data in Psychiatry, a field where non-compliance with treatment protocol is the norm rather than the exception.

e-mail: omalley@hcp.med.harvard.edu

15. Environmental Statistics

SPATIAL ASSOCIATION BETWEEN SPECIATED FINE PARTICLES AND HUMAN HEALTH EFFECTS SPECIATED FINE PARTICLES AND HUMAN HEALTH EFFECTS

Montserrat Fuentes*, North Carolina State University
Hae-Ryoung Song, North Carolina State University
Sujit Ghosh, North Carolina State University
David Holland, U.S. EPA

Particulate matter (PM) has been linked to a range of serious cardiovascular and respiratory health problems. Some of the recent epidemiologic studies suggest that exposures to PM may result in tens of thousands of excess deaths per year, and many more cases of illness among the US population. The main objective of our research is to quantify uncertainties about the impacts of fine PM exposure on mortality. We develop a multivariate spatial regression model for better estimation of the mortality effects from fine PM and its components across the coterminous US. Our approach adjusts for meteorology and other confounding influences, such as socioeconomic factors, age, gender and ethnicity, characterizes different sources of uncertainty of the data, and models the spatial structure of several components of fine PM. We consider a flexible Bayesian hierarchical model for a space-time series of (mortality) counts by constructing a likelihood based version of a generalized Poisson regression model. The model has the advantage of incorporating both over and under dispersion in addition to correlations that occur in space and time. We apply these methods to daily mortality county counts, measurements of total and several components of fine PM from national monitoring networks in the US, and the output of deterministic air quality models.

e-mail: fuentes@stat.ncsu.edu

LONGITUDINAL AND SPATIAL ANALYSIS OF SHORT-TERM RESPIRATORY HEALTH EFFECTS OF AIR POLLUTION

Paul Rathouz*, University of Chicago
Vanja Dukic, Dept. of Health Studies and CISES
Dana Draghicescu, CISES
Edward Naureckas, Dept. of Medicine and CISES
Xiaoming Bao, CISES
John Frederick, Dept. of Geophysical Sciences and CISES
Alexis Zubrow, CISES

Most recent studies of environmental effects on respiratory diseases have focused on aggregate-data time series analyses, with outcomes and exposure variables summarized by daily averages over an entire region of interest. The work we present here uses longitudinal person-level asthma-related health outcomes, and air quality and other exposure measures aggregated at the ZIP code level (or other fixed areal unit). The goal is to identify linkages between acute respiratory disease outcomes, such as asthma-prescription refills or ED visits, and measures of air quality at the ZIP code level, while exploiting the longitudinal data structure to increase statistical efficiency and to control for person-level confounding factors. We approach this problem via a case-crossover analysis for the person-level data, coupled with a model for underlying spatial variability in disease severity and disease sensitivity to fluctuations in air pollution. The methods are applied to a data set of Chicago Medicaid asthma sufferers over a three-year period from 1995—98. This research is supported by EPA grant R-82940201-0. The abstract does not necessarily reflect EPA views.

e-mail: prathouz@health.bsd.uchicago.edu

FUNCTIONAL DATA ANALYSIS METHODS IN THE ENVIRONMENTAL SCIENCES

Wendy Meiring*, University of California, Santa Barbara

Functional data are observations from curves or surfaces. For example a balloon-based ozonesonde measures the ozone partial pressure profile as the sonde ascends through the atmosphere. Each ozonesonde flight provides observations from one ozone partial pressure profile as a function of altitude. A sequence of sonde launches provides observations from a time series of ozone profile ‘curves’, with each curve a function of altitude. The shape of these curves evolves over time in response to complex dynamical and chemical processes. We present functional data analysis methodology to estimate altitude-dependent non-linear time trends and other space-time modes of variability in observations from a time series of ozonesonde flights. Due to the large number of observations, we combine dimension reducing functional basis approximations, with flexible additive models on the low-dimensional basis function coefficient scale. We fit additive coefficient models composed of cubic and periodic splines, as special cases of smoothing spline anova models (SSANOVA) and discuss extensions. We provide preliminary estimates of uncertainty of altitude dependent non-linear time trends and Quasi-Biennial Oscillation effects, building on SSANOVA Bayesian confidence intervals. We discuss continued improvements in uncertainty measures.

e-mail: meiring@pstat.ucsb.edu

SEPARABLE APPROXIMATIONS OF SPACE-TIME COVARIANCES

Marc G. Genton*, North Carolina State University

Statistical modeling of space-time data has often been based on separable covariance functions, that is covariances that can be written as a product of a purely spatial covariance and a purely temporal covariance. The main reason is that the structure of separable covariances dramatically reduces the number of parameters in the covariance matrix and thus facilitates computational procedures for large space-time data sets. In this talk, we discuss separable approximations of space-time covariances. In particular, we describe the nearest (in the Froebinius norm) Kronecker product approximation of a space-time covariance matrix. The algorithm is simple to implement and preserves symmetry and positive definiteness of the solution. The separable approximation allows for fast kriging of large space-time data sets. We present several illustrative examples and an application to the Irish wind speed data.

e-mail: genton@stat.ncsu.edu

16. Addressing the Looming Shortage of Statisticians

THE POTENTIAL FOR AP STATISTICS TO LEAD TO MORE STATISTICIANS

Linda J. Young*, University of Florida

Since the first AP Statistics Exam was offered to 7,500 students in 1997, the program has experienced rapid growth with more than 60,000 students anticipated to take the exam in 2004. The statistical background of the student who does well on the AP Statistics will be briefly reviewed. Opportunities to recruit these students to the profession will be discussed. The impact this influx of students should have on undergraduate education will be addressed.

e-mail: LYoung@biostat.ufl.edu

TWO PROGRAMS TO ATTRACT STUDENTS TO STATISTICS/BIOSTATISTICS

Dennis D. Boos*, North Carolina State University

This talk will focus on two efforts for recruiting students to graduate study in statistics and biostatistics: 1. StatFest and 2. The Summer Institute for Training in Biostatistics (SIBS). StatFest refers to a series of one-day conferences sponsored by the ASA Committee on Minorities in Statistics (and others) to attract minority students to graduate school and ultimately to careers in statistics. These conferences feature panels of statisticians from industry, government, and academia, and also panels of current graduate students and directors of graduate programs. SIBS is a new program funded by the National Heart, Lung and Blood Institute (NHLBI) to attract quantitatively-oriented undergraduates to graduate study in biostatistics. Basically, the demand for biostatisticians is growing, and this program seeks to remedy the supply side of the equation by teaching highly qualified undergraduates some basics of biostatistics and the role it plays in medical discoveries. Starting in the summer of 2004, there will be 6 week training programs at Boston University, North Carolina State University (jointly with the Duke Clinical Research Institute), and The University of Wisconsin. This talk will give an interim report on the progress of these summer programs.

e-mail: boos@stat.ncsu.edu

CREATING AND SUSTAINING A DYNAMIC UNDERGRADUATE STATISTICS PROGRAM

Bruce J. Collings*, Brigham Young University

The Department of Statistics at Brigham Young University was created in 1960 with no students and one faculty member. The first student graduated with a BS in Statistics in 1962, at which time the department had three faculty members, plus two part-time faculty from the College of Business. By the mid 1970s the department had about a dozen faculty and about twenty each of graduate students (MS) and undergraduate students. By the mid 1980s there were about fifteen faculty, twenty graduate students and fifty undergraduates. The number of undergraduates increased to about 100 by 1990 and to about 180 at present. While some of this growth can be attributed to fortuitous circumstances, much of it can be attributed to positive steps taken by the Department. This presentation will discuss some of these steps, including recruitment, retention and program evolution.

e-mail: collings@byu.edu

WHERE ARE WE AND WHAT IS NIH DOING ABOUT IT?

Dennis O. Dixon*, NIH, NIAID

Many are convinced there is a serious shortage of well-trained biostatisticians, a conviction supported by a modest amount of empirical evidence that will be summarized in the presentation. The National Institutes of Health and its component institutes and centers recognize how critically important it is to create a larger workforce with advanced training in biostatistics and related fields. In fact, the recently announced NIH Roadmap for discovery and translation of new knowledge in the prevention, detection, diagnosis and treatment of disease cannot succeed without sufficient participation by biostatisticians, bioinformaticists, modelers, and others. This talk will highlight new and standing programs that aim to respond to these needs. Among the new programs are summer institutes to encourage undergraduate mathematics students to consider careers in biostatistics (created by the National Heart, Lung and Blood Institute) and multidisciplinary clinical research career development programs aimed at those already holding doctoral degrees. There is plenty of room for other initiatives, even those that do not need large amounts of funding.

e-mail: dd23a@nih.gov

17. Survival Methods with Random Effects, Noncompliance, or Longitudinal Data

SEMPARAMETRIC LINEAR TRANSFORMATION MODELS WITH RANDOM EFFECTS FOR CLUSTERED FAILURE TIME DATA

Donglin Zeng*, University of North Carolina
Danyu Lin, University of North Carolina
Xihong Lin, University of Michigan

We propose a class of semiparametric linear transformation models with random effects for the regression analysis of clustered or correlated failure time data. This class of models relates an unknown transformation of the failure time linearly to the covariates and random effects, and includes semiparametric proportional hazards and proportional odds models with random effects as special cases. We identify a set of sufficient conditions on the transformation and the distribution of the random effects such that the nonparametric maximum likelihood estimation achieves the semiparametric efficiency bound. We develop the corresponding likelihood-based inference procedures. Simulation studies demonstrate that the proposed methods perform well in practical situations. An illustration with a well-known diabetic retinopathy study is provided.

e-mail: dzeng@bios.unc.edu

A CORRECTED PSEUDO-SCORE APPROACH FOR ADDITIVE HAZARDS MODEL WITH LONGITUDINAL COVARIATES MEASURED WITH ERROR

Xiao Song*, University of Washington
Yijian Huang, Fred Hutchinson Cancer Research Center

In medical studies, it is often of interest to characterize the relationship between a time-to-event and covariates, not only time-independent but also time-dependent. Time-dependent covariates are generally measured intermittently and with error. Recent interests focus on the proportional hazards framework, with longitudinal data jointly modeled through a mixed effects model. However, approaches under this framework depend on the normality assumption of the error, and might encounter intractable numerical difficulties in practice. This motivates us to consider an alternative framework, that is, the additive hazards model, under which little has been done when time-dependent covariates are measured with error. We propose a simple corrected pseudo-score approach for the regression parameters with no assumptions on the distribution of the random effects and the error beyond those for the variance structure of the latter. The estimator has an explicit form and is shown to be consistent and asymptotically normal. We illustrate the method via simulations and by application to data from an HIV clinical trial.

e-mail: songx@u.washington.edu

BIVARIATE TRUNCATED FAILURE TIME DATA

Rinku Sutradhar*, University of Waterloo

Richard J. Cook, University of Waterloo

Incomplete failure time data often arise when observations are only available if they are in pre-specified truncation intervals. Data of this nature raise a variety of challenges. Turnbull (1976) developed a self-consistency algorithm for non-parametric estimation of the failure time distribution for truncated data. Our focus is on problems involving bivariate truncated failure time data where interest lies in the degree of association between paired failure times. For this problem one can use either the conditional (random effect) or marginal approaches for estimation and inference (Clayton, 1978). The former model treats the association of the failure times as arising due to shared unknown factors, whereas the latter model considers the joint distribution as a function of the marginal distributions and association parameters. We present a likelihood approach for regression based on a conditional formulation using parametric and piecewise constant baseline hazard functions. Results from simulation studies are reported. The proposed methods are illustrated using a motivating data set on age of first hospitalization for mental illness for siblings where these hospitalizations were restricted to lie between 1926 and 1943.

e-mail: rsutradh@math.uwaterloo.ca

NONPARAMETRIC ESTIMATION OF THE BIVARIATE RECURRENCE TIME DISTRIBUTION

Chiung-Yu Huang*, University of Minnesota

Mei-Cheng Wang, Johns Hopkins University

This paper considers statistical models in which two different types of events, such as the diagnosis of a disease and the remission of the disease, occur alternately over time and are observed subject to right censoring. We propose nonparametric estimators for the joint distribution of bivariate recurrence times and the marginal distribution of the first recurrence time. In general, the marginal distribution of the second recurrence time cannot be estimated due to an identifiability problem, but a conditional distribution of the second recurrence time can be estimated nonparametrically. In the literature, statistical methods have been developed to estimate the joint distribution of bivariate recurrence times based on data on the first pair of censored bivariate recurrence times. These methods are inefficient in the model considered here because recurrence times of higher orders are not used. Asymptotic properties of the estimators are established. Numerical studies demonstrate the estimator performs well with practical sample sizes. We apply the proposed method to a Denmark psychiatric case register data set for illustration of the methods and theory.

e-mail: cyhuang@biostat.umn.edu

A GENERALIZED RESIDUAL FOR MULTIVARIATE SURVIVAL AND FRAILTY MODELS

Kirby L. Jackson*, University of South Carolina

A generalized survival residual appropriate for frailty models is proposed. This residual generalizes the usual Cox-Snell residuals to the case of correlated survival data and can be used in the investigation of a variety of frailty models now in use. This residual is shown to have a multivariate exponential distribution with marginal the unit exponential and association parameters based upon the parameters of the original frailty model. The two most common frailty models, those based on either the gamma distribution or the positive stable distribution are relatively easy to examine using these residuals. The distribution of this residual is given in terms of Laplace transforms of functions of the observed survival times and possible measures of association based on this model are discussed. The bivariate version of this distribution can also be generated as an Archimedean copula. Smoothed goodness of fit tests based on this residual can be derived and investigation of the appropriateness of the frailty models examined. This residual may have major implications for the analysis of frailty assumptions and definition and estimation of dependence parameters in multivariate survival models.

kjackson@sc.edu

AN IMPROVED SMALL-SAMPLE VARIANCE ESTIMATOR FOR THE ANALYSIS OF CLUSTERED SURVIVAL DATA

Rong Zhou*, Medpace, Inc.
Jorge G. Morel, P&G Pharmaceuticals, Inc

When analyzing survival data with correlated/clustered events per subject, the intra-subject correlation should be properly taken into account in the analysis. The jackknife method (Lin and Wei, 1989) provides a robust estimator of the variance for correlated data when the assumption of independence for the ordinary Cox's model (proportional hazard) is not satisfied. In small samples, however, the jackknife estimator of the variance leads to inflated Type I error rates. We propose an adjusted robust method that reduces the small-sample bias in the Type I error rates. The adjustment is easy to implement while preserving the asymptotic properties of the robust jackknife variance estimator. The effectiveness of this correction is shown via simulation using correlated survival data. Two examples with correlated survival data are analyzed to demonstrate how this adjustment is a useful addition to current methodology. One example is from a satiety study where "time to hunger" is the efficacy variable, and the other is from a skin safety study where "time to irritation" is the main endpoint.

e-mail: r.zhou@medpace.com

STRUCTURAL PROPORTIONAL HAZARDS MODELS FOR CAUSAL INFERENCE IN CLINICAL TRIALS

Els JT Goetghebeur*, Ghent University
Tom Loeys, Harvard School of Public Health

Deviations from assigned treatment occur often in clinical trials. In such a setting, the traditional intent-to-treat analysis does not measure biological efficacy but rather programmatic effectiveness. For the all-or-nothing compliance situation, Loeys and Goetghebeur (2003) proposed a Structural Proportional Hazards method. It allows for causal estimation in the complier subpopulation provided the exclusion restriction holds: randomization per se has no effect unless exposure has changed. This assumption is typically made with structural models for noncompliance but questioned when the trial is not blinded. In this paper we extend the structural PH model to allow for an effect of randomization per se. This enables analyzing sensitivity of conclusions to deviations from the exclusion restriction. In a colo-rectal cancer trial we find the causal estimator of the effect of an arterial device implantation to be remarkably insensitive to such deviations.

e-mail: els.goetghebeur@ugent.be

18. Epidemiologic Methods

MULTIPOINT LINKAGE DISEQUILIBRIUM MAPPING USING CASE-CONTROL DESIGNS

Kung-Yee Liang, National Health Research Institutes
Yen-Feng Chiu*, National Health Research Institutes

Case-control study has been and continues to be one of the most popular designs in epidemiology. More recently, this design has been adopted to test for candidate genes when searching for disease genetic etiology. In this paper, we present a multipoint linkage disequilibrium (LD) mapping approach with the focus on estimating the location of the target trait locus. It builds upon a representation, which shows that the difference between a case and a control in probabilities of carrying the target allele of a marker is proportional to that of the trait locus and that the proportionality factor is simply a measure of LD between the trait locus and the marker. Our method has the desired properties that (i) it is robust in that no assumption on the underlying mechanism (e.g., single locus) is needed, (ii) it provides an estimate of the disease locus along with sampling uncertainty to help investigators to narrow chromosomal regions and (iii) a single test statistic is provided to test for LD in the framed region rather than testing the hypothesis one marker at a time. Our simulation work suggests that the proposed method performs well in terms of bias and coverage probability. Extension of the proposed method to account for confounding and genetic heterogeneity is discussed.

e-mail: yfchiu@nhri.org.tw

CASE-CONTROL STUDIES OF GEN-ENVIRONMENT INTERACTIONS WITH MISSING DATA

Christine Spinka*, Texas A&M University
Nilanjan Chatterjee, National Cancer Institute
Raymond J. Carroll, Texas A&M University

The risks of developing many complex diseases are believed to be determined by the joint effects of genetic factors and environmental exposures. We address the problem of maximum likelihood estimation of the gene-environment association with disease in case-control studies. We assume that in the underlying population, the distributions of the genetic factors and the environmental covariates are independent. We develop a semiparametric maximum likelihood estimation procedure to estimate the logistic regression parameters that utilizes the gene-environment independence assumption, accounts for the possibility of missing genetic data, and treats the distribution of the environmental covariates nonparametrically. A simple estimation procedure and asymptotic theory are also studied using the profile likelihood. The method is illustrated using simulations and is applied to a real data set exploring the interaction between diplotype and environment in disease risk.

cspinka@stat.tamu.edu

CASE-COHORT DESIGNS AND ANALYSIS FOR CLUSTERED FAILURE TIME DATA

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

Case-cohort design is an efficient and economical design to study risk factors for infrequent disease in a large cohort. It involves the collection of covariate data from all failures ascertained throughout the entire cohort, and from the members of a random subcohort selected at the onset of follow-up. In the literature, the case-cohort design has been extensively studied, but was exclusively considered for univariate failure time data. In this paper, we propose case-cohort designs adapted to multivariate failure time data. An estimation procedure with the independence working model approach is used to estimate the regression parameters in the marginal proportional hazards model, where the correlation structure between individuals within a cluster is left unspecified. Statistical properties of the proposed estimators are developed. The performance of the proposed estimators and comparisons of statistical efficiencies are investigated with simulation studies. A data example is used to illustrate the proposed methodology.

e-mail: lus2@umdnj.edu

SEMI-PARAMETRIC SURVIVAL ANALYSIS OF CASE-CONTROL FAMILY DATA

Lu Chen*, University of Washington
Li Hsu, Fred Hutchinson Cancer Research Center

Case-control family data consists of information on families sampled based on cases and controls. The retrospective sampling of the families and the dependence among family members beyond measured covariates present unique challenges for analyzing such data. Moreover, some important covariates, such as genotype for known breast cancer susceptible genes BRCA1/2, might only be collected for the cases and the controls but not for the relatives. We use age at breast cancer diagnosis as the outcome variable and work within the framework of Cox model and frailty model. I will present semi-parametric methods for the simultaneous estimation of regression parameters and any residual dependence beyond measured covariates. I will also discuss approaches for extending these methods to situation with missing genotype in relatives. I'll present results from a simulation study on the performance of these methods. Finally I'll illustrate the method on a case-control family data of breast cancer.

e-mail: luchen@u.washington.edu

A COMPARISON OF METHODS USED TO ASSESS CONFOUNDING AND MEDIATION IN LOGISTIC REGRESSION

Lorraine G. Ogden*, University of Colorado
Leann Myers, Tulane University

Methods for assessing the difference between a crude and adjusted parameter estimate in logistic regression have evolved in two different disciplines with distinct purposes: in epidemiology as methods to assess confounding and in the social sciences as statistical tests for mediation. We will first distinguish between the classical and operational definitions of confounding and discuss approaches for assessing confounding which have emerged from these definitions. We will then consider statistical tests for mediation and discuss their applicability as statistical tests for confounding. Finally, we examine the performance of these methods using simulated data under various scenarios (different sample sizes and with the presence/absence of classical confounders, operational confounders, and mediators) and make recommendations on the most appropriate methods given the sample size and the primary objective of the analysis.

e-mail: lorri.ogden@uchsc.edu

REGRESSION ADJUSTMENTS TO ESTIMATORS OF AVERAGE CAUSAL EFFECTS THAT USE STRATIFICATION OR WEIGHTING VIA THE PROPENSITY SCORE

Jared K. Lunceford*, Merck Research Laboratories
Marie Davidian, North Carolina State University

With observational data, treatment exposure may be associated with covariates that are also associated with potential response, and groups may be seriously imbalanced in these factors leading to biased causal inferences in the absence of adjustment. Two popular approaches to correcting for these imbalances use stratification and weighting based on estimated propensity scores. Adding standard regression adjustments to both the stratified or weighted methods offers additional improvement in performance. We describe theoretical properties of these regression-adjusted methods, highlight their implications for practice, and present comparisons of performance that provide guidance for practical use.

e-mail: jared_lunceford@merck.com

MULTIVARIATE GRAPHICS FOR FDA OF HORMONE PATTERNS IN THE MENOPAUSAL TRANSITION

Peter M. Meyer*, Rush University Medical Center
Judith L. Luborsky, Rush University Medical Center
Sioban Harlow, University of Michigan
Daniel S. McConnell, University of Michigan
Imke Janssen, Rush University Medical Center

Female hormones follow typical patterns of cyclical variability during premenopausal years. Functional Data Analysis (FDA) techniques provide a convenient framework in which to describe the patterns reflected by hormone levels assayed from daily urine samples. During the menopausal transition hormone patterns deviate from standard patterns and move toward other patterns more typical of post-menopause. Using a dataset from the Studies in Women's Health Across the Nation (SWAN) containing four different hormone values assayed from daily urines for 280 completed cycles of premenopausal and early-perimenopausal women, we present a variety of FDA multivariate graphical techniques for displaying typical premenopausal patterns and the variations from those patterns that are common during the menopausal transition.

e-mail: pmeyer@rush.edu

19. Linkage Disequilibrium and QTL Analysis

HIGH RESOLUTION LD MAPPING OF QTL BY MULTIPLE MARKERS

Jeesun Jung, Texas A&M University
Ruzong Fan*, Texas A&M University

Using multiple bi-allelic markers, regression models are proposed to fine map quantitative trait loci (QTL) using linkage disequilibrium (LD) analysis. The objective is to fully use marker information for fine mapping of QTL. In the model, the genetic effect is decomposed into orthogonal additive and dominant effects. Based on analytical formula, the non-centrality parameters of test statistics are calculated to test the genetic effects and linkage disequilibriums. By power and sample size comparison, the merits of the proposed method are investigated. It can be shown that the multiple marker LD analysis has high power in fine mapping QTL.

e-mail: rfan@stat.tamu.edu

TAGGING CANDIDATE GENES WITH SPECTRAL DECOMPOSITION

Zhaoling Meng*, Merck & Co., Inc.
Dmitri V. Zaykin, GlaxoSmithKline

Association study with high density markers typed within candidate genes is currently considered as one of most promising approaches for disease gene mapping. Single nucleotide polymorphisms (SNPs) are the most frequently utilized markers because of their high density and wide spread occurrence in human genome. Success of locating disease genes relies on high density of SNPs so that correlations between the unobserved alleles at a disease gene and alleles of nearby SNPs is sufficiently high. The expected value of this correlation, or linkage disequilibrium (LD), increases with the proximity of the disease gene and an SNP. In principle, this allows using SNPs as proxies for the alleles of the disease gene. However, LD among SNP markers themselves implies that SNPs in high LD with each other carry somewhat redundant information. This may incur unnecessary genotyping burden without proportional increase in the information gain. Therefore, genotyping cost can be sensibly reduced by selecting a subset of SNPs based on LD from a larger set typed on a relatively small set of samples. Only this informative subset may later be retyped in a larger sample of individuals. We apply a method based on the spectral decomposition of the marker pairwise LD matrix of the candidate gene and select non-redundant markers to balance information gain and genotyping cost. We test our method on genes with different LD patterns, with different numbers of SNPs typed and data with different sample sizes. Furthermore, we compare our method to an existing two-step approach that defines high LD regions (so called haplotype blocks) using D' and selects tagging SNPs within each block.

e-mail: zhaoling_meng@merck.com

FINE-SCALE MAPPING OF DISEASE GENES BASED ON LINKAGE DISEQUILIBRIUM

Na Li*, University of Minnesota
Matthew Stephens, University of Washington

Genetic mapping is based on the fundamental idea that the probability of recombination between markers is higher for markers that are further apart, and lower for those that are closer. When markers are very close, the probability of observing any recombination within a family (even a large one) becomes extremely small, which limits the resolution of pedigree based linkage analysis. For fine-scale localization, a population based approach (e.g., a case-control design) is not only more powerful but sometimes necessary. Apparently unrelated individuals are in fact related through distant co-ancestry and the effect of recombinations through many generations is reflected on the patterns of linkage disequilibrium (LD, association of marker alleles at different loci in the population) in the current generation. Current fine-mapping approaches based on this concept have several limitations. In particular, they do not properly take into account the background LD among the controls. Here we introduce a new method based on the general model for LD, Product of Approximate Conditionals (PAC) model (Li and Stephens, 2003). Simulations and real data (if available) are used to evaluate its performance.

e-mail: nali@umn.edu

GENETIC ASSOCIATION MAPPING UNDER FOUNDER HETEROGENEITY VIA HAPLOTYPE SIMILARITY ANALYSIS IN CANDIDATE GENES

Kai Yu*, Washington University
Charles Gu, Washington University
Mike Province, Washington University
Chengjie Xiong, Washington University
D.C. Rao, Washington University

Taking advantage of increasingly available high-density SNP maps within genes and across genome, several methods were proposed for candidate gene association studies using multiple closely linked SNP markers. One practical challenge faced by existing methods is founder heterogeneity, that is, not all case chromosomes from affected individuals inherited a disease-susceptibility mutation from the same ancestral chromosome. Assuming the existence of a common founder from which a relatively large proportion of case chromosomes is derived, we developed a procedure to remove case chromosomes not likely descending from the common founder so the effect of founder heterogeneity is reduced. This allows us to identify nested subsets of case chromosomes that are most likely to share a common ancestral chromosome. Association between the candidate region and the disease variant is then evaluated by comparing haplotype similarity among chromosomes in each identified subset to that among controls. The overall statistical significant level is accessed using a permutation procedure. Simulation studies suggest that this approach has the right Type I error rate, and improved power over existing methods.

e-mail: kai@wubios.wustl.edu

A UNIFYING MODEL FOR FINE MAPPING OF QUANTITATIVE TRAIT LOCI AFFECTING CD4+ T CELL AND HIV-1 DYNAMICS UNDER HIGHLY ACTIVE ANTIRETROVIRAL THERAPY

Zuoheng Wang*, University of Florida
Rongling Wu, University of Florida

The identification of specific genes (i.e., quantitative trait loci or QTL) that, carried by hosts, control dynamic changes of CD4+ T cells and HIV-1 viral loads is of utmost importance for designing personalized drugs for curbing AIDS. In this article, we present a unifying statistical model for simultaneous mapping of QTL underlying CD4+ T cell and HIV-1 dynamics based on the epidemiological property of the disease. We assume that each of the two dynamic processes is controlled by one or more QTL associated with a haplotype block comprised of single nucleotide polymorphisms (SNPs). We devise an EM-simplex hybrid algorithm to estimate the population frequencies of QTL alleles, marker-QTL linkage disequilibria, and QTL effects reflected by dynamic parameters of curves. We formulate two important hypothesis tests to test whether a QTL pleiotropically affects HIV-1 and CD4+ T cell dynamics and what is the relative contribution of pleiotropic effects and close association to the correlations between these two processes. Our model also allows for tests of a number of clinically important variables in an HIV/AIDS trial. The statistical properties of our model are examined through simulation studies and its clinical implications are discussed.

e-mail: zwang2@stat.ufl.edu

A STATISTICAL GENETIC MODEL FOR INTEGRATING GROWTH AND DEVELOPMENT

Min Lin*, University of Florida
Rongling Wu, University of Florida

In biology, there are two fundamentally important questions: (1) how a given amount of energy is allocated between vegetative growth and sexual reproduction to maximize the fitness of organisms? (2) Are there specific genes or quantitative trait loci (QTL) that control the integration of growth and development during ontogeny? The identification of QTL for growth process is statistically challenged due to its time series or longitudinal features. In this article, we present a new statistical model for functional mapping of QTL that affect growth trajectories by modeling the mean vector and (co)variance structure based on universal biological principles. By incorporating the genetic machinery of reproductive behaviors into the functional mapping framework, this model is extended to detect specific QTL that govern both growth and developmental processes through either pleiotropic effects or close linkage, or both. Our model is derived within the maximum likelihood context, implemented with an EM-simplex hybrid algorithm. An example from a forest tree was used to demonstrate the usefulness of our model. The implications of this model for integrating vegetative growth and sexual reproduction to gain better insights into comprehensive biology are discussed.

e-mail: mmin@stat.ufl.edu

20. Large Scale Multiple Testing

A BAYESIAN PERSPECTIVE ON FALSE DISCOVERY AND FALSE NON-DISCOVERY RATES

Jie Chen*, Merck Research Laboratories
Sanat K. Sarkar, Temple University

The classical definition of the FDR can be understood as the frequentist risk of false rejections conditional on the unknown parameter θ . From Bayes decision-theoretic point of view, it seems natural to take into account the uncertainty in both parameter and data. In this spirit, we propose the Average FDR (AFDR) and Average FNR (AFNR) approaches in which the frequentist risks of false rejections and false non-rejections are averaged out with respect to some prior distribution of parameter θ . A linear combination of the AFDR and AFNR, called the Average Bayes Error Rate (ABER), is considered as the overall risk of misclassification. Finally, the AFDR and AFNR are illustrated in an example where objective and subjective Bayesian hypotheses are formulated. Some useful formulas for the AFDR and AFNR are further developed for normal samples with normal priors. Critical values for the data at which the ABER is minimized are presented.

e-mail: jie_chen@merck.com 21

IMPROVING ESTIMATION OF THE FALSE DISCOVERY RATE

Stan B. Pounds*, St. Jude Children's Research Hospital
Cheng Cheng, St. Jude Children's Research Hospital

Microarray technology allows investigators to simultaneously measure the expressions of thousands of genes. These data are often analyzed by testing a given null hypothesis for each gene measured. The false discovery rate (FDR), which is roughly the proportion of significant findings that arise solely by chance, is a very practical way to account for the massive multiplicity encountered in this area. Several methods consider the p-values obtained for each gene's differential expression as a sample from a mixture distribution consisting of a null component and an alternative component. These methods estimate the FDR as a function of the p-value by estimating and partitioning (EP) the p-value distribution. An EP algorithm is proposed that uses loess to smooth a set of transformed spacings of p-values to construct an estimate of the probability density function (PDF). This PDF estimator is called a spacings loess histogram (SPLOSH). An example and a simulation study reveal that SPLOSH produces more reliable FDR estimates than other EP methods, including Benjamini and Hochberg's simple slope method, Storey's q-value, and Pounds and Morris' beta-uniform mixture model.

e-mail: stanley.pounds@stjude.org

AN EMPIRICAL BAYES ADJUSTMENT TO MULTIPLE P-VALUES

Somnath Datta, University of Georgia
Susmita Datta*, Georgia State University

In recent microarray experiments thousands of gene expressions are simultaneously tested in comparing samples (e.g., tissue types). Application of a statistical test (e.g. t-test) would result a p-value for each gene reflecting the amount of statistical evidence present in the data that the given gene is differentially expressed. We show how to use these p-values across the genes using the method of empirical Bayes estimation so that each gene borrows evidence of differential expression (or nondifferential expression, whatever the case may be) of all others on the microarray. A new set of accept/reject decisions are reached for the differential expressions using the adjusted p-values through a resampling based step-down p-value calculation that protects the analyst against the overall (familywise) type 1 error rate. The utility of the empirical Bayes adjustment is illustrated through simulation experiments by various performance measures such as sensitivity, specificity, false discovery rate and false non-discovery rate of the overall testing mechanism with and without the empirical Bayes adjustment.

e-mail: sdatta@mathstat.gsu.edu

NEW PROCEDURES FOR CONTROLLING FALSE DISCOVERY RATE

Zhongfa Zhang*, Case Western Reserve University
Jiayang Sun, Case Western Reserve University

Multiple testing procedures are useful for identifying disease genes and for locating sources of activation in a brain image. In this paper, we present a general theoretical result about False Discovery Rate (FDR), explore a functional relationship between the family-wise error rate (FWER) and FDR, and develop two new procedures for controlling FRD. Our new procedures are compared with Storey et al's (2003) procedure. Confidence bands around the expected FDR are provided. Some practical implications of the general theory are illustrated. An application to leukemia data is also demonstrated.

e-mail: zxz8@po.cwru.edu

A MIXTURE MODEL FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN DNA MICROARRAY

Jiangang Liao*, UMDNJ
Yong Lin, Cancer Institute of New Jersey
Zachariah Selvanayagam, University of Medicine and Dentistry of New Jersey
Joe Shih, University of Medicine and Dentistry of New Jersey

DNA microarray is often used to identify genes that express differentially between two types of tissues. With the simultaneous comparisons on thousands of genes, however, adjusting for multiple testing becomes critically important. This paper derives a flexible and stable mixture model for adjusting for multiple testing by requiring the distribution of p-values for differentially expressed genes to be stochastically smaller than the distribution of p-values for non-differentially expressed genes. The proposed model provides a unified framework for estimating the proportion of differentially expressed genes, the local positive discovery rate, positive false discovery rate or false discovery rate as well as other quantities of interest such as the positive and negative predictive values and the sensitivity and specificity. An R function implementing the proposed method is provided on the web and can be readily used for data analysis. **KEY WORDS:** False discovery rate; Proportional hazard model; Smoothing; Stochastic order.

e-mail: jg_liao@yahoo.com

ANALYSIS OF MULTIFACTOR DESIGNS WITH LARGE NUMBER OF FACTOR LEVELS

Haiyan Wang*, Penn. State Univerisity
Michael G. Akritas, Penn. State Univerisity

Testing methods for ANOVA designs where some of the factors have a large number of levels have received a lot of attention recently (Boos and Brownie 1995, Akritas and Arnold 2000, Bathke 2002, Akritas and Papadatos 2003, Wang and Akritas 2002). Most attention, however, has been restricted to procedures using the original observations, either in the balanced homoscedastic case or with no more than two factors. In this paper we consider hypotheses testing in possibly unbalanced and heteroscedastic multi-factor designs. Such procedures require strong moment assumptions and are sensitive to outliers. Thus we also develop (mid)-rank procedures for the same general setting. The main asymptotic tools are the asymptotic rank transform and H₀'s projection method. Simulation results show that the present rank statistics outperform those based on the original observations, in terms of both Type I and Type II errors. A real data set from a microarray experiment is analyzed with the suggested procedures.

e-mail: hwang@stat.psu.edu

SHARP SIMULTANEOUS INTERVALS FOR THE MEANS OF SELECTED POPULATIONS WITH APPLICATION TO MICROARRAY DATA ANALYSIS

Jing Qiu*, Cornell University
J. T. Gene Hwang, Cornell University

Simultaneous inference is a challenge when the number of populations is large. In some situations including microarray experiments, the scientists are particularly interested in the populations with the most extreme observations. In these cases, we can just focus on a few (K) selected observations and do inference about the corresponding populations. An Empirical Bayes approach is used to construct confidence intervals for the K selected means with good coverage probabilities. K can be prechosen or estimated from data. The resultant intervals have length comparable to that of K -dimensional simultaneous intervals, which are shorter (and could be much shorter) than the traditional N -dimensional Bonferroni's bounds. This saving together with the borrowing strength effect of Empirical Bayes results in very sharp simultaneous confidence intervals. These confidence intervals can also be used to conduct hypotheses testing. In one application to microarray data, we show that they detect 17% more significant genes than the Bonferroni's simultaneous bounds.

e-mail: jq14@cornell.edu

21. Causal Inference

CAUSAL INFERENCE IN HYBRID INTERVENTION TRIALS INVOLVING TREATMENT CHOICE

Qi Long*, University of Michigan
Roderick J. Little, University of Michigan
Xihong Lin, University of Michigan

Randomized allocation of treatments is a cornerstone of experimental design, but has drawbacks when a limited set of individuals are willing to be randomized, or the act of randomization undermines the success of the treatment. Choice-based experimental designs allow a subset of the participants to choose their treatments. We discuss here causal inferences for experimental designs where some participants are randomly allocated to treatments and others receive their treatment preference. This paper was motivated by the "Women Take Pride" (WTP) study (Janevic et al., 2001), a doubly randomized preference trial (DRPT) to assess behavioral interventions for women with heart disease. We propose a model for inference about preference effects for a DRPT and develop an EM algorithm to compute maximum likelihood estimates of the model parameters. The method is illustrated by analyzing compliance of the WTP data. Our results show that there were strong preference effects in the WTP study, that is, women assigned to their preferred treatment were more likely to comply.

e-mail: qlong@umich.edu

EXAMINING SENSITIVITY TO THE EXCLUSION RESTRICTION FOR INSTRUMENTAL VARIABLES ESTIMATORS OF TREATMENT EFFECT

Joseph W. Hogan*, Brown University
Tony Lancaster, Brown University
Jason Roy, University of Rochester
Donald Alderson, Brown University

The method of instrumental variables can be used to obtain consistent estimates of a causal treatment effect from observational data. Success of the method relies on using a valid instrument, defined as a variable that is correlated with receipt of treatment, and conditional on treatment, uncorrelated with underlying potential outcomes. The second requirement is sometimes called an ‘exclusion restriction’, and cannot be checked empirically. In this paper we describe a method for assessing the range of large-sample bias in the estimated treatment effect when the exclusion restriction is violated. We focus on the simple case of binary treatment and binary instrument, which illuminates some key points related to whether one assumes homogeneous or heterogeneous treatment effects. The sensitivity parameters have physical interpretations on the scale of the outcome variable (e.g. differences in mean CD4 for certain subpopulations), which provides contextual intuition about the exclusion restriction and facilitates clear discussion with subject matter experts about whether the assumption is met. Our method is used to quantify a range for the causal effect of receiving antiviral therapy, using data from a longitudinal cohort study of HIV-infected women.

e-mail: jhogan@brown.edu

MODIFICATIONS TO MARGINAL STRUCTURAL MODELS AND ASSOCIATED WEIGHTED ESTIMATION

Marshall M. Joffe*, University of Pennsylvania
Jill Knauss, University of Pennsylvania
Bruce Robinson, University of Pennsylvania
Harold I. Feldman, University of Pennsylvania

Hernan et al. (2000) recently proposed a new class of models for estimating the causal effect of a time-varying exposure called marginal structural models (MSMs). Estimating parameters of MSMs involves inverse probability of treatment weighted estimation. Sometimes, the weights in these models can be too variable to obtain useful inference. We propose two modifications to the weighting scheme which may allow greater efficiency in estimation. The first also involves modification to and the second additional restrictions on the models. We illustrate these points with data from an observational study of the effect of intravenous iron dosing in hemodialysis patients on mortality.

e-mail: mjoffe@cceb.upenn.edu

CAUSAL RANDOM EFFECTS LINEAR MODEL USING NAGELKERKE'S INSTRUMENTAL VARIABLE APPROACH IN A LONGITUDINAL SETTING

Julia Y. Lin*, University of Pennsylvania
Thomas Ten Have, University of Pennsylvania

PROSPECT (Prevention of Suicide in Primary Care Elderly: Collaborative Trail) is a multiple-site study examining the effectiveness of utilizing Depression Specialists in educating patients, families, and primary-care physicians about depression and monitoring adherence to treatment on clinical depression to prevent suicide in the elderly. To account for non-adherence to randomized assignment to the Depression Specialist, we extended the instrumental variable (IV) approach described in Nagelkerke et al. (2000) for cross-sectional data to longitudinal data with time-varying non-adherence measure using a random effects linear model. Both the intervention and the usual-care groups showed improvement in depression indicated by decreases in Hamilton Depression (HAMD) scores over time. However, the intervention group showed greater improvement in depression than the usual-care group. Furthermore, the longitudinal IV approach indicated that the effect of the intervention seemed to be driven by prescribed medication rather than psychotherapy.

e-mail: jlin@cceb.upenn.edu

EXPLORING INTRINSIC DEPENDENCE WITH EL CID

Tailen Hsing, Texas A&M University
Li-yu Liu*, Texas A&M University
Marcel Brun; Texas A&M University
Edward R. Dougherty, Texas A&M University

We introduce a new measure of dependence, the coefficient of intrinsic dependence or CID, between two sets of variables. CID is model-free, invariant under monotone transformations of the variables, and is capable of fully differentiate different degrees of dependence. The estimation of CID can be carried out nonparametrically based on the ranks of the data, and the MSE of the estimator can be seen to decrease quickly with sample size. The applications that we discuss include testing for dependence, variable selection, parameter estimation in generalized linear models, and genetics data analysis.

e-mail: daisy@stat.tamu.edu

A MIXED-EFFECT PROPENSITY ADJUSTMENT FOR TREATMENT EFFECTIVENESS ANALYSES OF ORDINAL DOSES

Andrew C. Leon*, Cornell University
Don Hedeker, University of Illinois at Chicago
Jedediah Teres, Cornell University

A simulation study evaluates the propensity adjustment in a mixed-effects framework. The data analytic strategy involves two stages: 1) a model of propensity for treatment intensity examines characteristics of subjects who receive various doses using mixed-effects ordinal logistic regression. 2) the effectiveness of ordinal doses is compared in a mixed-effects survival model of time to recovery. Initially, stratified effectiveness analyses are conducted separately for each propensity quintile. If there is not a propensity by treatment interaction, quintile-specific results are pooled. Type I error, statistical power, and bias reduction are examined. Simulation specifications vary N, ICC, and odds ratios for both the propensity and treatment models. The simulation study shows that this strategy substantially reduces bias. Both statistical power and type I error rates were acceptable. However, there was a tendency toward non-convergence for N=100. This data analytic strategy is useful for observational studies of treatment effectiveness. It can adjust for self-selection bias, incorporate multiple observations per subject, and compare effectiveness of ordinal doses.

e-mail: acleon@med.cornell.edu

22. Clinical Trials: Adaptive Design and Optimization

ADAPTIVE DESIGN WITH COVARIATES

George Sirbu*, Michigan State University

The potential benefits of adaptive allocation for clinical trials were recognized quite early. Implementation of these application schemes might ease the ethical problem involved in trials on human subjects. Our work concentrates on inference for covariate-adaptive and response-adaptive randomization procedures. A widely used approach for the covariate problem is stratification. This approach however, is only applicable to qualitative covariates, or discretized quantitative covariates, and few in number or else the number of strata grows exponentially. In many respects it is more natural to perform an adjustment using a regression model that allows for both qualitative and quantitative covariates simultaneously. The problem consists of choosing in a sequential manner one of two treatments while we continuously observe the information from the process - the response and the covariates. We study both aspects of the problem: consistency for inference and optimality properties of the design. We prove strong consistency for maximum quasi-likelihood estimators of regression parameters in generalized linear regression models with minimum conditions of smoothness for the link function. The consistency results are general enough to allow us to design the allocation in a large variety of situations. Properties of these designs are described and compared.

e-mail: sirbugeo@msu.edu

POWER COMPARISONS OF ADAPTIVE TWO-STAGE DESIGNS

Cyrus R. Mehta*, Cytel Software Corporation
Anastasios Tsiatis, North Carolina State University
Yuliya Lokhnygina, North Carolina State University

The traditional design for a randomized clinical trial requires prior specification of a clinically meaningful treatment difference, to be detected with some desired power, in order to determine the sample size. Often the choice for such a clinically meaningful treatment difference is not straightforward. For this reason there has been considerable recent research on adaptive two-stage designs in which the treatment difference observed at the first stage, possibly combined with external evidence, is used to re-estimate the sample size for the second stage. We will compare the power curves of several such adaptive designs including proposals by Prochan and Hunsberger (1995), Cui, Hung and Wang (1999), and Muller and Schafer (2001). We will benchmark the performance of these power curves relative to an optimal two-stage adaptive design developed by Tsiatis and Lokhnygina (2003).

e-mail: mehta@cytel.com

AN OPTIMAL BIASED COIN ADAPTIVE DESIGN FOR HETEROSKEDASTIC OUTCOMES

Thomas E. Gwise*, University of Virginia
Feifang Hu, University of Virginia

This paper introduces a family of D-optimal adaptive designs for continuous or binary heteroskedastic outcomes with variances assumed to be unknown. The design can be viewed as a generalization of Atkinson's biased coin design (1982). Asymptotic properties are obtained and some advantages are discussed. We also report some simulation results.

e-mail: teg7a@virginia.edu

OPTIMAL TWO-STAGE ADAPTIVE DESIGNS

Yuliya Lokhnygina*, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

We derive optimal two-stage group sequential designs for normally distributed data which achieve the minimum of a mixture of expected sample sizes at the range of plausible values of a normal mean. Unlike standard group sequential tests, our method is adaptive in that it allows the group size at the second stage to be a function of the observed test statistic at the first stage. Although two-stage adaptive designs have been proposed previously in the literature, the rationale for these designs have been ad hoc. Using optimality criteria, we construct two-stage adaptive designs which we show have advantage over other popular adaptive methods. The employed computational method is a modification of the backward induction algorithm applied to a Bayesian decision problem.

e-mail: ylokhny@ncsu.edu

A FLEXIBLE METHOD FOR DESIGN OF TWO-STAGE ADAPTIVE PROCEDURES

Tatsuki Koyama*, Vanderbilt University Medical Center
Leon J. Gleser, University of Pittsburgh
Allan R. Sampson, University of Pittsburgh

We introduce a framework for designing a two-stage adaptive procedure. Our technique enables one to construct a two-stage adaptive procedure that controls many aspects of the design including the Type I error rate, the power and the maximum total sample size. Our framework also enables us to systematically compare a number of previously proposed two-stage adaptive procedures by placing them under the same formulation. We conduct an ANOVA type study to understand the effects of different components of the design specification on the performance characteristics of the design. Stage I specification components, namely, sample size, Type I error rate and Type II error rate are found to be much more important than the ‘conditional power functions’ for Stage II despite the fact that the literature pays a fair amount of attention to the choice of conditional power functions.

e-mail: tatsuki.koyama@vanderbilt.edu

DESIGNING NONLINEAR EXPERIMENTS USING NONENUMERATIVE STRATEGIES

A. John Bailer*, Miami University
Stephen Wright, Miami University

The specification of sampling times for the study of a non-linear response provide a context in which non-enumerative strategies are applied for selecting times. A start-stop experiment used environmental toxicology provides a backdrop for this design discussion. A full enumeration of the design space is conducted to establish the target value of an objective function defined by a D-optimality criterion. The greedy heuristic rapidly converged to a design that was very close to the true optimal value while a stochastic optimization strategy generated a less optimal, although still competitive, solution. The sensitivity of the suggested design to variance heterogeneity, time grid resolution, interval specification of planning values for parameters and other factors are also explored.

e-mail: baileraj@muohio.edu

A NOTE ON COMPARING EFFICACY PROPORTIONS IN TWO-STAGE CLINICAL TRIALS

Kallappa M. Koti*, U.S. Food and Drug Administration

Clinical trials comparing safety and efficacy of combination of induction and maintenance therapies are being favored due to practical reasons. However, there appears to be lack of appropriate data analytic methodology. In this note, a statistical test to compare two treatment policies in terms of their efficacy proportions that have intent to treat interpretation is proposed. The critical region of the test is determined by iteration that is performed using the SAS/IML subroutine QUAD. The proposed test has desirable power when the initial sample size is as small as 100.

e-mail: koti@cber.fda.gov

23. Challenging Issues in the Analysis of Longitudinal/Clustered Data

SEMIPARAMETRIC LONGITUDINAL/SPATIAL MODELING OF BINARY OUTCOMES, WITH APPLICATION TO ABERRANT CRYPT FOCI IN COLON CARCINOGENESIS EXPERIMENTS

Tatiana V. Apanasovich*, Ph.D., Texas A&M University
 Raymond J. Carroll, Texas A&M University

Our work is directed towards the analysis of aberrant crypt foci (ACF), which are morphologically changed colonic crypts known to be precursors to colon cancer. The colon is laid out as a gridded rectangle and the occurrence of an ACF within the grid is noted. The biological question of interest is whether these binary responses occur at random: if not, this suggests that the effect of environmental exposures is localized regionally. We derive the score test for CAR correlation models, and show that this test also arises from a modification of the score test for M^latern correlation models. Robust methods are used to lower the sensitivity to regions where there are few ACF. We cast the problem as a longitudinal/spatial binary regression with underlying Gaussian latent process. Marginal probabilities of ACF indicators are modeled semiparametrically, using fixed-knot penalized regression splines and single-index models. We model underlying latent process nonstationary as the convolution of latent local stationary processes. The dependency of the correlation function on location is also modeled semiparametrically. We fit the models using pairwise pseudolikelihood methods. Assuming that the underlying latent process is strongly mixing we prove asymptotic normality and derive the optimal rate of convergence for penalty parameters.

e-mail: tanya@stat.tamu.edu

ESTIMATION OF REGRESSION MODELS FOR THE MEAN OF REPEATED OUTCOMES UNDER NON-IGNORABLE NON-MONOTONE NON-RESPONSE

Stijn Vansteelandt*, Ghent University
 Andrea Rotnitz, Harvard School of Public Health
 James M. Robins, Harvard School of Public Health

We propose a new class of models for making inference about the mean of a vector of repeated outcomes when the outcome vector is incompletely observed in some study units and missingness is non-monotone. Each model in our class is indexed by an unidentified selection bias function which quantifies the residual association of the outcome at each occasion t and the probability that this outcome is missing after adjusting for variables observed before time t and for the past non-response pattern. Selection bias functions equalling zero encode the investigator's a priori belief that non-response of the next outcome does not depend on that outcome after adjusting for the observed past. We call this assumption sequential explainability. Because all models in our class are non-parametric, they are ideal for conducting sensitivity analyses aimed at evaluating the impact that different departures from sequential explainability have on inference about the marginal means of interest. We discuss estimation about these means under each model in our class. We extend our proposed class of models to incorporate: 1) data configurations with baseline covariates and, 2) a parametric model for the conditional mean of the outcome vector given baseline covariates. We discuss estimation in this setting.

e-mail: stijn.vansteelandt@ugent.be

ACCOUNTING FOR RESPONSE CORRELATION IN VARYING-COEFFICIENT MODELS

Naisyin Wang, Texas A&M University
Jeng-Min Chiou*, National Health Research Institutes, Taiwan

This paper considers a nonparametric varying coefficient model in a longitudinal data setting. An estimating procedure which accounts for the dependency of within- subject responses is presented. The major gain is in the improvement of estimation efficiency. Asymptotic properties of this estimator are provided. The numerical efficacy of the proposed methodology is demonstrated through a small simulation study and the analysis of a real data set.

e-mail: jmchiou@nhri.org.tw

DYNAMIC CRITERIA AND LONGITUDINAL ACCURACY

Patrick J. Heagerty*, University of Washington

Longitudinal marker measurements are often used to guide medical decisions such as the timing and choice of intervention. A basic scientific goal is to use the available information to accurately predict those subjects who are likely to experience a key clinical event such as cancer onset or death. With censored survival endpoints prognostic accuracy may be summarized extensions of sensitivity and specificity which are commonly used for binary response models. In this talk we propose new time- dependent accuracy summaries based on time-specific versions of sensitivity and specificity calculated over risk sets. The accuracy summaries are connected to a previously proposed global concordance summary which is a variant of Kendall's tau, yet allow simple extension to longitudinal covariate or marker processes. Standard Cox regression output can be used to obtain estimates of time- dependent sensitivity and specificity, and time- dependent receiver operating characteristic (ROC) curves. Semi- parametric estimation methods appropriate for both proportional hazards and non-proportional hazards data are introduced and evaluated in simulations.

e-mail: heagerty@u.washington.edu

24. Statistical Issues in Emerging Diseases

BSE AND PRION DISEASES IN NORTH AMERICA - WHAT ARE THE RISKS?

Steven A. Anderson*, U.S. Food and Drug Administration

Prion diseases or transmissible spongiform encephalopathies (TSEs) include bovine spongiform encephalopathy (BSE), a disease of cattle and the human disease counterparts - variant Creutzfeldt Jakob Disease (vCJD) and classic Creutzfeldt Jakob Disease (CJD). This rare group of diseases is characterized by a neurodegenerative condition that is fatal and because the infectious agent is potentially transmissible, they pose a significant public health concern. There is a theoretical risk that these human diseases could be transmitted via blood but to date no cases of transmission via blood or blood products have been observed. However, animal studies suggest that the prion infectious agent may be transmitted via blood products. We will demonstrate the application of probabilistic modeling in evaluating the potential public health risks posed by this route of transmission and application of the methodology in evaluating the effectiveness of various risk reduction measures.

e-mail: andersonst@cber.fda.gov

ESTIMATED RISK OF WEST NILE VIRUS TRANSFUSION TRANSMISSION THROUGH VOLUNTARY BLOOD DONATION IN THE UNITED STATES, 2002

Brad J. Biggerstaff*, National Center for Infectious Diseases Centers for Disease Control and Prevention
Lyle R. Petersen, National Center for Infectious Diseases Centers for Disease Control and Prevention

Human West Nile virus (WNV) infection has spread across the United States since the virus's introduction to the US, recognized in 1999. Epidemics of WNV neuroinvasive disease and fever raise concern that transmission of WNV may occur through voluntary blood donations; transfusion-associated transmission of WNV was documented for the first time in 2002. Onset dates from the 2002 WNV neuroinvasive disease cases and historical data on WNV viremia in humans are used to estimate the risk of transfusion-associated WNV transmission. Resampling methods are used and related to an inhomogeneous Poisson process approach. Average and maximum risk of transfusion-associated WNV transmission (per 10,000 donations) during the epidemic period for the whole US were estimated as 0.36 (95% CI 0.35 – 0.37) and 1.55 (95% CI 1.43 – 1.76), respectively. For selected, high-incidence states, the average and maximum risk estimates ranged 2.12 – 4.76 and 4.34 – 10.46, respectively; for selected metropolitan areas they ranged 1.46 – 12.33 and 3.02 – 21.32, respectively.

e-mail: bbiggerstaff@cdc.gov

CONTAINING PANDEMIC OR BIOTERRORIST INFLUENZA WITH ANTIVIRALS

M. Elizabeth Halloran*, Emory University
Ira M. Longini, Emory University
Yang Yang, Emory University
Azhar Nizam, Emory University

We developed a new model to assess the best use of antiviral agents to reduce the effect of the first wave of pandemic influenza in the entire community. We used the model to investigate the effectiveness of antiviral prophylaxis targeted at close contacts. We first analyzed data from two household studies of prophylactic and therapeutic use of antivirals to obtain estimates of the antiviral efficacy for susceptibility (AVEs) and for infectiousness (AVEi). We compare targeted antiviral prophylaxis with vaccination strategies. We model an influenza pandemic or bioterrorist attack for an agent similar to A(H2N2) that caused the Asian influenza pandemic of 1957-1958. Using targeted antiviral prophylaxis (TAP), if 80% of exposed people maintained prophylaxis for eight weeks, targeted antiviral prophylaxis has potential as an effective measure for containing influenza until adequate quantities of vaccine were available.

e-mail: mehallo@sph.emory.edu

25. Statistical Models for Complex Data in Agricultural and Environmental Studies

A SEMIPARAMETRIC STOCHASTIC MIXED MODEL FOR INCREMENT-AVERAGED DATA WITH APPLICATION TO CARBON SEQUESTRATION IN AGRICULTURAL SOILS

F. Jay Breidt*, Colorado State University
Nan-Jung Hsu, National Tsing Hua University
Stephen Ogle, Colorado State University

Adoption of conservation tillage practice in agriculture offers the potential to reduce greenhouse gas emissions. Studies comparing conservation tillage methods to traditional tillage pair fields under the two management systems and obtain soil core samples from each field. Within each core, carbon stock is recorded at multiple depth increments. These data represent not the instantaneous value at a particular depth, but the total or average over the increment at that depth. A semi-parametric mixed model is developed for such increment-averaged data. The model uses parametric fixed effects to represent covariate effects, random effects and a stochastic process to capture within-core correlation, and an integrated, smooth function to describe effects of depth. The instantaneous depth function is estimated using penalized maximum likelihood. Variance components and the smoothing parameter are estimated using restricted maximum likelihood. The methodology is applied to the problem of estimating change in carbon stock due to change in tillage practice.

e-mail: jbreidt@stat.colostate.edu

FACTOR ANALYSIS MODELS FOR POLLUTION SOURCE APPORTIONMENT

William F. Christensen*, Brigham Young University

Receptor modeling involves the statistical analysis of air quality data for the purpose of apportioning pollutants to the sources from which they were emitted. Interest in receptor models has increased as regulatory agencies have begun utilizing source apportionment studies to develop and evaluate environmental policy. Factor analysis has been used for over 30 years to simultaneously estimate pollution source profiles (factor loadings) and pollution source contributions (factor scores), but often with severe limitations for interpretation. For example, traditional exploratory factor analysis suffers from factor indeterminacy and does not yield pollution source profiles and pollution source contributions that are nonnegative. We consider a confirmatory factor analysis approach which utilizes partial pollution source profile information, yields physically interpretable solutions, and accounts for correlation structure when carrying out statistical inference.

e-mail: william@stat.byu.edu

A SPATIAL-TEMPORAL AUTO-LOGISTIC REGRESSION MODEL FOR THE ANALYSIS OF THE SOUTHERN PINE BEETLE OUTBREAKS

Jun Zhu*, University of Wisconsin – Madison
Hsin-Cheng Huang, Institute of Statistical Science, Academia
Sinica Chi-tung Wu, Feng Chia University

The Southern Pine Beetles have caused severe damage to pine forests in the southern United States and hence are of great concern. Research has found that the outbreaks are influenced by factors such as host volumes, physiographic properties of the fields, and seasonal temperature. Further, outbreaks of the Southern Pine Beetles in forests throughout the southern United States show visible spatial and temporal patterns. We develop a general spatial-temporal auto-logistic regression model for the presence and absence of an outbreak, which captures spatial dependence by a Markov random field and captures temporal dependence by a Markov chain. For statistical inference, we use maximum pseudo-likelihood, which is computationally efficient, for parameter estimation, and develop a Markov chain Monte Carlo (MCMC) algorithm for prediction. We apply our method to study the outbreaks in North Carolina and evaluate the potential of the model in capturing correlation across space and over time.

e-mail: jzhu@stat.wisc.edu

PREDICTION OF RIVERINE SEDIMENT CATEGORIES USING A SPATIAL MIXTURE MODEL

Mark S Kaiser*, Iowa State University
Kyoji Furukawa, Iowa State University

Many biotic responses in river ecosystems that are of interest to aquatic ecologists are related to the underlying sediment structure in the riverbed. Prediction of sediment type over the spatial extent of a river reach holds the potential for increasing our ability to both understand and model the levels of biotic responses. We develop a spatial mixture model for this purpose consisting of a conditionally independent multinomial observation process mixed over a conditionally specified spatial process that possesses Dirichlet-like behavior. Analysis of many conditionally specified models is made complex because the corresponding joint distribution is not available in closed form. Not only is the joint unavailable in closed form for our new mixture model, but the same is true for the basic conditional specifications used to formulate the model. Nevertheless, we show that estimation and prediction are possible with this model, and illustrate this through prediction of sediment categories in the upper Mississippi river.

e-mail: mskaiser@iastate.edu

26. The Future of Biostatistical Science

Panel Discussion:

Donald A. Berry, M.D. Anderson Cancer Center
David P. Harrington, Dana-Farber Cancer Center and Harvard University
George W. Williams, Amgen, Inc.

27. Bayesian Methods in Survival Analysis

MULTIRESOLUTION MODELS IN MULTICENTER SURVIVAL DATA ANALYSIS

Peter Bouman, Northwestern University
Vanja Dukic*, University of Chicago
James Dignam, University of Chicago
Xiao-Li Meng, Harvard University

In multicenter studies one has to make inference about a population survival curve based on possibly heterogeneous survival data collected from individual enrolled centers. This talk will be about a flexible Bayesian method proposed by Bouman, Dukic

vdukic@health.bsd.uchicago.edu

BAYESIAN ESTIMATION OF MULTIVARIATE FRAILTY MODELS FOR EXCHANGEABLE SURVIVAL DATA

Catalina Stefanescu*, London Business School
Bruce Turnbull, Cornell University

Correlated exchangeable time-to-event data may be encountered in many areas, such as studies of disease occurrence in families, developmental toxicity experiments, or group randomized trials. When data are clustered, modelling the correlation within groups of observations is necessary for the validity of the statistical inference. This paper investigates a multivariate lognormal frailty model for the correlation structure within clusters, and proposes a Bayesian approach to estimation. While likelihood inference is computationally difficult, the Bayesian framework derives naturally from the inherent hierarchical structure of the model and may be implemented by means of a Gibbs sampling approach. The methodology is illustrated with several applications.

e-mail: cstefanescu@london.edu

A SIMPLE APPROACH TO FITTING BAYESIAN SURVIVAL MODELS

Paul Gustafson*, University of British Columbia
Dana Aeschliman, University of British Columbia
Adrian R. Levy, University of British Columbia

There has been much recent work on Bayesian approaches to survival analysis, incorporating features such as flexible baseline hazards, time-dependent covariate effects, and random effects. Some of the proposed methods are quite complicated to implement, and we argue that as good or better results can be obtained via simpler methods. In particular, the normal approximation to the log-gamma distribution yields easy and efficient computational methods in the face of simple multivariate normal priors for baseline log-hazards and time-dependent covariate effects. While the basic method applies to piecewise-constant hazards and covariate effects, it is easy to apply importance sampling to consider smoother functions.

e-mail: gustaf@stat.ubc.ca

A NEW MODEL FOR LONGITUDINAL AND SURVIVAL DATA WITH A CURE FRACTION

Ming-Hui Chen Chen*, University of Connecticut
Joseph G. Ibrahim, University of North Carolina
Debajyoti Sinha, Medical University of South Carolina

We develop a new joint cure rate model for longitudinal and survival data. The model allows for multiple longitudinal markers as well as a cure structure for the survival component based on the promotion time cure rate model. Several characteristics and properties of the new model are discussed and examined. A real dataset from a melanoma clinical trial is given to demonstrate the methodology.

e-mail: mhchen@stat.uconn.edu

28. Analysis of Imaging Data

COVARIATE ADJUSTMENT IN PARTIAL LEAST SQUARES FOR THE EXTRACTION OF THE SPATIAL-TEMPORAL PATTERN FROM POSITRON EMISSION TOMOGRAPHY DATA

Lei Xu*, Merck & Co., Inc.
Sati Mazumdar, University of Pittsburgh
Julie Price, University of Pittsburgh

This work extended the partial least squares methodology that has been used to extract the spatial-temporal pattern from sequential positron emission tomography (PET) image data. The extension adjusted for covariate effects with possible spatial and temporal elements. The methodology involved derivation of multi-dimensional latent variables to explain PET image data and the development of an iterative algorithm. An illustration of the extended method is provided that shows its application to a sequential PET data set that was acquired to study the temporal and spatial effects of amphetamine on cerebral blood flow. Simulation studies were performed to validate the performance of the new method.

e-mail: lei_xu2@merck.com

ESTIMATION IN NONLINEAR REGRESSION WITH APPLICATIONS TO KINETIC MODELING WITH PET IMAGING DATA

Todd Ogden*, Columbia University
Thaddeus Tarpey, Wright State University

In some nonlinear regression situations, one or more of the parameters in the expression for the regression function are estimated from a separate data source. In such a case, the typical estimation procedure is to estimate the appropriate parameters from the separate data, then to plug these estimated values into the expression for the regression function in order to estimate the other parameters. This situation arises frequently in compartment modeling when there is an external 'input function' to the system. Issues in estimation of parameters and their standard errors will be discussed. Both asymptotic and bootstrap-based approaches will be presented and compared. This work is applicable in a variety of real-data situations. One important application to be discussed is the estimation of kinetic parameters in an alternative 'model-free' formulation of a standard pharmacokinetic model for PET imaging data. In this case (and in standard pharmacokinetic fitting cases), the model is fit to each of several regions of interest in the brain and the 'input function' is estimated from data obtained from an arterial line. Approaches to estimation of parameters and their standard errors in this situation will be presented and discussed.

e-mail: to166@columbia.edu

METABOLITE CURVE FITTING IN PET IMAGING STUDIES OF NEURORECEPTOR MAPPING

Songmei Wu*, Columbia University
Todd Ogden, Columbia University
Ramin V. Parsey, NYSPI
John Mann, NYSPI

The objective of PET neuroreceptor imaging study is to obtain the information on the distribution of the target receptor throughout the brain. More specifically, the quantification is achieved by estimation of parameters in a brain region of interest (ROI), such as total tissue equilibrium volume of distribution (VT), binding potential (BP), etc. Compartment models are often used for the estimation, while an arterial plasma function for ligand concentration is required as the input function to the compartment model. Due to the metabolism of the ligand, the observed ligand concentration function is indistinguishable from that of the metabolites concentration. Therefore the observed ligand plasma function must be corrected before it can be used as the input function. Many types of models have been used for the correcting purpose. Different ligands may have different metabolism rates and need to be adjusted by different models. Some candidate models, all fit by nonlinear regression methodology, will be examined. The data set contains 94 subjects but no more than 6 observations per subject. Various diagnostics are used to assess and compare the models. The corresponding impacts on the parameter estimates of several ROIs are to be shown and compared. These models may also be fit using nonlinear mixed-effect techniques.

e-mail: sw289@columbia.edu

EXPLORING SPATIAL RELATIONSHIPS IN FUNCTIONAL NEUROIMAGING DATA

Rajan S. Patel*, Emory University
F. Dubois Bowman, Emory University
Chengxing Lu, Emory University

Functional neuroimaging using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) makes it possible to explore the interplay between spatially distinct brain regions associated with substance abuse, behaviors, emotions, or symptoms of psychiatric disorders. Neuroimaging data pose major challenges for clustering algorithms due to the intensive computations required, the massive number of voxels, the multiple sources of noise, and the range of possible characteristics of physiological clusters. We present a 2D and 3D clustering and visualization tool, which we use to compare several existing and novel clustering techniques and stopping criteria. We demonstrate the use of our clustering tools and conduct the comparative analysis using PET functional neuroimaging data from a study that examines functional relationships in brain activity related to the use of ethanol. Our results show distinct superiority of the hierarchical clustering algorithms over the partitioning algorithms. Our new variable linkage method produces results that are competitive among existing hierarchical algorithms, however also providing computational advantages for three dimensional clustering.

e-mail: rajan@alumni.rice.edu

A TEMPLATE METHOD FOR COMBINING FMRI CONNECTIVITY ACROSS SUBJECTS

Raymond G. Hoffmann*, Medical College of Wisconsin
Natalia S. Lawrence, Institute of Psychiatry London, England
Elliot A. Stein, NIDA-IRP

Determination of effective connectivity in fMRI studies is complicated by the variability between subjects. A template method using jump markov processes is proposed to determine the common connectivity pattern as well as determining the variability among subjects. An empirical Bayes approach to determining the prior distribution of the connectivity is used together with a physiological template to determine the posterior in terms of a finite mixture of normal distributions. The method is applied to subjects imaged with a 3T Bruker Medspec scanner performing 8 blocks of a rapid visual information processing (RVIP) task. Data from each subject is conformed to a standard stereotaxic (Talairach- Tourneaux) space.

e-mail: hoffmann@mcw.edu

BRAIN FUNCTION, THE MOVIE: OPTICAL IMAGING OF BRAIN ACTIVITY

Kary L. Myers*, Carnegie Mellon University

Scientists use functional neuroimaging techniques to locate and quantify areas of stimulus-induced brain activity. A particular technique, optical imaging, involves making a ‘movie’ of the active brain by recording high resolution video data, where a typical 10 minute experiment produces over 5 gigabytes of data. Optical imaging offers better spatial resolution than any other in-vivo method, making it a powerful tool for mapping activation in the brain. However, optical imaging data typically have a signal-to-noise ratio of 0.001, mostly due to physiological changes unrelated to the stimulus of interest. This noise effectively masks the activated regions. In this talk I will discuss ways of using simultaneously recorded physiological data like blood pressure readings to reduce noise in the optical imaging movies, thus allowing scientists to make better maps of the brain.

e-mail: kary@stat.cmu.edu

29. Clinical Trials: Non-Parametric Methods

NONPARAMETRIC PREDICTION OF EVENT TIMES IN RANDOMIZED CLINICAL TRIALS

Gui-shuang Ying*, University of Pennsylvania
 Daniel Heitjan, University of Pennsylvania
 Tai-Tsang Chen, Columbia University

In clinical trials with planned interim analysis, it can be valuable for logistical reasons to predict the times of landmark event such as the 50th and 100th event. Bagiella and Heitjan (Statistics in Medicine 2001; 20: 2055-2063) proposed a parametric prediction model for failure-time outcomes assuming exponential survival and Poisson enrollment. When little is known about the distributions of interest, there is concern that parametric prediction methods may be biased and inefficient if their underlying distributional assumptions are invalid. We propose nonparametric approaches to make point and interval predictions for landmark dates during the course of the trial. We obtain point predictions using the Kaplan-Meier estimator to extrapolate the survival probability into the future, selecting the time when the expected number of events is equal to the landmark number. To construct prediction intervals, we use a simulation strategy based on the Bayesian bootstrap. Monte Carlo results demonstrate the superiority of the nonparametric method when the assumptions underlying the parametric model are incorrect. We demonstrate the methods using data from a trial of immunotherapy of chronic granulomatous disease.

e-mail: gsyng@mail.med.upenn.edu

A NEW BINOMIAL VARIANCE ESTIMATOR FOR THE KAPLAN-MEIER SURVIVAL FUNCTION

Craig B. Borkowf*, Centers for Disease Control and Prevention

In this talk we propose a new hybrid piecewise Binomial variance estimator for the Kaplan-Meier survival function. The mean proportion of this estimator equals the survival function on the interval up to the median survival time, then becomes constant and equal to 0.5 on the next interval, and finally decreases again when a certain criterion is met. Also, the effective sample size equals the number of subjects not censored prior to that time. In addition, we consider an adjusted hybrid piecewise Binomial variance estimator that modifies the regular estimator for small sample sizes. We present a simulation study to compare the performance of these new hybrid piecewise Binomial variance estimators to the Greenwood and Peto variance estimators for small sample sizes. We show that these new variance estimators give better variance estimates than these traditional variance estimators, and hence confidence intervals constructed with these estimators have more nominal coverage rates. Indeed, the Greenwood and Peto variance estimators can substantially underestimate the true variance in the left and right tails of the survival distribution, even with moderately censored data. Finally, we illustrate the use of these new and traditional variance estimators on an example from a leukemia clinical trial.

e-mail: CBorkowf@cdc.gov

A NONPARAMETRIC TEST FOR COMPARING SEVERAL SURVIVAL CURVES WITH INTERVAL CENSORED DATA

Dae Ryong Kang*, Yonsei University College of Medicine, South Korea
Jinheum Kim, University of Suwon
Chung Mo Nam, Yonsei University College of Medicine, South Korea

Interval-censored survival data often occurs in medical or health studies that entail a periodic follow-up. We propose a nonparametric test for comparing several survival curves with interval-censored data. Our test do not need so-called EM algorithm as in Finkelstein and Wolfe (1985), Finkelstein (1986) and Sun (1996, 1997) because we introduce an uniform weight only depending on the size of risk set at an observed failure time instead of weight involving estimated survivals and thereby can save much time in computational work. The results of a simulation study comparing the proposed test with the usual logrank test and the Finkelstein (1986);'s score test present that our proposed test is very satisfactory in terms of size and power. Also, the proposed method was applied to a breast cosmesis data (Finkelstein and Wolfe, 1985) and a lung cancer post-operative treatment study taken from Yonsei Cancer Center. Keywords : Interval-censored data, Weighted logrank test, EM algorithm

e-mail: cosakang@yumc.yonsei.ac.kr

PAIR-WISE COMPARISON ESTIMATION OF CLUSTERED NONPARAMETRIC TRANSFORMATION MODELS

Yan Li*, Rush University Medical Center
Daniel Zelterman, Yale University

In survival analysis, various parametric, semiparametric, and nonparametric models are employed to estimate incidence rates. The semiparametric transformation model includes the proportional hazards model and the proportional odds model as special cases. We propose an M- estimator for the nonparametric transformation model to analyze clustered failure time data. In our model, the latent error distribution is unspecified, and the unknown transformation function can be heterogeneous across different clusters. Our new estimator makes a nonparametric pair-wise comparison of all subjects within the same cluster. It takes censoring and clustering into account, and incorporates the maximum rank correlation (MRC) estimator as a special case. We illustrate the method with an application to an HIV pregnancy study, which matches each HIV+ woman with three HIV- women on potential confounding factors. In this dataset, HIV infection is associated with a decrease in pregnancy rates and it appears that this decrease in fertility is negatively confounded by the illicit drug use.

e-mail: yan_li@rush.edu

NONPARAMETRIC MODEL SELECTION FOR COX'S PROPORTIONAL HAZARD MODEL

Chenlei Leng*, University of Wisconsin
Hao Helen Zhang, North Carolina State University
Yi Lin, University of Wisconsin
Grace Wahba, University of Wisconsin
I. J. Schoenberg, University of Wisconsin

We consider Cox proportional hazard models using smoothing splines. By imposing a novel penalty instead of the usual square penalty in smoothing spline ANOVA cases, our method shrinks functional components and produces some components that are exactly zeros. It is a variant of the 'COSSO' proposal by Lin and Zhang (2002) and our method extends it to nonparametric Cox proportional hazard models. An efficient algorithm based on a reformulation of the penalized likelihood is given. To choose the smoothing parameter, an approximation to the leave-out-one likelihood cross validation score criterion is derived. Both simulations and real examples suggest our proposal is very powerful for model selection and component estimation in survival analysis.

e-mail: chenlei@stat.wisc.edu

MARGINAL ESTIMATION IN MULTISTAGE MODELS USING CURRENT STATUS DATA

Somnath Datta, University of Georgia
Rajeshwari Sundaram*, University of North Carolina-Charlotte

In a multistage model, individuals are observed over time in various stages. Estimation of marginal quantities such as the stage occupation probabilities, marginal hazards and stage waiting time distribution functions are of interest for answering questions such as: what is the probability that a randomly selected person is in a certain stage at a certain time point, what is the rate (hazard) at which person in a certain stage moves to another stage, and so on. In the case of complete data such quantities are easily estimated using their empirical versions. This talk will address these estimation problems in current status data without making any structural assumptions (such as Markov or semi-Markov) about the multistage process. A simulation study and a real data example will be presented.

e-mail: rsundara@uncc.edu

30. Linkage Analysis

UNIFICATION OF VARIANCE COMPONENTS AND HASEMAN-ELSTON REGRESSION FOR QUANTITATIVE TRAIT LINKAGE ANALYSIS

Wei-Min Chen*, Johns Hopkins University
Karl W. Broman, Johns Hopkins University
Kung-Yee Liang, Johns Hopkins University

Two of the major approaches for linkage analysis with quantitative traits in humans include variance components and Haseman-Elston regression. Previously, these have been viewed as quite separate methods. We describe a general model, fit by use of generalized estimating equations (GEE), for which the variance components and Haseman-Elston methods (including many of the extensions to the original Haseman-Elston method) are special cases, corresponding to different choices for a working covariance matrix. We also show that the regression-based test of Sham et al. (2002) is equivalent to a robust score statistic derived from our GEE approach. These results have several important implications. First, this work provides new insight regarding the connection between these methods. Second, asymptotic approximations for power and sample size allow clear comparisons regarding the relative efficiency of the different methods. Third, our general framework suggests important extensions to the Haseman-Elston approach which make more complete use of the data in extended pedigrees and allow a natural incorporation of environmental and other covariates.

e-mail: wechen@jhsph.edu

A LATENT VARIABLE MODEL OF LINKAGE ANALYSIS FOR ORDINAL TRAITS

Rui Feng*, Yale University
Heping Zhang, Yale University

Linkage analysis is routinely used to map genes for human diseases and conditions. The existing linkage analysis methods require, however, that the disease or condition must either be dichotomized (as yes or no) or measured by a quantitative trait such as blood pressure for hypertension. On the other hand, many diseases and conditions of medical interest are recorded in ordinal scales, such as cancer stage and the severity of mental health and behavioral conditions. The objective of this work is to establish a framework to conduct linkage analysis for ordinal traits. We propose a latent variable, proportional odds model to analyze general pedigree data. Simulation studies are carried out to assess the power of this new model, and they reveal that the power of our model can increase up to ten folds of that using the standard linkage analysis method. Simulation results also suggest that our model is robust with regard to parameter misspecifications.

e-mail: rui.feng@yale.edu

ASYMPTOTIC EQUIVALENCE OF SEVERAL AFFECTED SIBSHIP STATISTICS IN NONPARAMETRIC LINKAGE ANALYSIS

Xiaoling Wu*, Johns Hopkins University
Daniel Q. Naiman, Johns Hopkins University

Affected sibship statistic which measures allele sharings of affected sibships is pivotal in nonparametric linkage test. We investigate the distribution and cumulants of several commonly used affected sibship statistics (including S-pairs and S-all, Whittemore and Halpern, Biometrics 1994) when sibships are of size larger than 2. These statistics are asymptotically equivalent to each other and they all have the same covariance structure between markers after normalization. The result leads to an Ornstein-Uhlenbeck process approximation to the affected sibship statistics process along the genome, which is essential in exploring the validity of multiple linkage testing. The conclusion above is based on Markov chain Monte Carlo simulation of unrelated sibships' allele sharings and analysis of genome scan data provided by GAW (Genetic Analysis Workshop) 13. We adopted our own efficient multipoint algorithm to decide the probability distribution of identical by descent (IBD) sharings in the GAW 13 data. By observing the behavior of their Edgeworth approximations, we also demonstrate that the total score statistic from sibling pairs samples summed over the sample at a single marker converges to standard normal faster than summed over sibling triple samples (or quadruples, etc) which contains the same number of affected siblings.

e-mail: xlwu@jhu.edu

TESTING FOR HOMOGENEITY IN GENETIC LINKAGE ANALYSIS

Yuejiao Fu*, University of Waterloo
Jiahua Chen, University of Waterloo
Jack D. Kalbfleisch, University of Michigan

We apply the modified likelihood ratio test to two binomial mixture models arising from genetic linkage analysis. The resulting limiting distribution of the test statistic for both models is all found to be a mixture of chi-squared distributions. A consideration of random family sizes for both models gives similar results. We also explore the power function under local alternatives. Simulation studies show that the modified likelihood ratio test is more powerful than other methods under a variety of model specifications.

e-mail: y6fu@math.uwaterloo.ca

THE IBD DISTRIBUTION OF A SIB PAIR AND THE LIMITING DISTRIBUTION OF TEST STATISTIC IN ASP LINKAGE ANALYSIS

Jiahua Chan, University of Waterloo
Mary E. Thompson, University of Waterloo
Zeny Z Feng*, University of Waterloo

In affected-sib-pair linkage analysis, the IBD data of markers are collected. Holmans (1993) showed that, when a disease is Mendelian and the recombination rates of male and female are the same, the IBD distribution of a marker satisfies the “possible triangle constraint”. By a simple direct approach, we show this conclusion is true in general regardless of genetic models behind it. Under the hypothesis of no linkage, the probabilities of a sib pair sharing 0, 1 and 2 alleles IBD are $1/4$, $1/2$ and $1/4$. The limiting distribution of the log-likelihood ratio test statistic without the possible triangle constraints is Chi-square distributed with 2 df. With the possible triangle constraint, the test statistic is stochastically smaller and is no longer Chi-square distributed. The corresponding test may offer greater statistical power. We derive the limiting distribution of the log-likelihood ratio test statistic under the possible triangle constraint.

e-mail: zzqfeng@math.uwaterloo.ca

VARIANCE-COMPONENT LINKAGE METHODS FOR COMPLEX TRAIT DATA

Michael P. Epstein*, Emory University
Xihong Lin, University of Michigan
Michael Boehnke, University of Michigan

Variance-component (VC) linkage analysis is a powerful and flexible tool for detecting genes that influence complex quantitative traits. However, VC methods typically assume that the trait of interest follows a multivariate normal distribution within a family. Studies have shown that violation of the VC normality assumption can lead to biased parameter estimates and elevated type I error rates. Therefore, application of the VC method to binary, count, or gamma-distributed trait data will be difficult. Given the appealing design of the VC method, we have extended it to accommodate non-normal trait data using the generalized-linear-mixed-modeling (GLMM) framework. Using our proposed method, one can perform VC linkage analyses for binary, count, or gamma-distributed data under both random and non-random sampling designs. Like the traditional VC method, our general framework has a flexible modeling structure, which allows for testing of multiple genetic and environmental effects and interactions. We illustrate the utility of these extensions using simulated data and quantitative trait data from the Finland-United States Investigation of NIDDM (FUSION) study.

e-mail: mepstein@genetics.emory.edu

QUANTITATIVE TRAIT LINKAGE ANALYSIS USING GAUSSIAN COPULAS

Mingyao Li*, University of Michigan
Peter X-K Song, York University, Canada
Goncalo R. Abecasis, University of Michigan
Michael Boehnke, University of Michigan

Variance-component (VC) methods have been playing a central role in the mapping of quantitative trait loci. One of the key assumptions for the traditional VC methods is that the trait follows a multivariate normal distribution. Violation of this assumption usually results in loss of power and bias in estimation. Here, we present an extension of the traditional VC models through Gaussian copulas, which enables us to accommodate a variety of continuous and discrete traits. Given a marginal distribution of the trait, the copula VC method models specify joint distributions of the trait in that the variance and covariance are formulated in a desired structure. Our method allows full likelihood inference and hence provides a powerful tool to conduct QTL analysis. In addition, the traditional VC method is a special case of our method. We compare the performance of the copula and traditional VC methods by simulation studies. Our results suggest that the copula VC method generally leads to increase in linkage power as compared to the traditional VC method, and the extent of the power increase depends on the degree to which the trait distribution differs from normality. We apply the method to the Australian twin skin cancer data, in which the number of moles are measured for twins.

e-mail: myli@umich.edu

31. Survey Research Data

EXTENSIONS OF MANGAT'S (1994) RANDOMIZED-RESPONSE MODEL

Jong-Min Kim*, University of Minnesota at Morris
Joshua M. Tebbs, Kansas State University
Seung-Won An, Oklahoma State University

The randomized-response technique can be an effective survey method when collecting sensitive information. In this paper, we extend the model proposed by Mangat (1994, *Journal of the Royal Statistical Society, Series B*, 56, 93-95) in two ways. First, we propose a Bayesian version of Mangat's (1994) model which is applicable to situations wherein prior information is available on the sensitive characteristic prevalence. We show that our Bayesian approach can provide greatly-improved estimators when compared to the maximum likelihood estimator in the Mangat (1994) procedure; furthermore, our approach provides estimators guaranteed to lie within the parameter space. Second, we extend the Mangat (1994) procedure to include data from a stratified-sampling protocol and show that both of our stratified procedures (one non-Bayesian and one Bayesian) are more efficient than the one initially proposed by Mangat (1994) for a single population.

e-mail: jongmink@mrs.umn.edu

USE OF AUXILIARY VARIABLES IN MARKOV SAMPLING FOR FINITE POPULATION

G.K. Balasubramani*, University of Pittsburgh
Stephen R. Wisniewski, University of Pittsburgh

Suresh Chandra et al. (1992) have introduced the Markov sampling scheme for finite population and studied the performance of the Horvitz-Thompson estimator in Markov sampling for the two types of deterministic population, viz., linear and exponential. The auxiliary information has used to obtain the estimate and the mean square errors of Markov sampling was studied by Suresh Chandra and Sampath (1992) and infer that the Markov sampling with auxiliary information performs better than other sampling schemes. This paper describes the method of Markov ratio, and Markov regression estimators and derived the variances of the following four models: (i) linear (ii) quadratic (iii) periodic and (iv) Exponential. The variances of the Horvitz- Thompson estimator of the population total in Markov sampling, simple ratio estimator, regression estimator, Markov ratio, Markov weighted ratio and Markov weighted regression estimators are evaluated for different values of population and sample sizes through Monte Carlo simulation studies. The empirical studies also made along the lines of the computer based numerical evaluation of the variances of the estimators. The investigation revealed that the weighted Markov regression estimators provide minimum mean square error among other estimators for all the four models considered for the study.

e-mail: balagk@edc.gsph.pitt.edu

A SIMPLE SAMPLE SIZE FORMULA FOR REGRESSION-BASED REFERENCE LIMITS

Carine Bellera*, McGill University
James Hanley, McGill University

Reference values (such as the 5% or the 95% reference limits) are widely used in areas such as anthropometry, medicine and clinical chemistry. These parameters describe the distribution of a quantitative variable in a healthy population. They are often a smooth function of age. Thus, rather than estimate reference values separately for each age, it is more economical, and common, to use simple (or multiple) linear regression to estimate reference limits for each age. The variability of such estimates, and the sample sizes needed to estimate the parameters with a desired level of precision have been addressed previously. However, the available methods to determine the sample size required for such studies are overly complex, and not very transparent. We propose a simple, and intuitive, formula to determine sample size/precision. We present calculations based on samples with a uniform age distribution. However, the way we derive the formula it makes it easy to evaluate the effect on precision of different age-sampling strategies.

e-mail: carine.bellera@mail.mcgill.ca

ISSUES IN THE ANALYSIS OF COMPLEX SURVEY DATA: DESIGN BASED VS. MODEL BASED APPROACHES

Marianne H. Bertolet*, Carnegie Mellon University
Howard Seltman, Carnegie Mellon University
Joel Greenhouse, Carnegie Mellon University
Kelly Kelleher, The Ohio State University

The differences between design and model based analyses of complex surveys can be subtle and confusing. This paper describes the differences between the underlying assumptions, explains how these differences propagate in regression analyses, and contains examples using the National Survey of Child and Adolescent Well-Being (NSCAW). Briefly, the core differences lie in the origin of the variability of the data, with the design based approach placing the variability in the randomization distribution and the model based approach placing the variability in the definition of the model. This difference manifests itself in many areas, including the target of inference and the interpretation of the results. The use of weights in survey analysis can cause much confusion. Design based analyses must use the weights to provide unbiased estimators. However, in model based analyses, the debate over the use of the sampling weights continues, as explained in this paper. The examples from NSCAW take three approaches to determining models, and compare them using design based, model based, weighted and unweighted techniques.

e-mail: mhb2@stat.cmu.edu

FLEXIBLE MATCHING IN SAMPLE SURVEYS CONTAINING OBSERVATIONAL STUDIES

Ben B. Hansen*, University of Michigan

In education research, medicine, administrative studies of class equity, and other areas, one finds sample surveys aiming to compare two predesignated groups along measures to be obtained from the sample. As an observational study, the comparisons thus enabled call for statistical adjustment to account for confounding variables; but as the result of a probability sample, the data being used in these comparisons call for an analysis that respects the survey's design. These demands are often at odds. This talk discusses advantages of anticipating a binary comparison at the design stage, using frame data and a flexible matching routine (optimal full matching or matching with a variable number of controls) to create matched sets that will in turn be treated as clusters in the selection of the sample. The advantages include increased likelihood of obtaining comparable subgroups, even in a self-weighting sample; support for design-based inference about 'causal' parameters (e.g. the effect of treatment on the treated); and implicit adjustment for potentially identifying variables that avoids the using them directly, thus easing both IRB approval and the creation of public-use files. The argument draws on the speaker's experiences with a small gender-equity study.

e-mail: ben.b.hansen@umich.edu

DOES WEIGHTING FOR NONRESPONSE INCREASE THE VARIANCE OF SURVEY MEANS?

Roderick J. Little*, University of Michigan
Sonya Vartivarian, University of Michigan

Nonresponse weighting is a common method for handling unit nonresponse in surveys. A widespread view is that the weighting method is aimed at reducing nonresponse bias, at the expense of an increase in variance. Hence, the efficacy of weighting adjustments becomes a bias-variance trade-off. This note suggests that this view is an oversimplification - nonresponse weighting can in fact lead to a reduction in variance as well as bias. A covariate for a weighting adjustment must have two characteristics to reduce nonresponse bias – it needs to be related to the probability of response, and it needs to be related to the survey outcome. If the latter is true, then weighting can reduce, not increase, sampling variance. A detailed analysis of bias and variance is provided in the setting of weighting for an estimate of a survey mean based on adjustment cells. The analysis suggests that the most important feature of variables for inclusion in weighting adjustments is that they are predictive of survey outcomes; prediction of the propensity to respond is a secondary, though useful, goal. **KEY WORDS:** missing data, nonresponse adjustment, sampling weights, survey nonresponse

e-mail: rlittle@umich.edu

BIAS CORRECTION IN RARE EVENTS DATA

Bin Wang*, University of South Alabama
Yu Xu, University of South Alabama
Candice M Ross, University of South Alabama
Kun Huang, University of South Alabama

In a study of rare events data, binary dependent variables usually contains much more zeros (cancers, software bugs, wars, epidemiological infections) than ones. Standard statistical procedures without any correction may underestimate the probability of occurrence of rare events. In addition, the data from a survey may subject to a selection bias because some questions in the questionnaire touch sensitive issues. This study incorporates appropriate corrections to a logistic regression model. The performance of the new model will be illustrated via simulation studies and improvements of sampling designs will be suggested as well.

e-mail: bwang@jaguar1.usouthal.edu

32. Gene Selection Methods

COMPARISON OF GENE SELECTION METHODS IN CLASS PREDICTION

Shibing Deng*, University of North Carolina at Chapel Hill
 Young Truong, University of North Carolina at Chapel Hill

Microarray technology is widely used in cancer research and one of its important applications is to predict clinical outcomes, such as tumor classes. Due to the large amount of data and the noises associated, it is desirable to select a small set of genes that have good prediction power. There are several commonly used gene selection methods in the literatures, and most of them are univariate approaches. They select gene one at a time and do not take the effect of other genes into account. We propose a couple of multivariate approaches to select predictive genes and compare genes selected by these methods with those by univariate methods in terms of consistency, prediction accuracy and robustness. In applying to several tumor microarray datasets, we find genes selected by multivariate methods have some advantages in predicting tumor classes.

e-mail: sdeng@bios.unc.edu

PROBABILITY BASED CLASS PREDICTION IN TOXICOGENOMICS

Nandini Raghavan*, J&J, PRD L.L.C.
 Dhammika Amaratunga, J&J, PRD L.L.C.
 James J. Colaianne, J&J, PRD L.L.C.

This work was motivated by the following problem. Suppose that you have data on various classes of compounds (say, non-toxic, toxic type I, toxic type II, etc.). Given data on a new compound, can you assign the new compound a probability-based score as to which class it belongs? Statistically, this is a familiar problem of classification, but our problem had some unusual characteristics: (i) fuzzy class boundaries, (ii) differential misclassification costs, (iii) the number of potential predictors exceeds the number of available samples by orders of magnitude, (iv) new compounds may not belong to any of the classes for which we have data. We describe a three-pronged approach to gene-selection which we used to elicit the most relevant features for differentiating between classes. We also compare the performance of various classifiers, including linear discriminant analysis and a Bayesian classifier that we developed to address the requirements of the application described above. Our methodology gave results that reproduced well in an independent test set. This problem arose in a toxicogenomics context. One focus of current research in this area is to use data from DNA microarrays to develop gene expression “signatures” for sets of compounds known to induce a particular toxic response.

e-mail: nraghava@prdus.jnj.com

A SEMIPARAMETRIC APPROACH FOR DETECTING PROGNOSTIC GENES ASSOCIATED WITH SURVIVAL OUTCOME IN MICROARRAY STUDIES

Kouros Owzar*, Duke University Medical Center
Sin-Ho Jung, Duke University Medical Center
Pranab Kumar Sen, University of North Carolina, at Chapel Hill

A challenging and crucial issue in large scale studies involving microarrays, is to detect those genes which are prognostic with respect to a censored clinical outcome variable such as time to death. In this specific context, a gene is said to be prognostic if its corresponding expression is associated, in some statistical sense, with the target outcome variable. A comprehensive methodology should properly quantify the association between each gene and the outcome variable, under the auspices of a bona-fide statistical model, and outline robust estimation methods for the parameters. For the purpose of statistical inference, given that the number of genes is typically large, the methodology should also admit flexibility in the admission of adjustments of multiplicity. What is to be outlined in this talk is a semi-parametric method which addresses all of these issues. Application of the method to a case study will be presented.

e-mail: kouros.owzar@duke.edu

MODELING STRATEGIES FOR NEURAL NETWORKS BASED ON AIC CRITERIA AND GENETIC ALGORITHMS

Doug Landsittel*, University of Pittsburgh
Harshinder Singh, West Virginia University and NIOSH Biostatistics Branch
Vincent C. Arena, University of Pittsburgh
Stewart J. Anderson, University of Pittsburgh

Although neural networks are commonly implemented for classification and prediction, selection of the optimal model remains a challenging task due to the complexity of their model architecture and lack of established selection algorithms. This study will investigate a model selection strategy for feed forward neural networks with a binary outcome using the AIC and genetic algorithms. To utilize the AIC, effective model dimension is quantified by a generalized measure for degrees of freedom. The general approach is to fix the number of hidden units, weight decay, and any other variations in the model structure, and then search the space of possible models (across differing subsets of explanatory variables) via genetic algorithms and the AIC. After determining the optimal subset of explanatory variables, the number of hidden units, weight decay, and other variations in model structure are then systematically varied across some predetermined range, and the AIC is evaluated to determine the resulting optimal model. Results are illustrated on a set of markers shown to be predictive of early stage ovarian cancer.

e-mail: landsittel@upci.pitt.edu

SOME NEW RESULTS IN MULTIVARIATE ANALYSIS

Samuel W. Wu*, University of Florida
Hongying Li, University of Florida
George Casella, University of Florida

It is well known that in a general multi-parameter setting, there may not exist any unique best test. More importantly, unlike the univariate case, the power of different test procedures could vary remarkably in multivariate analysis. In this talk we discuss some new findings on combining univariate tests for multivariate problems. Bahadur relative efficiencies are derived for six methods of combining non-independent univariate tests which are assumed to be asymptotically normal. Relationship is established between Fisher's method of combining tests and a new class of tests that have best average power for multivariate linear hypotheses. A Monte Carlo study indicate that the small-sample power of the combination methods compare very well to the Hotelling type tests in many cases. Applications in microarray analysis, QTL detection, and composite analysis of 5 aphasia trials are also considered.

e-mail: samwu@biostat.ufl.edu

33. Generalized Linear Models

A HYBRID METHOD FOR THE ANALYSIS OF POISSON AND OVER-DISPERSED POISSON COUNT DATA WITH ZERO-INFLATION

Grace X. Liu*, Merck, Inc.
Dirk F. Moore, UMDNJ

In this paper we consider a modification of the EM algorithm for modeling count data with zero inflation. In this modification, a probability model based on the Poisson-gamma (negative binomial) distribution is used for the E-step of the algorithm, and a generalized estimating equation method is used in place of maximum likelihood to estimate the regression and over-dispersion parameters. We use the method to accommodate excess zeros found in two numbers of classical data sets, one on lung cancer deaths in Welsh nickel refinery workers, and one on doctoral publication data.

e-mail: grace_liu@merck.com

INVESTIGATION OF THE BEHAVIOR OF POISSON REGRESSION WITH VERY SMALL COUNTS WITH APPLICATION TO MORTALITY DATA ANALYSIS

Yue Cui*, University of Minnesota
Melanie M. Wall, University of Minnesota

The first concern of this work is to detect racial disparities in cancer mortality based on mortality data collected for the five counties of the Twin Cities Metro area. In order to do this, we need to get population at risk stratified to match that of the deaths. By different stratification methods, we end up with mortality data at different geographic levels. Different models are used to fit data at each level. Pearson and Deviance statistics are used to check model fit. But opposite conclusions will be made from Pearson and Deviance if we compare the statistics with 1 to conclude over or under dispersion of data for the models considered. In an attempt to better understand the behavior of those statistics, we present a simulation study to examine the sampling distribution of Pearson and Deviance and also derive some analytic results which validate the simulation results.

e-mail: yuecui@biostat.umn.edu

ASSESSING POISSON-NESS: HOW FISHY IS THE MOUSE?

Michael A. Newton*, University of Wisconsin-Madison
David Hastie, Bristol University

In a recent Nature Genetics article, K. Haigis and W.F. Dove used special chromosomal constructs to understand factors which affect the inactivation of the adenomatous polyposis coli (Apc) gene during intestinal tumor growth in the mouse. The experiment produced four groups of tumor counts which seemed to exhibit purely Poisson variation. This was a striking feature because although such variation is expected on first principles, extensive historical data from genetically identical (but chromosomally different) mice shows extra-Poisson variation. The experiment raised a statistical question regarding the Poisson-ness of the tumor count distributions. After considering several hypothesis-testing strategies and permutation schemes, we developed calculations in the context of a multi-group negative-binomial model in which both shape and rate parameters are allowed to vary among groups. We used both AIC/BIC model selection and reversible jump MCMC to assess the Poisson-ness of the tumor counts and thus to quantify evidence in favor of the null hypothesis.

e-mail: newton@stat.wisc.edu

QUANTILE DISPERSION GRAPHS FOR POISSON REGRESSION MODELS

Siuli Mukhopadhyay*, University of Florida
Andre I. Khuri, University of Florida

Unlike linear models, optimal designs for generalized linear models depend on the unknown parameters of the fitted model. This dependence problem causes a great difficulty in the construction of designs, and some prior knowledge of the parameters is needed. In this article, a graphical technique is proposed for comparing and evaluating designs for a log-linear Poisson model. Quantiles of the scaled mean square error of prediction are obtained within the region of interest, R . For a given design these quantiles depend on the parameters. Plots of the maxima and minima of the quantiles, over a subset of the parameter space, produce the so-called quantile dispersion graphs(QDGs). The use of QDGs provides a convenient technique for evaluating and comparing designs for GLMs. They provide information concerning the quality of prediction of designs and its sensitivity to the model's parameter values. The robustness of the QDGs to the form of the link function, will also be discussed.

e-mail: smukhopa@stat.ufl.edu

ANATOMICAL LOCATION OF ACTIVE SPINAL NEURONS USING C-FOS GENE EXPRESSION

Xiaofeng Wang*, Case Western Reserve University
Kenneth J. Gustafson, Case Western Reserve University and the Cleveland VA FES Center
Michael A. Moffitt, Case Western Reserve University
Scott Snyder, Case Western Reserve University
Warren M. Grill, Case Western Reserve University and the Cleveland VA FES Center
Jiayang Sun, Case Western Reserve University

The spinal cord contains the neural circuitry required to generate coordinated movements, but detailed maps of the spinal neural circuitry are not available. A spatial organization of receptive fields and a modular organization of the flexion withdrawal reflex system have been recently supported. However, the location and topographical organization of interneurons involved in flexion reflex pathways have not been systematically examined. In this paper, we present our work to determine the anatomical locations of spinal neurons involved in the hindlimb flexion withdrawal reflex using expression of the immediate early gene *c-fos*. A Poisson regression model is developed to fit the data. A nonparametric mapping with a FDR controlling procedure is given to determine locations within the spinal cord where stimulation caused a significant increase in the number of active spinal neurons. The FDR controlling procedure overcomes the multiplicity effect from the hundreds of null hypotheses involved above. The results suggest that the region of increased spinal activity upon stimulation is similar in shape to a traditional motoneuron population. Parallel 3-D distributions are observed for the two stimulated nerves in the upper laminae. The results also support the spatial/modular organization of the flexion withdrawal reflex system within the spinal cord.

e-mail: Xiaofeng@cwru.edu

ESTIMATION OF THE CHANGE-POINT IN GENERALIZED LINEAR MODELS

Hongling Zhou*, Johns Hopkins University
Kung-Yee Liang, Johns Hopkins University

There are many applications of change-point problems in medical research as well as in other scientific fields. Statistical models which involve change-points are termed as change-point models. Much of the existing research is focused on change-point models with normally distributed response variables. In a broader framework, we consider generalized linear models with a single and abrupt change. In this type of the estimation problems, the usual asymptotic theory becomes inappropriate since the likelihood function lacks smoothness with respect to the change-point. We propose to replace the likelihood-based objective function with a sufficiently smooth function whose almost sure limit is the same as that of the likelihood-based objective function. The asymptotic properties of the estimated change-point and regression coefficients are presented. The estimation procedure is exemplified on the EURAMIC (EUROpean study in Antioxidants, Myocardial Infarction, and breast Cancer) data.

e-mail: hzhou@jhsph.edu

TESTS FOR PAIRWISE MULTIPLE COMPARISONS OF COEFFICIENTS OF VARIATION

Yvonne M. Zubovic*, Indiana University Purdue University-Fort Wayne
Chand K. Chauhan,, Indiana University Purdue University-Fort Wayne

Consider the situation in which characteristics, such as weight or body length, are observed for several populations of an animal. The populations may differ with respect to age, gender, species, or habitat, and interest lies in comparing the relative variability of the groups. Several tests have been proposed in the literature for comparing k population coefficients of variation, where k is at least two. When the hypothesis of equal coefficients of variation is rejected, a natural step in the analysis is to identify which coefficients differ from the rest. In this paper, we consider multiple comparison procedures to test for differences between each possible pair of coefficients of variation. Several competing test statistics are proposed, based on the sample coefficient of variation, using different estimators for its variance. The methods are demonstrated using data on body measurements for different groups of lizards. To evaluate the statistical properties of these procedures, we investigate the distribution of the test statistics. Simulation results are presented to demonstrate the family Type I error rates as well as Type II error rates of the various statistics.

e-mail: zubovic@ipfw.edu

34. Nonparametric Regression Techniques Applied to Analysis of Survival Data

MARGINAL HAZARD MODELS WITH VARYING-COEFFICIENTS FOR MULTIVARIATE FAILURE TIME DATA

Jianwen Cai*, University of North Carolina at Chapel Hill

Jianqing Fan, Princeton University

Haibo Zhou, University of North Carolina at Chapel Hill

Yong Zhou, University of North Carolina at Chapel Hill

Statistical inference for the marginal hazard models with varying-coefficients for multivariate failure data is studied. A local pseudo-partial likelihood procedure is proposed for estimation of the unknown coefficients and the intercept function. A weighted average estimator is also proposed in an attempt to improve the efficiency of the estimator. The consistency and the asymptotic normality of the proposed estimators are established and the standard error formulas for the estimated coefficients are derived and empirically tested. To reduce the computational burden of the maximum local pseudo-partial likelihood estimator, a simple and useful one-step estimator is proposed. Statistical properties of the one-step estimator is established and simulation studies are conducted to compare the performance of the one-step estimator to the maximum local pseudo-partial likelihood estimator. The results show that the one-step estimator can save computational cost without deteriorating its performance both asymptotically and empirically. A data set from the Busselton Population Health Surveys is analyzed to illustrate our proposed methodology.

e-mail: cai@bios.unc.edu

ESTIMATION IN ACCELERATED FAILURE MODELS

Robert L. Strawderman*, Cornell University

A transparent extension of the accelerated failure time model is proposed for recurrent event data. Semiparametric efficient estimation is considered in the context of estimating the corresponding intensity function for recurrent event counting processes. As expected, semiparametric efficient estimation of the regression parameter involves the derivative of the log of the baseline intensity function. Smoothing the nonparametric component of the model using penalized splines is therefore considered. Given an estimate of the regression parameter, it is shown that one may estimate the spline component by fitting a suitably constructed Poisson generalized linear mixed model. An iterative procedure is thereby obtained for jointly estimating the regression parameter and smoothed baseline intensity in AFT-type models via maximum likelihood.

e-mail: rls54@cornell.edu

ANALYSIS OF THE ADDITIVE HAZARDS MODEL WITH MEASUREMENT ERRORS - NONPARAMETRIC CORRECTION METHOD

Jianguo (Tony) Sun*, University of Missouri
Liuquan Sun, University of Missouri

Measurement errors occur in many situations and one common method to deal with it is to assume that there exists a validation data set. However, this may not be the case sometimes and instead, there may only exist repeated measurements or observations on the variable with measurement errors. This paper considers one such situation in survival analysis where the objective is regression analysis of failure time data with the additive hazards model and measurement errors on covariates. For inference about regression parameters, an estimation procedure is developed by using a nonparametric-correction approach under the additive measurement error model. The resulting estimators are shown to be consistent and asymptotically normal. A corresponding modified estimator for the cumulative baseline hazard function is also presented.

e-mail: tsun@stat.missouri.edu

35. Statistics in Genetics

HUMAN QUANTITATIVE TRAIT LOCUS MAPPING WITH SELECTED SAMPLES

Eleanor Feingold*, University of Pittsburgh
Jin P. Szatkiewicz, University of Pittsburgh

Human geneticists have traditionally focused on binary disease traits, but in the last ten years there has been a dramatic increase in the number of studies seeking to map genes for “quantitative” (continuous) traits. This has caused a (slightly lagged) boom in the development of statistical methods for such studies. Most of the new methods assume at some level that the dataset is a population sample from a Gaussian distribution. But population sampling is arguably the exception rather than the rule in human genetic studies. Selected sampling can induce substantial departures from normality, which calls into question the validity and power of many mapping methods. This talk will start by reviewing the most important types of selected samples in human quantitative trait mapping studies, and will then discuss which statistical methods are useful for which types of samples.

e-mail: feingold@pitt.edu

ASSESSING THE POTENTIAL IMPACT OF POPULATION STRATIFICATION AND INBREEDING ON COMMONLY USED STATISTICAL METHODS FOR ASSOCIATION STUDIES

Zhaohai Li*, George Washington University
Joseph Gastwirth, George Washington University
Mitchell Gail, National Cancer Institute, National Institutes of Health

In their seminal article Risch and Merikangas (1997) argued that association studies provide greater power than linkage ones to detect genes underlying complex disease. Other authors have observed that population stratification or inbreeding in the study population increases the false positive rate of the standard methods. Based on C.C. Li's (1969) classic model of the genotype probabilities incorporating the degree, F , of population stratification or inbreeding, we derive explicit formulas for the bias and overdispersion induced by these factors on allele or genotype based tests. When there is population substructure its effect on bias depends on the relationship between the disease risks and the allele frequencies in the subpopulations.

e-mail: zli@gwu.edu

SNP SELECTION AND HAPLOTYPE ANALYSIS IN IDENTIFYING DISEASE GENES

Hongyu Zhao*, Yale University School of Medicine

With the availability of millions of genetic markers in the human genome and the ever-decreasing genotyping costs, it is becoming feasible to genotype tens of thousands of markers throughout the genome in search for genetic variants underlying human diseases. In addition, it is possible to saturate candidate regions with many tightly linked genetic markers. The analysis of such data poses many statistical challenges, including the control of population stratification, the analysis of haplotypes, the relative merit of various genotyping strategies, multiple comparisons, and other statistical issues. In this presentation, we will focus on two issues in the collection and analysis of haplotypes from population-based samples: marker selections and the analysis of case-control data.

e-mail: hongyu.zhao@yale.edu

GENETIC DISSECTION OF COMPLEX TRAITS

Kai Yu *, Washington University School of Medicine

Statistical methods, optimal study designs, and efficient analysis strategies for finding complex disease genes will be reviewed. Issues relating to genome-wide linkage scans and multiple testing will be discussed, identifying the implications of false positives and false negatives. Lumping and splitting approaches for pooling multiple studies and meta-analysis will be discussed. By pooling data and/or results from multiple studies on a given disease/trait, one may attain a large enough sample size to be able to subdivide the data into multiple relatively more homogeneous subgroups. Each subgroup may still have a sufficiently large sample size within each of which the genetic dissection may be more successful. Analyses within subgroups, created using “tree linkage” or other methods, should enhance gene finding. The methods are not yet optimal, but the future holds promise. We believe that such strategies are essential for the genetic dissection of complex traits. After all, for most complex traits, the question is not whether there are genes, only when and how they might be found.

e-mail: rao@wubios.wustl.edu

36. Assessing Efficacy in a Changing Background

HETEROGENEITY OF EFFICACY ASSESSMENT ACROSS SPACE AND TIME: STATISTICAL AND HEALTHCARE UTILIZATION IMPLICATIONS

Demissie Alemayehu*, Pfizer and Columbia University

This talk reviews several statistical problems and healthcare policy issues that arise when clinical trials that assess efficacy are conducted and/or analyzed at different places or at different times. Factors that contribute to the temporal and spatial evolution of treatment effects are identified, and the implications for the design of such experiments, the appropriate analysis, and interpretation and use of the results are discussed. Illustrative examples are given from diverse therapeutic areas, including infectious diseases, cardiovascular and oncology.

e-mail: alemad@pfizer.com

COMPARING HYPERTENSIVE THERAPIES OVER TIME AND PLACE

When many interventions are available for the same condition it is unlikely that definitive randomised trials will be available for all pairwise comparisons. Formal or informal use of indirect comparisons is unavoidable, but there are important concerns about their reliability. I will discuss the new technique of network meta-analysis for summarising and assessing the consistency of information from a large network of randomised trials. A published network meta-analysis of treatments for hypertension will be used to motivate the method and illustrate its benefits and limitations.

e-mail: tlumley@u.washington.edu

LESSONS IN DRUG EFFICACY COMPARISONS FROM METROLOGY

David Banks*, Duke University

The problem of evaluating drug efficacy involves a changing target and networks of comparisons. Specifically, many pathogens evolve immunity, and the efficacy judgments entail historical benchmarks. Thus there might be clinical data on treatment A versus a control in 1980, treatment A versus treatment B in 1990, and treatment B versus treatment C in 2000. This kind of situation is quite similar to issues that arise in metrology, where uncertainty must be propagated across a chain of comparisons. The metrology community has developed various procedures, such as the Mandel bundle-of-lines technique, for modeling chains or networks of comparisons, using standards that are compared at different times and replicated in different ways. This talk describes some of the work in the metrology community and how it can be adapted to the problem of assessing drug efficacy over time.

e-mail: banks@stat.duke.edu

37. The Causes and Consequences of HIPAA: A Statistician's Perspective (HIPAA: Health Insurance Portability and Accountability Act) (Roundtable)

HIPAA DATA STANDARDS AND HEALTH INFORMATION: THE GOOD, THE BAD, AND THE UGLY

Christopher G. Chute*, Mayo Clinic

The scope and eventual spectrum of HIPAA data standards encompass vastly more than privacy rules, which understandably dominate researchers' discussions. Explicitly or implicitly coupled with the HIPAA standards that impose new restrictions are data interchange standards which purport to enhance data comparability, exchange, and interoperability (albeit within privacy regulations). An overview of the 'other HIPAA' standards, including transactions codes, planned coding systems such as ICD-10-CM, SNOMED-CT, RxNorm, and LOINC, together with standardized electronic health record standards under federally sponsored development by the Institute of Medicine and HL7, paint a potentially bountiful future for health statistics operations. Anticipating future health data standards in present data collection methods or clinical trials designs may finally permit a HIPAA dividend for research involving health data.

e-mail: chute@mayo.edu

PRACTICAL ISSUES OF SHARING DATA AND BIOLOGICAL SPECIMENS UNDER HIPAA

Gerald J. Beck*, Cleveland Clinic Foundation

HIPAA stipulates that Protected Health Information (PHI), including research data and specimens, being transferred outside the Covered Entity must address patient privacy and confidentiality. In order to comply with the HIPAA regulation, sharing data and biological specimens among investigators not at the same institution have become a challenge. Patient privacy and confidentiality must be addressed to be in compliance. This presentation will describe the impact of these rules and strategies for implementing them, particularly as they apply to NIH Data Coordinating Centers of multi-center studies. During the course of these studies, and certainly after the study ends, there is a need to make the data available to outside investigators. Also, many of these studies include central repositories for the storage of biological specimens (e.g., tissue, blood, DNA, etc.). The issues involved and the procedures developed for several studies to meet HIPAA guidelines will be presented.

e-mail: gbeck@bio.ri.ccf.org

SHARING AND PUBLICATION OF DATA ON INDIVIDUAL PATIENTS: SCIENTIFIC, ETHICAL AND REGULATORY ISSUES

Andrew J. Vickers*, Memorial Sloan-Kettering Cancer Center

The ease with which statisticians can share raw data on individual patients has been dramatically altered by changes in technology, allowing electronic transfer of data from peer to peer and publication of entire data sets on the Internet. Debate on data sharing and publication has lagged behind these technological innovations. I will argue that there are compelling scientific and ethical reasons to promote sharing and publication of data, and whilst ethical and regulatory safeguards are essential, they must be seen to be just one aspect of the data sharing process. Publication and sharing of raw research data offers tremendous promise but appears to be relatively rare: indeed, it appears commonplace for statisticians to refuse to share data without giving a reason and it appears likely that privacy concerns related to HIPAA with further dampen data exchange. It seems self-evident that data belong to patients, not to investigators, and that data produced with public funds should be shared for the public good. Nonetheless, only in late 2003 did the NIH mandate data sharing for funded projects, and even then, the policy applies only to large grants.

e-mail: vickersa@mskcc.org

38. Information Related Problems in Biology

INFORMATIVENESS OF GENETIC MARKERS FOR INFERENCE OF ANCESTRY

Noah A. Rosenberg, University of Southern California

Lei Li, University of Southern California

Ryk Ward, University of Oxford

Jonathan K. Pritchard*, The University of Chicago

Inference of individual ancestry is useful in various applications in genetics, including admixture and case-control mapping of disease genes. Using information-theoretic principles, we introduce a general measure, the informativeness for assignment, applicable to any number of potential source populations, for determining the amount of information that multiallelic markers provide about individual ancestry. In a worldwide human microsatellite data set, we identify markers of highest informativeness for regional ancestry, and for population ancestry within regions, and we describe the properties of such markers. Our results can aid in decisions about the type, quantity, and specific choice of markers for use in studies that require estimates of genetic ancestry.

e-mail: pritch@uchicago.edu

INFORMATION-DRIVEN MARKER SELECTION FOR LARGE SCALE GENOMIC STUDIES ON COMPLEX TRAITS

Tian Zheng*, Columbia University
Shaw-Hwa Lo, Columbia University

Many current studies on complex traits involve large number of markers. Before any valid detailed study can be performed, a screening procedure must be taken, where markers contributing no information but noise are identified and excluded. The marker-by-marker strategy used in most current efforts resulted in the loss of valuable information regarding the functional interaction among the genomic loci. This is especially important when dealing with traits due to epistatic genes, each of modest effect. Genes that contribute substantially to an interactive control of a trait might be missed by marker-by-marker methods due to their weak marginal effects. We present here a selection procedure based on an information-driven screening algorithm. The marker selection procedure starts with random subsets of the large marker set (thousands of loci). For each subset, the screening algorithm deletes noninformative markers from this 'local' set and drives the total set transmission disequilibrium information to a 'maximum', and return the markers retained as the 'local' important set. Markers with significant high return frequencies are selected for further detailed studies. The properties of this selection procedure will be demonstrated using simulated complex disease models.

e-mail: tz33@columbia.edu

A BAYESIAN FRAMEWORK FOR QUANTIFYING INCOMPLETE INFORMATION IN STATISTICAL AND GENETIC HYPOTHESIS TESTING

Augustine Kong, deCode Genetics, Iceland
Xiao-Li Meng*, Harvard University
Dan Nicolae, The University of Chicago

A basic approach in genetic linkage analysis is to test statistically whether a genetic marker is linked to a disease susceptibility gene. A common problem is that most genetic data sets are incomplete and investigators need to know how much information is available in the observed data for a chosen test, relative to the amount of information that would have been available if the data were complete. This relative information can be critical for the investigator's follow-up strategies (e.g., using more genetic markers with existing DNA samples versus collecting DNA samples from additional families). Reliably measuring the relative information in such studies turns out to be challenging, mainly because there are several competing criteria that a desirable measure should satisfy. For large samples, an effective likelihood-based framework is available. For small samples (e.g., with individual families in a genetic pedigree), a Bayesian approach appears to be necessary. This talk proposes to use a posterior measure of the flatness of the likelihood surface to quantify the relative information. The measure has all the desirable properties, but it requires a sensible prior specification. A possible empirical approach to deal with this problem will be discussed.

e-mail: meng@stat.harvard.edu

BAYESIAN ANALYSIS FOR GENERALIZED LINEAR MODELS WITH NONIGNORABLY MISSING
COVARIATES

Lan Huang*, University of Connecticut
Ming-Hui Chen, University of Connecticut
Joseph G. Ibrahim, University of North Carolina

We propose Bayesian methods for estimating parameters in generalized linear models (GLM's) with nonignorably missing covariate data. We show that when improper uniform priors are used for the regression coefficients of the multinomial selection model for the missing data mechanism, the resulting joint posterior will always be improper regardless of whether proper or improper priors are specified for the regression parameters, of the GLM or the parameters of the covariate distribution. To overcome this problem, we propose a novel class of proper priors for the regression coefficients in the selection model for the missing data mechanism. In addition, we propose the weighted L measure and extend the Deviance Information Criterion for assessing whether the missing data mechanism is ignorable or nonignorable. A novel Markov chain Monte Carlo sampling algorithm is also developed for carrying out posterior computation. Several simulations are given to investigate the performance of the proposed Bayesian criteria as well as the sensitivity of the prior specification. Real datasets from a melanoma cancer clinical trial and a liver cancer study are presented to further illustrate the proposed methods.

e-mail: lan@merlot.stat.uconn.edu

SEMIPARAMETRIC MAXIMUM LIKELIHOOD FOR MISSING COVARIATES IN PARAMETRIC REGRESSION

Zhiwei Zhang*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh

Parametric regression models are widely used in biometrics research. This paper is concerned with statistical inference under such models with some covariates missing at random. If the always observed covariates are discrete or can be discretized, we propose a semiparametric maximum likelihood method which requires no parametric specification of the selection mechanism or the covariate distribution. Simple conditions are given under which the semiparametric maximum likelihood estimator (MLE) exists. For ease of computation, we also consider a restricted MLE which maximizes the likelihood over covariate distributions supported by the observed values. The two MLEs are asymptotically equivalent and strongly consistent for a class of topologies on the parameter set. Upon normalization, they converge weakly to a zero-mean Gaussian process in a suitable space. The regression parameter estimate, in particular, achieves the semiparametric information bound, which can be consistently estimated by perturbing the profile log-likelihood. Furthermore, the profile likelihood ratio statistic is asymptotically chi-squared. An EM algorithm is proposed for computing the restricted MLE and for variance estimation. Simulation results suggest that the proposed method performs reasonably well in moderate-sized samples. An application is described.

e-mail: zhzst5@pitt.edu

SEMIPARAMETRIC MODELS FOR MISSING COVARIATES IN GENERALIZED LINEAR MIXED MODELS

Qingxia Chen*, University of North Carolina at Chapel Hill

Joseph Ibrahim, University of North Carolina at Chapel Hill

We consider a class of semiparametric models for the covariate distribution and missing data mechanism for missing covariate and/or response data for generalized linear models and generalized linear mixed models. Ignorable and Nonignorable missing covariate and/or response data is considered. A generalized additive model (GAM) is considered for the covariate distribution and/or the missing data mechanism. Penalized regression splines are used to express the GAM's as a generalized linear mixed effects model, in which the variance of the corresponding random effects provides an intuitive index for choosing between the semiparametric and parametric model. Maximum likelihood estimates are obtained via the EM algorithm. Simulations are given to demonstrate the methodology and two real datasets from a melanoma cancer and breast cancer clinical trials are analyzed using the proposed methods.

e-mail: qchen@bios.unc.edu

INFERENCE ABOUT THE PROPENSITY PENALIZED SPLINE PREDICTION METHOD FOR MULTIVARIATE DATA WITH MISSING VALUES

Hyonggin An*, University of Michigan

Roderick J. Little, University of Michigan

Little and An (2003) developed a robust model-based method for handling the multivariate missing data, that deals with the so-called “curse of dimensionality”. In their approach, they modeled the variables with missing values as a penalized spline of a particular function of the covariates most sensitive to model misspecification, namely the response propensity score. They called this method a propensity penalized spline prediction (PPSP). The main idea of PPSP is to focus on specifying the relationship with the propensity score correctly, since misspecification of that relationship leads to bias. In this presentation, we discuss variance estimation methods for the PPSP estimator. Since the PPSP estimator is based on a single imputation method, variance estimates incorporate the imputation uncertainty. In addition, we need to account for the variability due to the estimation of response propensity scores. We develop a bootstrap method and a multiple imputation method for variance estimation that take these additional sources of variability into account. We also conduct a simulation study to examine the performance of these variance estimation methods.

e-mail: hyongg@umich.edu

SEMPARAMETRIC BAYESIAN ANALYSIS OF MATCHED CASE CONTROL STUDIES WITH MISSING EXPOSURE

Samiran Sinha*, University of Florida
Bhramar Mukherjee, University of Florida
Malay Ghosh, University of Florida
Bani K. Mallick, Texas A&M University
Raymond J. Carroll, Texas A&M University

The paper considers Bayesian analysis of matched case-control problems when one of the covariates is partially missing. Within the likelihood context, the standard approach to this problem is to posit a fully parametric model among the controls for the partially missing covariate as a function of the covariates in the model and the variables making up the strata: sometimes, the strata effects are ignored at this stage. Our approach differs not only in that it is Bayesian, but far more importantly in the manner in which we treat the strata effects. We assume a Dirichlet process prior with a base measure normal distribution for the stratum effects and estimate all the parameters in a Bayesian framework. Two matched case-control examples and a simulation study are considered to illustrate our methods and the computing scheme.

e-mail: ssinha@stat.ufl.edu

IMPUTATION OF MISSING GENOTYPIC AND BIOMARKER DATA TO IMPROVE A RISK MODEL FOR LUNG CANCER

Carol J Etzel*, M.D. Anderson Cancer Center
Mei Lu, M.D. Anderson Cancer Center
Hui Zhou, M.D. Anderson Cancer Center
Qingy Wei, M.D. Anderson Cancer Center
Xifeng Wu, M.D. Anderson Cancer Center
Margaret Spitz, M.D. Anderson Cancer Center

Our goal is to develop a multiple imputation procedure to impute missing genotypic (categorical) and biomarker (continuous) data to assess the risk for lung cancer while improving power to detect risk factors and precision to estimate the effects of these factors while controlling for type 1 error associated with non-risk factors. Using a sample of 278 newly diagnosed lung cancer patients recruited from UT M. D. Anderson Cancer Center, Houston and 316 healthy controls that had complete data on two biomarkers (Bleomycin-induced chromosome breaks and DNA repair capacity) as well as a panel of markers from three metabolic polymorphic gene families, we generated data sets with missing data. Using PROC MI in SAS, we formed complete data sets via multiple imputation and evaluated risk for lung cancer using a logistic model. We observed that this method performed adequately for the continuous biomarkers, but was not optimal for the categorical genotypes. However, this method did not allow us to incorporate prior information for any of the genetic variables, which could enhance the imputation of the missing data. Our next step is to implement a Bayesian-based imputation method in order to incorporate population-level (prior) information for the genetic variables into our modeling.

e-mail: cetzel@mdanderson.org

CONDITIONAL INFERENCE METHODS FOR POISSON REGRESSION MODELS WITH MISSING DATA AND FEEDBACK

Jason Roy*, University of Rochester
Don Alderson, Brown University
Joseph W. Hogan, Brown University
Karen Tashima, Brown University

A strategy to account for serial correlation in longitudinal studies is to assume observations are independent, conditional on unit-specific nuisance parameters. One method for estimation of these models is maximum likelihood, where the nuisance parameters are referred to as random effects and are intergrated out of the likelihood. An alternative is conditional inference, where we eliminate the nuisance parameters from the likelihood by conditioning on a sufficient statistic. An advantage of conditional inference methods over parametric random effects models is all patient-level time-invariant factors (both measured and unmeasured) are accounted for in the analysis. A limitation with standard conditional inference methods is it assumes missing data are missing completely at random and does not allow 'feedback.' We develop new estimation strategies that allow for feedback and missing at random dropouts. We apply the methods to estimated the effect of protease inhibitor (PI) use on the rate of emergency department visits among HIV-infected women.

e-mail: jason_roy@URMC.rochester.edu

40. Survival Analysis I

ANALYSIS OF TIME-TO-EVENT DATA WITH INCOMPLETE EVENT ADJUDICATION

Thomas D. Cook*, University of Wisconsin, Madison
Michael R. Kosorok, University of Wisconsin, Madison

In many multicenter, randomized clinical trials, the primary outcome is the time to the first of a number of possible clinical events. An event classification committee (ECC) may be convened to determine whether events which have been reported by investigators meet the predetermined criteria for primary endpoint events. When interim analyses are performed in such trials, the final classification for many reported events will not be known. Failure to account for the uncertain status of these events may result in incorrect interim analysis. The probability that an unadjudicated event will be confirmed as a primary event can typically be estimated from those events for which adjudication is complete. We show that if each unadjudicated event is weighted according to the probability that it will be the first primary event, consistent estimates of survival probabilities and regression parameters, and unbiased logrank tests of treatment differences can be performed. Moderate sample consistency of point estimates and variance estimates is verified by simulation.

e-mail: cook@biostat.wisc.edu

PARAMETRIC ESTIMATION OF CUMULATIVE INCIDENCE FUNCTION

Jong-Hyeon Jeong*, University of Pittsburgh

In clinical data collected from phase III clinical trials to treat breast cancer, a recurrence of the original cancer, a new primary cancer, or a death are often observed as multiple events per patient. Radiation oncologists, however, may be interested in comparing patterns of recurrences alone to identify a subgroup of patients who need to be treated by a radiation therapy after a surgery. The cumulative incidence function provides unbiased estimates in this case for the cumulative probability of recurrences occurring in the presence of other competing events such as a new primary cancer or a death. In this talk, a new flexible parametric inference procedure is introduced to improve efficiency of the nonparametric estimates of the cumulative incidence function. This parametric procedure will be applied to several datasets from the National Surgical Adjuvant Breast and Bowel Project (NSABP).

e-mail: jeong@nsabp.pitt.edu

FAILURE TIME ANALYSIS OF HIV VACCINE EFFECTS ON VIRAL LOAD AND TREATMENT INITIATION

Peter B. Gilbert, Fred Hutchinson Cancer Research Center and University of Washington
Yanqing Sun*, University of North Carolina at Charlotte

Classically designed Phase III vaccine efficacy trials allow evaluation of a vaccine effect that may lower susceptibility of vaccinees to acquire HIV infection, but do not allow direct evaluation of vaccine effects on secondary transmission and/or progression to clinical disease. We consider a surrogate endpoint of disease/transmission, the level of plasma HIV RNA measured post-HIV infection, to assess the later effects. The assessment of viral load, however, is complicated by the fact that some trial participants will likely receive antiretroviral therapy (ART) following the diagnosis of HIV infection. A composite endpoint is considered, which is defined as the first occurrence of virologic failure (a rise in HIV viral load above a pre-specified failure threshold $x_{\{v1\}}$ copies/ml) or initiation of ART, whichever occurs first. The vaccine efficacy parameter, $VE(\tau, x_{\{v1\}})$, is defined as one minus the ratio of cumulative probabilities of the composite endpoint (vaccine/placebo) occurring by τ months post infection diagnosis. We develop a general procedure for constructing simultaneous confidence bands that applies to both cases of $x_{\{v1\}}$ spanning discrete levels and spanning a continuous range. The methods are applied to analyze data from the world's first HIV vaccine efficacy trial conducted by VaxGen.

e-mail: yasun@uncc.edu

A SAMPLE SIZE FORMULA FOR THE SUPREMUM LOG-RANK STATISTIC

Kevin H. Eng, Brown University
Michael R. Kosorok*, University of Wisconsin-Madison

An advantage of the supremum log-rank over the standard log-rank statistic is an increased sensitivity to a wider variety of stochastic ordering alternatives. In this paper, we develop a formula for sample size computation for studies utilizing the supremum log-rank statistic. The idea is to base power on the proportional hazards alternative, so that the supremum log-rank will have the same power as the standard log-rank in the setting where the standard log-rank is optimal. This results in a slight increase in sample size over that required for the standard log-rank. For example, a 5.733 percent increase occurs for a two-sided test having type-I error 0.05 and power 0.80. This slight increase in sample size is offset by the significant gains in power the supremum log-rank test achieves for a wide range of non-proportional hazards alternatives. These results should facilitate the wider use of the supremum log-rank statistic in clinical trials.

e-mail: kosorok@biostat.wisc.edu

POWER CONSIDERATION IN SURVIVAL ANALYSIS WITH UNCERTAIN ENDPOINTS

Hongwei Wang*, Rutgers University
Cong Chen, Merck Research Laboratories
Steven M. Snapinn, Merck Research Laboratories

In endpoint clinical trials, uncertainty in classifying endpoints may dilute the observed treatment difference. The conventional procedure is to classify each potential endpoint as true or not, and use the confirmed endpoints in a Cox-regression analysis. Snapinn (1998) proposed a weighted method to incorporate information on endpoint uncertainty into a modified Cox-regression analysis. We explore the theoretical properties of the two approaches, and develop closed-form expressions for power calculations. The predicted power is close to the power empirically estimated by Snapinn (1998) in simulation studies, and it is further confirmed that the weighted method is more powerful in detecting a treatment difference compared to the conventional one. This gives us strong confidence to implement it in a planned cardiovascular outcome study.

e-mail: hwang@stat.rutgers.edu

JOINT MODELING SURVIVAL AND RECURRENT EVENT DATA

Xuelin Huang*, The University of Texas
M.D. Anderson Cancer Center

Patients receive different treatments after their disease recurrences. It is necessary to account for the effects of treatments after recurrences (secondary treatments) when analyzing the effects of initial treatments. It is not appropriate to use secondary treatments as time-dependent covariates in a Cox proportional hazards model since only those patients whose disease recur (i.e., in worse conditions) receive secondary treatments. In this talk, I will propose and compare two approaches dealing with this problem. The first one is an extension of the method by Lunceford, Davidian and Tsiatis (2002, *Biometrics* 58, 48-57). The second approach uses a bivariate frailty model. A retrospective study of soft tissue sarcoma is presented to illustrate the methods.

e-mail: xlhuang@mdanderson.org

SEMIPARAMETRIC EFFICIENT ESTIMATION OF SURVIVAL DISTRIBUTION IN TWO-STAGE RANDOMIZATION DESIGNS

Abdus S. Wahed*, University of Pittsburgh
Anastasios A. Tsiatis, North Carolina State University

Two-stage randomization designs are common in evaluation of combination therapies where patients are initially randomized to an induction therapy and then depending upon their response and consent, are randomized to a maintenance therapy. Comparison of all possible treatment strategies to find the best combination that leads to the largest patient benefit is the main goal of such trials. Since the second randomization is done only among those who responds, usual survival analysis techniques do not apply. Lunceford et. al. (2002, *Biometrics*, 58, 48-57) introduced consistent estimators for the survival distribution under different treatment policies. Wahed and Tsiatis (2003) presented the most efficient regular estimator when the data is complete. In this paper we derive the most efficient estimator when the data are subject to censoring and show how to improve efficiency further by taking into account additional information from auxiliary variables. Large sample properties of the proposed estimator are derived and comparisons with other estimators are made using simulation. We apply our estimators to a leukemia clinical trial data set that motivated this study.

e-mail: wahed@pitt.edu

A COMPARATIVE STUDY OF THE USE OF GAM AND GLM IN AIR POLLUTION RESEARCH

Shui He*, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh
Vincent C. Arena, University of Pittsburgh

GAMs (Generalized Additive models) were used as a standard analytic tool in time-series studies of air pollution and health during the last decade. Currently, concerns were raised in the air pollution research community related to the use of GAMs in the assessment of pollution-health outcome associations. Concurvity is one of these concerns. Previous studies and our results suggest that statistical software such as S-plus can seriously overestimate the fitted model parameters and underestimate their variances of GAM in the presence of concurvity. We explore an alternate class of models, the class of generalized linear models with natural cubic splines (GLM+NS), that may not be affected by concurvity. We make a systematic comparison between GLM+NS and GAM with smoothing splines (GAM+S) in the presence of varying degrees of concurvity. Our results suggest that GLM+NS performs better than GAM+S with medium-to-high concurvity in the data.

e-mail: shh10@pitt.edu

RELATING PM_{2.5} CONCENTRATION LEVELS TO MORTALITY USING AN EXPOSURE SIMULATOR

Catherine Calder*, Ohio State University;
Christopher Holloman, Battelle Memorial Institute;
Steven Bortnick, Battelle Memorial Institute;
Warren Strauss, Battelle Memorial Institute;
Michele Morara, Battelle Memorial Institute

Since the EPA began widespread monitoring of PM_{2.5} concentration levels in the late 1990s, the epidemiological community has performed several observational studies relating PM_{2.5} levels directly to various health endpoints including mortality and morbidity. Recent research, however, suggests that human exposure to the constituents of PM_{2.5} may differ significantly from ambient (or outdoor) PM_{2.5} levels measured by monitors since people spend a great deal of time in various indoor environments. To address this concern, we propose a three-stage Bayesian hierarchical model as an alternative to the Poisson Generalized Additive Model (GAM) that is traditionally used to characterize the relationship between PM_{2.5} levels and health endpoints. Our approach includes a spatial model relating monitor readings to average county PM_{2.5} levels and an exposure simulator which links average ambient PM_{2.5} levels to average personal exposure using activity pattern data. We apply our model to a study population in North Carolina where we explore the effect of PM_{2.5} exposure on cardiovascular mortality over a three-year period.

e-mail: calder@stat.ohio-state.edu

A SIMPLE STATISTICAL AIR DISPERSION MODEL AND ITS PREDICTION

Tae-Young Heo*, North Carolina State University
Jacqueline M. Hughes-Oliver, North Carolina State University

We suggest a method for combining air dispersion models with statistical models in order to test, evaluate, and reduce uncertainty due to point source effects. In environmental science, there are a variety of computational air dispersion models. These models usually only provide deterministic predictions or estimates without measures of prediction or estimation uncertainty. By introducing error components in air dispersion models, we obtain improved prediction of concentrations, we reduce prediction uncertainty, and we are able to statistically test the impact of a point source. Application is made to three real datasets on experimental trace fields, namely, the Kincaid, Indianapolis, and Prarie Grass projects.

e-mail: theo@unity.ncsu.edu

BIOLOGICALLY BASED MODEL FOR ESTIMATING THE FORMATION RATE OF BIOMARKER FROM BUTADIENE AIR EXPOSURE: FREQUENTIST AND BAYESIAN APPROACHES

Sohee Park*, Harvard School of Public Health;
Louise M. Ryan, Harvard School of Public Health

We present a biologically motivated model to estimate the formation rate of hemoglobin adducts (biomarker) from air exposure to Butadiene. This calibration model expresses the change in hemoglobin adducts at time $t-n$ and t as the weighted sum of daily air exposure to Butadiene during a time interval of n days, while accounting for the formation and removal rate (kinetics) of red blood cells. We will compare the results from a classical mixed modeling approach and a Bayesian modeling approach.

e-mail: shpark@hsph.harvard.edu

LINEAR REGRESSION WITH TYPE I INTERVAL- AND LEFT-CENSORED DATA

Mary-Lou Thompson, University of Washington
Kerrie P. Nelson*, University of South Carolina

Laboratory analysis in a variety of contexts may result in left- and interval-censored measurements. We develop and evaluate a maximum likelihood approach to linear regression analysis in this setting and compare this approach to commonly used simple substitution methods. We explore via simulation the impact on bias and power of censoring fraction and sample size in a range of settings. The maximum likelihood approach represents only a moderate increase in power, but we show that the bias in substitution estimates may be substantial.

e-mail: kerrie@stat.sc.edu

42. Longitudinal Data Analysis I

A RESTRICTED MULTIVARIATE RANDOM-EFFECTS MODEL WITH APPLICATIONS TO ASSESSING RADIATION THERAPY FOR BRAIN TUMORS

Hongbin Fang *, University of Maryland;
Guoliang Tian, University of Maryland;
Xiaoping Xiong, St. Jude Children's Research Hospital;
Ming Tan, University of Maryland

In clinical studies, multiple endpoints are often measured for each patient longitudinally. The multivariate random-effects or random coefficient model has been a useful method for analysis. However, medical research problems may impose restrictions on the parameters of interests. For example, in a pediatric brain tumor study on radiation therapy, there is a natural ordering in the white matter relaxation time of brain tissues among different regions surrounding the primary tumor. Such parameter constraints should be accounted for in the analysis. This article proposes a class of multivariate random coefficient models with restricted parameters and derives its maximum likelihood estimates. We propose a modified EM algorithm for the quadratic optimization with linear inequality constraints necessary in deriving the maximum likelihood estimates. The method is applied to analyzing the pediatric brain tumor study.

e-mail: hfang@umm.edu

EXTENDING HISTORICAL FUNCTIONAL LINEAR MODEL USING PENALIZED REGRESSION SPLINES

Jaroslaw Harezlak*, Harvard School of Public Health
Nan M. Laird, Harvard School of Public Health

Functional data arise in many applications including human growth curves and continuous monitoring of biological functions of an organism. Current literature covers many applications where either predictors or responses are considered to be functions. However, there are relatively few methods available for the case where both predictors and responses are in functional forms. Malfait and Ramsay (2003) proposed a historical functional linear model for regression modelling of a functional response with a functional predictor. We extend their method by regularizing the regression surface fit using penalized regression splines. We also consider dependence of a response function on additional non-functional covariates. The method is applicable when individual functions are sampled at irregularly spaced time points. We compare the longitudinal and functional views of the regression problem. We illustrate our extension with a simulation study and apply it to the analysis of effects of pollution on the heart rate variability in an occupational setting.

e-mail: jharezla@hsph.harvard.edu

MODELING UNEQUALLY SPACED BIVARIATE GROWTH CURVE WITH A KALMAN FILTER APPROACH

Qianyu Dang *, University of Pittsburgh;
Stewart J. Anderson, University of Pittsburgh;
Sati Mazumdar, University of Pittsburgh

In many clinical studies, patients are followed over time with their responses measured longitudinally. Using mixed model theory, one can characterize these data using a wide array of across subject models. A state-space representation of the mixed effects model and use of the Kalman filter allows one great flexibility in choosing the within error correlation structure even in the presence of missing or unequally spaced observations. Furthermore, using the state-space approach, one can avoid inverting large matrices resulting in efficient computation. The approach also allows one to make detailed inference about the error correlation structure. We consider a bivariate situation where the longitudinal responses are unequally spaced and assume that the within subject errors follows a continuous first order autoregressive (CAR(1)) structure. Since a large number of nonlinear parameters need to be estimated, the modeling strategy and numerical techniques are critical in the process. We developed both a Visual Fortran® and a SAS® IML program for modeling such data. A simulation study was conducted to investigate the robustness of the model assumptions. We also use data from a psychiatric study data to demonstrate our model fitting procedure.

e-mail: qidst1@pitt.edu

A LATENT CLASS REGRESSION MODEL FOR LONGITUDINAL ORDINAL OUTCOMES

Don Hedeker*, University of Illinois at Chicago

A mixed-effects logistic regression model is described for analysis of repeated ordinal outcomes. Random effects are included in the model to account for the correlation of the repeated observations, however no distributional form is assumed for these random effects. Instead, by specifying the number of node points for the random-effects distribution, the values of the latent nodes and the density of the distribution at these nodes are estimated along with the model parameters. The nodes indicate the latent classes and so the model is quite similar to latent growth models, though in the proposed model covariates can also be used to determine these classes directly or the classes can be estimated adjusting for the covariates. For implementation, a full maximum likelihood solution is described. The proposed model will be compared to models assuming normally distributed random effects, as well as a semi-parametric model in which the values of the node points are not estimated but fixed in advance. Using data from an adolescent smoking study, analyses are presented that examine smoking change across time and the influence that peer smoking might play in this development.

e-mail: hedeker@uic.edu

LATENT PREDICTORS OF REPEATED CLINICAL MEASURES – APPLICATION TO BEHAVIORAL RISK FACTORS AND LUNG FUNCTION AMONG YOUTH WITH CYSTIC FIBROSIS

Melanie M. Wall * , University of Minnesota;
Joan Patterson, University of Minnesota

Generalized linear mixed models are well developed and suited for modeling repeated measurements. Observed predictors (e.g. age, weight) can be included very straightforwardly into the model for the mean structure or for the covariance structure of the random effects. But often there are predictors of interest that cannot be observed directly, in particular latent variables that are measured via some battery of indicators or questionnaire items which are best modeled using a factor analysis framework. This talk presents a repeated measures model that includes a factor analysis framework for incorporating latent predictors. The model is applied to data collected on youth with cystic fibrosis (CF). Researchers are interested in understanding the reasons for differences in the decline of lung function for CF patients. Because regular, aggressive daily treatment (including chest physiotherapy) can forestall pulmonary function deterioration, researchers hypothesize that certain risk-taking behaviors (e.g. cough suppression) and parental-child connectedness may be influential factors. Risk taking behaviors, personal and family factors are latent variables which were measured by a 40-item self report questionnaire. The model developed in this talk includes these latent factors as predictors for the repeated measures of lung function.

e-mail: melanie@biostat.umn.edu

A LATENT CHANGEPOINT MODEL USING A GENERALIZED ESTIMATING EQUATIONS APPROACH

Ji-Hyun Lee*, University of South Florida
Bahjat F. Qaqish, University of North Carolina at Chapel

We propose a latent changepoint model for the analysis of longitudinal biomarker data in relation to progression or recurrence of disease. A parametric model that contains a random changepoint in the expected biomarker values is considered. In this paper, estimation through generalized estimating equations is proposed. The procedure allows estimation of the biomarker trend over time and the changepoint distribution. We provide the details of the estimation procedure. Through a Monte Carlo simulation study, several aspects of the small sample performance of the estimates are investigated.

e-mail: LeeJ@moffitt.usf.edu

43. Clustering and Classification Methods

CORRECTING THE LOSS OF CELL-CYCLE SYNCHRONY IN CLUSTERING ANALYSIS OF MICROARRAY DATA USING WEIGHTS

Fenghai Duan*, Yale University;
Heping Zhang, Yale University

Due to the existence of the loss of synchrony in cell cycle data sets (e.g., yeast cell cycle data sets studied by Cho et al. (1998) and Spellman et al. (1998)), the standard clustering methods (e.g., k-means) that group ORFs (Open Reading Frames) based on similar expression levels are deficient unless the temporal pattern of the expression levels of the ORFs is taken into account. We propose a method to improve the performance of the k-means method by assigning a decreasing weight on its variable level and evaluating the λ -weighted k-means on Spellman et al.'s (1998) yeast cell cycle data set. The protein complexes in a public website are used as biological benchmarks. To compare the k-means clusters with the structures of the protein complexes, we measure the agreement between these two ways of clustering via the adjusted rand index. Our results show an exponential decreasing weight function that we assign to the variable level of k-means generally increases the agreement between protein complex and k-means clusters. We derive the temporal weight function through a Fourier transformation of the expression data and our results show that per-minute-synchrony-losing rate in the function is near the length of two cell cycles, consistent with the biological evidence.

e-mail: f.duan@yale.edu

AN UP AND DOWN REGULATION DIRECTED CLUSTERING OF TEMPORAL GENE EXPRESSIONS

Lang Li*, Indiana University
Stephen J. Metthew, Indiana University
Howard Edenberg, Indiana University

A two step clustering algorithm is proposed to cluster temporal gene expressions. This algorithm is driven by the hypothesis test at step one. The threshold is chosen such that 95% of null genes will be classified to the null profile. Only candidate profiles that have statistically significantly more genes than by chance alone will be further evaluated at step two, and false discovery rates for significant candidate profiles are calculated. At step two, overlapping probability matrix among the significant candidate profiles is generated through a parametric bootstrap, and it is employed to merge candidate profiles through our newly developed clustering algorithm. The performance of this clustering algorithm is demonstrated through both a real data analysis and statistical simulations.

e-mail: lali@iupui.edu

BUILDING MIXTURE TREES FROM BINARY SEQUENCE DATA

Shu-Chuan Chen * , National Health Research
Institutes Taiwan, ROC; Bruce G. Lindsay, Penn State University

Clustering methods have been broadly investigated in the last decade. Since the rapid progress of human genome sequencing, more efficient clustering methods are demanded. For example, a map of 1.42 million single nucleotide polymorphisms(SNPs) at a density of 1.91 kilobases is found by the international SNP map working group (Sachidanandam et al., 2001). Such an SNP map can be a public resource for defining nucleotide variation among human genome which is believed to be useful in identifying disease genes. In this talk, an ancestral mixture model is proposed for clustering binary sequences. This model has a natural relationship to the coalescent process of population genetics. The sieve parameter in the model plays an important role of time in the evolutionary tree of the sequences. By sliding the sieve parameter, one can create a hierarchical tree. Some properties of the proposed model will be discussed.

e-mail: shu@stat.psu.edu

CATS: CLUSTERING AFTER TRANSFORMATION AND SMOOTHING

Nicoleta Serban*, Carnegie Mellon University
Larry Wasserman, Carnegie Mellon University

CATS — Clustering After Transformation and Smoothing — is a technique for nonparametrically estimating and clustering a large number of curves. Our motivating example is a genetic microarray experiment but the method is very general. The method includes: transformation and smoothing multiple curves, multiple nonparametric testing for trends, clustering curves with similar shape, and nonparametrically inferring the misclustering rate.

e-mail: nserban@stat.cmu.edu

SETS OF COMPLEMENTARY LACK OF FIT TESTS BASED ON EDGE CLUSTERS

James W. Neill *, Kansas State University;
Forrest R. Miller, Kansas State University

In earlier work, we presented a maximin power clustering criterion for use with Christensen's tests for assessing regression function adequacy (Miller et al. 1998, 1999 and Neill et al. 2002). This previous work specifically addressed the problem of cluster selection for detecting between-cluster lack of fit. When applying maximin methods to the problem of detecting within-cluster lack of fit, one is led to consider edge clusters of graphs on the predictor variables. Tests based on edge clusters have the advantage of providing good power for the circumstance in which there exists between- or within- cluster lack of fit or a mixture of the two types of lack of fit, leading to the consideration of complementary tests. We have developed a powerful testing procedure which involves doing several sets of complementary tests, each based on a different edge clustering. For the case of nonreplication, a collection of edge clusterings can be determined from graphs based on families of parallel trend tubes in predictor space. The idea is to detect lack of fit along the tubes. In the case of replication, one obtains edge clusters consistent with the graph determined by the replicates. That is, there is an edge between every pair of replicate predictor settings.

e-mail: jwneill@stat.ksu.edu

A BAYESIAN MIXTURE MODEL RELATING DOSE TO CRITICAL ORGANS AND FUNCTIONAL COMPLICATION IN 3D CONFORMAL RADIATION THERAPY

Timothy D. Johnson^{*}, University of Michigan;
Jeremy M.G. Taylor, University of Michigan

A goal of radiation therapy is to deliver maximum dose to the target tumor while minimizing complications due to irradiation of critical organs. Technological advances in 3D conformal radiation therapy has allowed great strides in realizing this goal, however complications may still arise. Several mathematical models have been proposed that describe the relationship between dose and observed functional complication, however only a few published studies have successfully fit these models to data using modern statistical methods which make efficient use of the data. One complication following radiation therapy of head and neck cancers is the patient's inability to produce saliva. Although this may seem innocuous to the lay-person, xerostomia (dry mouth) leads to high susceptibility of oral infection. We present a dose-damage-injury model that can accommodate any of the various mathematical models relating dose to damage. The model is a non-linear, longitudinal mixed effects model where the outcome (saliva flow rate) is modeled as a mixture of a point mass at zero and a gamma distribution whose mean depends on both time and dose. Bayesian techniques are used to estimate the relationship between dose delivered to the parotid and the outcome—saliva flow rate.

e-mail: tdjtdj@umich.edu

HYPOTHESIS TESTING IN MIXTURE REGRESSION MODELS

Hongtu Zhu^{*}, Columbia University
Heping Zhang, Yale University

A difficult question commonly arising from many applications (e.g., human genetics, epidemiology) is whether the data at hand come from a homogeneous or heterogeneous population. Finite mixture models have been widely used to analyze the data from a heterogeneous population. In this talk, we establish asymptotic theory for both the maximum likelihood and maximum modified likelihood estimates in a class of mixture regression models. Moreover, under some mild conditions, we show that the best possible rate for estimating the mixing distribution is achievable for both the maximum likelihood and maximum modified likelihood estimates. We also derive the asymptotic distributions of two log-likelihood ratio testing statistics for testing homogeneity and propose a resampling procedure for approximating the p-value. Some simulation studies are conducted to investigate the empirical performance of the two testing statistics. Finally, two real data sets are analyzed to illustrate the applications of our theoretical results.

e-mail: hz2114@columbia.edu

A NON-LINEAR MIXED MODEL OF PARKINSON'S DISEASE PROGRESSION

Paulo Guimaraes * , Medical University of South Carolina;
Jordan Elm, Medical University of South Carolina;
Huang Peng, Medical University of South Carolina;
Yuko Palesch, Medical University of South Carolina;
Karl Kieburtz, University of Rochester;
Barbara Tilley, Medical University of South Carolina

The UPDRS measures parkinsonian symptoms and disabilities and is the most widely used scale to track an individual's progression of Parkinson's disease. However, this score also reflects the effects of symptomatic treatments, hampering its ability to serve as a tool for the evaluation of true disease progression. We propose a nonlinear growth model derived as the sum of two individual functions, one reflecting the long-run trend of the disease and the other the symptomatic (short-term) effect of a drug. This specification allows for a direct quantification of the drug symptomatic effect as well as the long-term rate of disease progression, which has heretofore been masked by short-term drug effects. In our application we use a nonlinear mixed effects model on longitudinal data from several clinical trials testing different compounds. The model provides a good fit and shows results that are consistent with clinical insights. Moreover, we find that despite differences in the extent of short-term symptomatic improvements, most of the data exhibit similar long-term disease growth rates.

e-mail: guimarap@musc.edu

RISK-ANALYSIS BASED MIXTURE MODEL SELECTION

Surajit Ray*, University of North Carolina
Bruce G. Lindsay, Pennsylvania State University

Multivariate mixture models provide a convenient method of density estimation and model based clustering as well as providing possible explanations for the actual data generation process. But the problem of choosing the number of components (k) in a statistically meaningful way is still a subject of considerable research. Available methods for estimating k include, AIC and BIC, estimating the number through nonparametric maximum likelihood, hypothesis testing and Bayesian approaches with entropy distances. We propose the idea of risk based model selection as an approach to estimating the number of components in a mixture model. The loss function is defined as the distance between the true density and the proposed mixture model, and we choose the k which give the minimum risk. To avoid multidimensional numerical integration in the risk calculation, the distance is chosen appropriately (depending on the model class). We will also show that by varying the tuning parameter in the distance we can analyze a data set at different levels of smoothness. Application of the risk based methods along with comparison to other methods are considered in the results section.

e-mail: surajit@stat.psu.edu

ASYMPTOTIC PROPERTIES OF MINIMUM DISTANCE ESTIMATOR IN BETA MIXTURE DISTRIBUTIONS

Xuejun Peng * , Cleveland Clinic Foundation;
Arnold Stromberg, University of Kentucky;
Constance Wood, University of Kentucky

We study the asymptotic consistency and normality of minimum Cramer von-Mises distance estimator (MDE) and Maximum likelihood estimator (MLE) in the context of mixture distributions. The robustness and efficiency is then compared between MDE and MLE. The special case with Beta mixture distribution is studied numerically. An example is given using data from microarray experiments. We conclude that MDE performs better than MLE under certain conditions.

e-mail: xpeng@bio.ri.ccf.org

OPTIMAL DATA-BASED HEART RATE CORRECTION FOR QT_i

Rene Kubiak*, Boehringer Ingelheim Pharmaceuticals, Inc.
Steve Weldon, Boehringer Ingelheim Pharmaceuticals, Inc.

The QT interval (QT_i) of an ECG is the interval from the beginning of the QRS complex to the end of the T wave. Drug-induced prolongation of QT_i can cause arrhythmias, which may lead to ventricular fibrillation and sudden death. Thus, QT_i response is studied as a safety parameter in both the preclinical and clinical phases of drug development. QT_i response is correlated with heart rate (HR). To determine drug effect on QT prolongation, it is necessary to evaluate QT_i response adjusted for HR, or the corresponding RR interval (RR_i). QT correction formulae available in the literature (Bazett, Fridericia etc.) usually do not eliminate the correlation between QT_i and RR_i. These methods developed for clinical analyses do not perform well in nonclinical research. The relationship between the QT_i and RR_i response is best-fitted by curves having an approximate hyperparabolic shape. Current approaches to the statistical analysis of QT/QT_c response are incorporating individual or population data-based correction on non-drug data from the particular study. With this optimal data-based correction, the correlation between QT_i and RR_i can be eliminated and the resulting corrected QT interval values can be evaluated independently from RR_i variations.

e-mail: jschwenk@rdg.boehringer-ingelheim.com

HISTORICAL-BASED FORMULA FOR HEART-RATE CORRECTED QT INTERVALS IN BEAGLE DOGS

Kevin Guo * , Eli Lilly and Company;
Wendell C. Smith, Eli Lilly and Company

The purpose of this study was to determine a database- driven QT correction formula for toxicology studies conducted in beagle dogs at Eli Lilly and Company, Greenfield Laboratories. The database for this evaluation consisted of all baseline (pretreatment) values from 14 studies in beagle dogs over a course of one year and half (November, 2000 to April, 2002). The final database included values from 416 baseline ECG recordings, one from each of the 416 beagle dogs (208 females, 208 males). The power function was used to derive a population-specific correction formula. The nonlinear least squares regression estimate of the correction factor for females and males combined was 0.20 with a 95% confidence interval from 0.18 to 0.23. Correction factors computed for females and males separately were not significantly different. The correlation coefficient of 0.0040 between RR and QT interval values after correction with the Lilly database- driven correction factor (0.20) was not significantly different from zero. In contrast, significant correlations between RR and QT interval values remained after corrections with both Fridericia and Bazett formulas, -0.421 and -0.72 , respectively.

e-mail: guo@lilly.com

45. What More Needs to be Done in the Graduate Training of Statisticians?

FROM GRADUATE SCHOOL TO INDUSTRY: SOFTENING THE BLOW

Janet Wittes * , Statistics Collaborative, Inc.

Students entering industry directly from graduate school often suffer from culture shock. The endpoints change from grades to dollars, values from individual achievement to group success, and the greatest respect from the person who sees problems most abstractly to the one who is best able to explain statistical principles to the statistically naïve. Many graduate programs include some experience with real applications through a consulting course or laboratory; however, the problems are often relatively simple and the datasets small. Rarely is the student confronted with the simultaneous requirements of science, regulation, and economics. Thrust into such an unfamiliar complex world, the freshly minted graduate student longs to be able to practice his newly acquired sophisticated statistical techniques; instead, the technical requirements of the job may demand only careful attention to denominators - or glorified sixth grade math. This talk addresses some approaches graduate programs might adopt to prepare their students for a quicker adjustment to the workplace without sacrificing the time the student needs to master necessary statistical theory and methods.

janet@statcollab.com

PREPARING STATISTICIANS FOR COLLABORATIVE AND INDEPENDENT RESEARCH AT THE NATIONAL INSTITUTES OF HEALTH

Mitchell H. Gail*, National Cancer Institute

My former boss, David Byar, used to say he liked to hire people who could teach him something and were smarter than he. Dave also said he read 100 medical papers for every statistical one when preparing for an important consultation. These attitudes have relevance to the training and selection of statisticians to work at the National Institutes of Health and other scientific governmental agencies. Students with sound and modern theoretical training and computational skills have much to teach their employers. Yet to be well adapted for work in a scientific research agency, one should also have a sense of the purpose and methods of scientific discovery, enjoy collaborative scientific work, and continuously learn about the science and required statistical methods. How to inculcate these skills and attitudes is a challenge to graduate education.

e-mail: gailm@mail.nih.gov

THE EXEMPLARY BIOSTATISTICIAN IN ACADEMIC MEDICAL CENTERS

Ralph G. O'Brien*, Cleveland Clinic Foundation

Today's ideal biostatistical scientist develops and maintains a broad range of technical skills and personal qualities in four domains: (1) numeracy, in mathematics and numerical computing; (2) articulacy and people skills; (3) literacy, in technical writing and in programming; and (4) graphicacy. Yet, many of these components are given short shrift in university statistics programs. How can the exemplary collaborating medical statistician become 'completely sufficient' across this range of skills? Is this even possible?

robrien@bio.ri.ccf.org

COMPARING MOUSE TISSUE-SPECIFIC GENE EXPRESSION DATASETS

Timothy R. Hughes * , University of Toronto;
 Wen Zhang, University of Toronto;
 Quaid Morris, University of Toronto

The spatial pattern of expression of a gene is an indicator of function and also provides a starting point for discovering regulatory mechanisms. Several groups, including ours, have generated large-scale expression profiles across major mouse organs and tissues, with each group using a different microarray technology. The question arises whether the different data sets agree and, if not, which measurements are most accurate. A number of experimental variables can contribute to lack of agreement, as can the continued absence of rigorous standards in mammalian gene nomenclature and sequences. Post-processing steps (e.g. ‘normalization’) are also important in this analysis, because the raw data from different array types can carry different meanings. In an initial comparison of ~1,100 unambiguously-identified genes measured in fifteen different tissues in each of three different studies, Spearman rank correlations revealed both reassuring similarities and major discrepancies among the different studies. We are now compiling literature-based ‘benchmark’ genes in order to resolve these differences and identify systematic sources of error. This effort is complicated by apparent lack of standardization of the meaning of ‘tissue- specificity’ in the literature. Nonetheless, we ultimately plan to assemble a composite data set with confidence weights on each gene in each tissue.

e-mail: t.hughes@utoronto.ca

LARGE-SCALE ANALYSIS OF GENE COEXPRESSION

Paul Pavlidis*, Columbia University

A common strategy in functional genomics is to infer gene function based on coexpression of mRNA using microarrays: genes that have similar expression profiles are predicted to have a functional relationship. I will discuss strategies to improve the quality of such inferences by analyzing gene coexpression in multiple data sets collected by numerous laboratories. The fundamental concept being considered is that coexpression observations made by multiple laboratories are more reliable than those made using a single data set. This task is made more complex by the variety of microarray designs and methods in use. Particular reference will be made to two important issues: validation of the analytical strategies, and the exploration of possible sources of artifactual coexpression. Our approach will be discussed in the context of a large database of mammalian gene expression data sets we have developed. Our findings demonstrate that combining coexpression analyses across data sets adds considerable value to existing data, and contribute insights into the reproducibility and reliability of microarray data.

e-mail: pp175@columbia.edu

MULTI-STUDY GENOMIC DATA ANALYSIS

Giovanni Parmigiani * , Johns Hopkins University;
Elizabeth S. Garrett, Johns Hopkins University;
Edward Gabrielson, Johns Hopkins University

Genomic data analysis investigates the transcriptional activity of a large number of genes measured simultaneously. Because of cost and limitations in the accessibility of biological samples, the majority of genomic investigations use a limited number of biological samples and focus on specific sample types. While this provides highly valuable insight on gene regulation, important biological and medical questions require comparison and integration of genomic information across studies, across measurement technologies, and across biological systems. Such integration is what we call “multi-study genomic data analysis”. Our ability to efficiently integrate and accumulate information from related genomic experiments will be critical in the success of the massive investment being made on genomic studies. Yet, use of multi-study analysis to date is limited compared to its potential. This presentation will review statistical and bioinformatics issues, present exploratory and model-based tools, and illustrate them in two projects on the molecular classification of cancer.

e-mail: gp@jhu.edu

STATISTICAL METHODS FOR COMBINING DATA FROM MULTIPLE CANCER STUDIES

Debashis Ghosh*, University of Michigan

The recent development of high-throughput technologies for measuring gene expression has allowed for scientists to study biochemical activity on a global basis. This has been extremely useful in cancer studies, where microarray experiments have led to the development of molecular subtypes of and candidate biomarkers of various. In this talk, we explore the investigation of higher-order hypotheses involving gene regulation in cancer by combining multiple cancer microarray datasets as well as external biological data. In particular, we explore the use of combining multiple microarray datasets from cancers of the same origin, cancers of different origin, in vitro and in vivo microarray data and chromosomal localization data. Statistical issues and methods for these problems are discussed. Various examples are used to illustrate the methodologies.

e-mail: ghoshd@umich.edu

47. Recent Developments in Survival Analysis

AN OVERVIEW OF JOINT MODELING OF SURVIVAL AND LONGITUDINAL COVARIATES MEASURED WITH ERROR

Anastasios A. Tsiatis*, North Carolina State University,
Marie Davidian, North Carolina State University

A common objective in longitudinal studies is to characterize the relationship between a failure time process and time-independent and time-dependent covariates. Often, the proportional hazards model is used for this purpose, which necessitates knowledge of the time-dependent covariates for each individual at each failure time. In most studies, the longitudinally-measured covariate is only collected at a finite set of time points and may be subject to measurement error and biological variation. ‘Naive’ methods, such as last value carried forward, lead to biased estimators for the underlying relationship of survival to the time-dependent covariates. Recently, joint modeling of both the survival distribution and the longitudinally measured time-dependent covariates as a function of a common set of random effects has been advocated for this problem. We consider such models, give the rationale for their use, and trace some of the analytic methods proposed for estimation.

tsiatis@stat.ncsu.edu

48. New Advances in Nonparametric Statistics

COMPARING NONPARAMETRIC PROCEDURES

Yang Yuhong*, Iowa State University

Nonparametric procedures relax stringent parametric assumptions. However different nonparametric procedures work well for different situations. When multiple nonparametric procedures are available, how should one select the best procedure for the data at hand? Is it better to combine the candidate procedures than selecting one of them? In this talk, we will address these questions in the context of nonparametric function estimation from both selection and risk perspectives.

e-mail: yyang@iastate.edu

SPLINE ESTIMATION IN PARTIALLY LINEAR MODELS FOR LONGITUDINAL DATA

Jianhua Huang *, The University of Pennsylvania Wharton School

We consider marginal semiparametric partially linear models for longitudinal data, where the models are specified totally by some conditional moment restrictions. We derive the semiparametric efficient score function and information bound for this problem without the multivariate Gaussian assumption. We propose an estimation procedure based on spline approximation of the nonparametric part of the model and an extension of the parametric marginal generalized estimating equations (GEE). It is shown that our estimator of the parametric part of the model is asymptotically normal and achieves the semiparametric efficiency bound. Our estimator of the nonparametric part is consistent and the rate of convergence is given. The proposed estimator takes into account the with-in subject correlation naturally as in the standard GEE and is straightforward to implement. This is a joint work with Liangyue Zhang.

e-mail: jianhua@wharton.upenn.edu

IMAGE SEGMENTATION FOR CDNA MICROARRAY DATA BASED ON LOCAL LINEAR KERNEL SMOOTHING

Peihua Qiu * , University of Minnesota;
Jingran Sun, University of Minnesota

Gene microarray data are widely used in pharmaceutical and clinical research. By comparing gene expression in normal and abnormal cells, microarrays can be used for identifying genes involved in particular diseases, and then these genes can be targeted by therapeutic drugs. Most gene expression data are produced from cDNA microarray images. A microarray image consists of thousands of spots, with individual DNA sequences printed at each spot first and then equal amount of cDNA samples from treatment and control cells mixed and hybridized with the printed DNA sequences. To obtain gene expression data, the image needs to be segmented first to separate foregrounds from backgrounds for individual spots, and then averages of the foreground pixels are used for computing the gene expression data. Several image segmentation procedures have been suggested and included in some software packages handling gene microarray data. In this paper, a new image segmentation methodology is proposed based on local linear kernel smoothing. Theoretical arguments and numerical studies show that it has some good statistical properties and would perform well in applications.

e-mail: qiu@stat.umn.edu

STATE OF THE ART IN NONPARAMETRIC LACK-OF-FIT TESTS

Jeffrey D. Hart*, Texas A&M University

The last fifteen years have seen a great deal of research on nonparametric methods for testing the fit of models for regression curves and probability densities. Most of this research has focused on smoothing-based methods, i.e., methods utilizing nonparametric function estimators, such as kernel estimators, local polynomials and Fourier series. This talk reviews many of the methods that have been proposed and discusses some of their empirical and theoretical properties. In the simple setting of testing for “no-effect” in regression, an attempt is made to categorize tests according to their power properties.

e-mail: hart@stat.tamu.edu

KERNEL ESTIMATION FOR STATIONARY PROCESSES

Wei Biao Wu*, The University of Chicago

We will discuss large sample properties of kernel estimators for causal processes. Under minimal conditions on bandwidths, we establish asymptotic normality and optimal uniform error bounds for a wide class of non-linear time series and short-memory linear processes. When the process is long-memory, we also depict the dichotomous and trichotomous phenomena for various choices of bandwidths, where the limiting distributions may not be Gaussian.

e-mail: wbwu@galton.uchicago.edu

COMBINING INFORMATION FROM MULTIPLE SURVEYS FOR SMALL-AREA ESTIMATION: A BAYESIAN APPROACH

Trivellore E. Raghunathan, University of Michigan
Dawei Xie, University of Michigan
Nathaniel Schenker*, National Center for Health Statistics
Van L. Parsons, National Center for Health Statistics
William W. Davis, National Cancer Institute
Eric J. Feuer, National Cancer Institute
Kevin W. Dodd, National Cancer Institute

Cancer surveillance research requires accurate and precise estimates of the prevalence of cancer risk factors and screening for small areas such as counties. Two popular data sources are: (1) the Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey, which includes almost every county in the United States, but which has lower response rates as is typical with telephone surveys and does not include subjects who live in households with no telephones; and (2) the National Health Interview Survey (NHIS), a face-to-face survey, which is smaller than the BRFSS and does not include as many counties, but which has higher response rates and includes both telephone and non-telephone households. A hierarchical Bayesian approach is used to combine information from the two surveys. The proposed model incorporates potential noncoverage and nonresponse biases in the BRFSS as well as complex sample design features of both surveys. A Markov Chain Monte Carlo method is used to fit the model using county-level design-based direct estimates from both surveys and covariates. Combined county-level estimates of the prevalences of smoking and common cancer screening procedures are obtained for the years 1997 – 2000. The combined estimates are different from those based on the BRFSS alone.

e-mail: nschenker@cdc.gov

MODELING THE COUNTY-LEVEL RESPONSE RATE FOR THE BEHAVIORAL RISK FACTORS SURVEILLANCE SYSTEM (BRFSS)

William W. Davis * , National Cancer Institute;
Ali H. Mokdad, Centers for Disease Control and Prevention;
Machell Town, Centers for Disease Control and Prevention;
Eric J. Feuer, National Cancer Institute

The Behavioral Risk Factors Surveillance System (BRFSS) is a large ongoing telephone survey of the health behaviors of US adults. For many states, it is the only source of data to monitor risk behaviors associated with the leading causes of death. This paper examines the determinants of the response rate at the county level. This model and analysis may be useful in improving telephone response rates and also in understanding the potential biases due to non-response in telephone surveys.

e-mail: davisbi@mail.nih.gov

A BAYESIAN APPROACH FOR COMBINING INFORMATION FROM MULTIPLE SURVEYS IN SMALL AREA ESTIMATION USING PUBLIC-USE DATA

Dawei Xie *, University of Michigan; Trivellore E. Raghunathan, University of Michigan

Cancer surveillance research requires accurate county level prevalence rates of screening and risk factors. Raghunathan et al (2003) developed a hierarchical Bayesian approach for obtaining such estimates by combining information from in-house data from the Behavioral Risk Factor Surveillance System (BRFSS) and the National Health Interview Survey (NHIS). Due to confidentiality concerns, however, it is not clear whether these model-based estimates obtained using in-house data can be released to researchers for use in their research. In public-use NHIS data, only large geographic areas are identified. In this article, we develop estimates at the MSA level using publicly available data and then use the estimated model, the county level covariates and the direct estimates from BRFSS to obtain county level estimates. A Markov Chain Monte Carlo (MCMC) method is used to generate the joint posterior predictive distribution of the county level unknown quantities. Yearly county level estimates for 49 states, District of Columbia and the whole state of Alaska in 1997-2000 are developed and compared to the model estimates from the in-house NHIS and BRFSS data. These estimates will also be compared with model-assisted estimates obtained by Elliott and Davis (2003), again using public-use data sets.

e-mail: xiedawei@umich.edu

OBTAINING CANCER RISK FACTOR PREVALENCE ESTIMATES IN SMALL AREAS: COMBINING DATA FROM THE BRFSS AND THE NHIS

Michael R. Elliott*, University of Pennsylvania School of Medicine

Cancer surveillance research requires accurate estimates of risk factors at the small area level. These risk factors are often obtained from surveys such as the National Health Interview Survey (NHIS) or the Behavioral Risk Factors Surveillance System (BRFSS). The NHIS is a nationally representative, face-to-face survey with a high response rate; however, it cannot produce state or sub-state estimates of risk factor prevalence because sample sizes are too small and small area identifiers are unavailable to the public. The BRFSS is a state-level telephone survey that excludes non-telephone households and has a lower response rate, but does provide reasonable sample sizes in all states and many counties and has publicly-available small area identifiers (counties). A novel extension of dual-frame estimation is proposed using propensity scores that allows the complementary strengths of each survey to compensate for the weakness of the other. This method is used to obtain county-level estimates of 1999-2000 male smoking rates and mammogram usage rates among females 40 and older. The method is evaluated using data from the Current Population Survey Tobacco Use Supplement,

e-mail: melliott@cceb.upenn.edu

ESTIMATING INCIDENCE RATES USING PEDIGREE SAMPLING

Steven N. MacEachern, Ohio State University;
Amy K. Ferketich * , Ohio State University;
Stanley Lemeshow, Ohio State University;
Lei Shen, Ohio State University;
Judith A. Westman, Ohio State University;
Albert de la Chapelle, Ohio State University;
Clara D. Bloomfield, Ohio State University

Little is known about cancer incidence in the Amish. The Amish Population Cancer Study was conducted to estimate the incidence and prevalence of cancers in the Amish, identify those at risk for developing cancer and provide education on cancer prevention and screening. In this study, 92 households were sampled from a total of 4,512 in Holmes County, Ohio during a six-month period in 2000. For each household sampled, a complete cancer family history was taken from the proband and his spouse. In some households, this “list” of relatives contained several hundred individuals and most people appeared on several household lists. This sampling procedure led to data on 28,201 adults who were alive during the reference period, 1996- 1999. During this time, 103 adult cancer cases occurred. Here, we will describe different methods of cancer incidence rate estimation, including ratio, probability proportional to size and distinct case estimation. We will also describe two empirical Bayes estimators. Then, we will present the results of the simulation studies conducted to study the properties of the various methods considered and, finally, we will provide suggestions on how to improve the sampling design to obtain data on cancer incidence among small, closed populations.

e-mail: ferketich.1@osu.edu

HIV/AIDS PROJECTION- A BACK CALCULATION APPROACH

Anbupalam Thalamuthu*, University of Pittsburg
Ravanan Ramanujam, Presidency College, Chennai, India

Backcalculation is a methodology to reconstruct the past human immunodeficiency virus (HIV) infection rates from the AIDS incidence data and incubation distribution. An important public health problem is how many individuals are infected with HIV. Most national surveillance data systems record only AIDS cases and not HIV infected cases. In this work we attempt to study the size of the HIV infection estimates under various parametric model assumption using backcalculation methodology and apply these models to estimate the infection levels using surveillance data at national level for India and regional level for Tamilnadu using the annual surveillance databases. Various alternative choices for the two major components of backcalculation methodology namely incubation period distributions and infection densities are analysed in detail. The uncertainties of the backcalculation estimates are quantified by real and simulated data sets.

e-mail: perumal_venkatesan@yahoo.com

A DYNAMIC GRAPHICAL MODEL FOR INFLUENZA SURVEILLANCE

Paola Sebastiani ^{*}, Boston University School of Public Health; Ken D. Mandy, Harvard Medical School; Peter Szolovits, Massachusetts Institute of Technology; Isaac S. Kohane, Harvard Medical School; Marco F. Ramoni, Harvard Medical School

The threat of bioterrorism, the recent Severe Acute Respiratory Syndrome (SARS) epidemic and the growing fear of an influenza pandemic similar to the 1918 Spanish flu highlight the need for surveillance systems able to provide early detection of epidemic events. This talk shows that the number of patients with respiratory syndromes in a pediatric emergency department predict influenza morbidity and mortality in the general population up to three weeks in advance. To support this claim, we have used probabilistic graphical models to discover the interplay among four data sources: the frequency of patients with respiratory syndromes at the emergency departments of a pediatric and a general hospital, the number of cases of influenza-like-illness in Massachusetts, and the number of deaths for pneumonia and influenza in New England. Our results show that pediatric patients with respiratory syndromes at the emergency department provide an early signal of impending influenza morbidity and mortality compared to the influenza-like-illness data currently collected through federal surveillance programs.

e-mail: sebas@bu.edu

CONSTRAINED BAYES ESTIMATES OF RANDOM EFFECTS WHEN DATA ARE SUBJECT TO A DETECTION LIMIT

Renee H. Moore^{*}, Emory University
Robert H. Lyles, Emory University
Amita K. Manatunga, Emory University

In many environmental and HIV epidemiologic studies, the prediction of random effects corresponding to subject-specific characteristics (e.g., mean exposure or HIV viral load levels) is of much interest. A common feature of data from such studies is that some values fall below a limit of detection associated with the measuring instrument. A widely accepted predictor of subject-specific random effects is the Bayes estimate, which has the well-known potential disadvantage of overshrinking toward the population mean. Alternative predictors, termed constrained Bayes estimates, have been shown to reduce this shrinkage in an appealing way. Although the Bayes estimate has been adapted to predict random effects when data contain non-detectable values, the constrained Bayes methodology has not been demonstrated in this setting. In this talk, we combine general methods found in the literature for computing constrained Bayes estimates with methods for computing Bayes estimates in the presence of non-detects. The resulting constrained Bayes predictor is compared with the Bayes predictor and ad hoc predictors of both types that fail to properly account for non-detects. We illustrate using real exposure data from a cohort of chemical factory workers. Key words: Censoring; Prediction; Shrinkage

e-mail: rhmoore@emory.edu

VALIDITY OF SHORT-TERM BETA-AGONISTS PRESCRIPTIONS AS A MARKER FOR ASTHMA

Xiaoming Bao * , University of Chicago;
Paul J. Rathouz, University of Chicago;
Vanja M. Dukic, University of Chicago

Asthma, a chronic disease that affects the smooth muscle in the airways, is a major public health problem, and an increasing concern in the United States. Prescription refills for medication, such as short-term Beta-agonists may be a sensitive indicator of asthma that can be easily obtained at the population level from insurance claims data. The goal of this study is to validate the use of short-term Beta-agonists as a marker for asthma. The target population is the Medicaid population in the City of Chicago. A Mantel-Haenszel method will be used to characterize the association between short-term Beta-agonist prescriptions and traditionally studied asthma outcomes as a function of time lag after adjusting for subject slow-varying time confounders. Bootstrap, a re-sampling method, will be performed to quantify uncertainty. As an alternative approach, the receiver operating characteristic (ROC) analysis will be used to assess the accuracy of use of short-term Beta-agonist prescriptions as an indicator for acute asthma events. Our research results demonstrate the validity of short-term Beta-agonist use as a marker for asthma. This research is supported by EPA grant R-82940201-0. The abstract does not necessarily reflect EPA views.

e-mail: xbao1@uchicago.edu

MULTIPLE CUTPOINT SELECTION FOR CATEGORIZING A CONTINUOUS PREDICTOR

Sean M. O'Brien*, National Institute of Environmental Health Sciences

We present a new approach for choosing both the number of categories and the location of category cutpoints when a continuous exposure variable needs to be categorized to obtain numerical summaries of the exposure effect. The optimum categorization is defined as the partition that minimizes a measure of distance between the true expected value of the outcome for each subject and the estimated average outcome among subjects in the same exposure category. To estimate the optimum partition, an efficient nonparametric estimate of the unknown regression function is substituted into a formula for the asymptotically optimum categorization. This new approach is easy to implement and outperforms traditional cutpoint selection methods in simulation studies.

e-mail: obrien4@niehs.nih.gov

INFECTIOUS DISEASE SURVEILLANCE SCHEME C: A MULTIVARIATE APPROACH

Bo Hong * , University of Alabama;
Michael J. Hardin, University of Alabama

A majority of the statistical research efforts on monitoring infectious disease surveillance have focused on univariate techniques, which are typically implemented on aggregated national disease data. However, most surveillance systems produce multivariate data: several reporting units are monitored simultaneously. The objective of this research is to design a multivariate statistical surveillance scheme that simultaneously monitors cases of infectious diseases on multiple geographic locations, and provides early signals for the potential onset of epidemics in time and space. Multivariate scheme may detect aberrations earlier than univariate ones because the data for infectious diseases from different locations are often cross-correlated. Further, surveillance data also exhibit positive autocorrelation. The literature has examined multivariate control charts for independent processes and univariate control charts for autocorrelated processes separately. In our work, we examine multivariate control charts for autocorrelated processes by modeling the processes with multiple time-series method and then implementing multivariate control charts on the residuals. Based on preliminary studies, our proposed scheme outperforms the univariate schemes when applied to simulated processes, and the CDC weekly reported cases of Chlamydia from 1998 to 2003.

e-mail: bhong1@bcc.cba.ua.edu

51. Bayesian Survival Analysis

JOINT LONGITUDINAL-SURVIVAL-CURE MODELS AND THEIR APPLIATION TO PROSTATE CANCER

Menggang Yu* , University of Michigan
Ngayee J. Law, Roche Molecular Systems, Inc.
Jeremy M.G. Taylor, University of Michigan
Sandler M. Howard, University of Michigan

In cancer research it is common for there to be long term survivors or cured patients and methods have been developed to analyze such data. In this article, we review both joint models for the analysis of longitudinal and survival data and cure models. We then present a joint longitudinal-survival-cure model to analyze data from a study of prostate cancer patients treated with radiation therapy. In this model each patient is assumed to be either cured or susceptible to clinical recurrence. The cured fraction is modeled as a logistic function of baseline covariates. The longitudinal PSA data is modeled as a non-linear hierarchical mixed model, with different models for the cured and susceptible groups. The clinical recurrences are modeled as a time-dependent proportional hazards model for those in the susceptible group. The baseline variables are covariates in both the failure time and longitudinal models. We use both a Monte Carlo EM algorithm and Markov chain Monte Carlo techniques to fit the model. The results from the two estimation methods are compared. We focus on both selected parameters of the model and derived interpretable quantities. We use both a Monte Carlo EM algorithm and Markov chain Monte Carlo techniques to fit the model. The results from the two estimation methods are compared. We focus on both selected parameters of the model and derived interpretable quantities.

e-mail: menggang@umich.edu

BAYESIAN ANALYSIS OF CONSTRAINED HAZARD FUNCTIONS

Gregg E. Dinse^{*}, National Institute of Environmental Health Sciences;
David B. Dunson, National Institute of Environmental Health Sciences

In survival analyses, it is often reasonable to place general restrictions on the hazard rates, such as assuming they are monotone, umbrella-shaped, or bathtub-shaped functions of time. We describe a new approach for Bayesian analysis of hazard functions under these types of shape constraints. Prior distributions are proposed for transformed increments on the cumulative hazard. These priors consist of mixtures of point masses and truncated normal densities. This structure facilitates efficient posterior computation via a data augmentation Gibbs sampler, which updates the high-dimensional vector of constrained parameters jointly. The algorithm is easy to implement for a variety of censoring structures and shape constraints. Generalizations for incorporating covariates are straightforward. We illustrate the approach through application to tumor data from a bioassay study.

e-mail: dinse@niehs.nih.gov

A CLASS OF BAYESIAN BOX-COX TRANSFORMATION HAZARD REGRESSION MODELS

Guosheng Yin^{*}, M. D. Anderson Cancer Center
Joseph G. Ibrahim, University of North Carolina at Chapel Hill

We propose a novel class of Box-Cox transformation models on the hazard functions for survival data. This class of models allows a very broad range of relationships between the baseline hazard and the hazard function. It includes the Cox proportional hazards model and the additive hazards model as two special cases. Several properties of the model are derived, and illustrations of the behavior of the Box-Cox transformation parameter are provided. A class of joint prior distributions is proposed for the model parameters. Due to the requirement of a nonnegative hazard function in the survival model, complex multidimensional nonlinear parameter constraints must be imposed in the model formulation. We propose an efficient Markov chain Monte Carlo (MCMC) computational scheme for sampling from the posterior distribution of the parameters. The joint priors are constructed through a conditional-marginal specification, in which the conditional distribution is univariate, and one which absorbs all of the non-linear parameter constraints. The marginal part of the prior specification is free of any constraints. This class of prior distributions allows us to easily compute the full conditionals needed for Gibbs sampling, incorporating the constraints, and hence implement the MCMC algorithm in a relatively straightforward fashion. This class of models is illustrated with a melanoma dataset.

e-mail: gsyin@mdanderson.org

BIVARIATE POSITIVE STABLE FRAILTY MODELS

Madhuja Mallick *, University of Connecticut;
Nalini Ravishanker, University of Connecticut

This presentation discusses a bivariate frailty model with piecewise exponential baseline hazard and a bivariate positive stable frailty distribution within a conditional proportional hazards model. The likelihood formulation and Bayesian inference procedure are presented. We illustrate our approach on a data set involving multivariate survival times.

e-mail: madhujamallick@hotmail.com

PROPORTIONAL MEAN REGRESSION MODELS FOR CENSORED DATA

Sujit K. Ghosh*, North Carolina State University
Subhashis Ghosal, North Carolina State University

A novel semiparametric regression model for censored data is proposed as an alternative to the widely used proportional hazards survival model. The proposed regression model for censored data turns out to be flexible and practically meaningful. Features include physical interpretation of the regression coefficients through the mean response time instead of the hazard functions. It is shown that the regression model obtained by a mixture of parametric families, has a proportional mean structure (as in an accelerated failure time model). The statistical inference is based on a nonparametric Bayesian approach that uses a Dirichlet process prior for the mixing distribution. Consistency of the posterior distribution of the regression parameters in the Euclidean metric is established under certain conditions. Finite sample parameter estimates along with associated measure of uncertainties can be computed by a MCMC method. Simulation studies are presented to provide empirical validation of the new method. Some real data examples are provided to show the easy applicability of the proposed method.

e-mail: sghosh@stat.ncsu.edu

SELECTING FACTORS PREDICTIVE OF HETEROGENEITY IN MULTIVARIATE EVENT TIME DATA

David B. Dunson, National Institute of Environmental Health Sciences;
Zhen Chen *, University of Pennsylvania

In multivariate survival analysis, investigators are often interested in testing for heterogeneity among clusters, both overall and within specific classes. We represent different hypotheses about the heterogeneity structure using a sequence of gamma frailty models, ranging from a null model with no random effects to a full model having random effects for each class. Following a Bayesian approach, we define prior distributions for the frailty variances consisting of mixtures of point masses at zero and inverse-gamma densities. Since frailties with zero variance effectively drop out of the model, this prior allocates probability to each model in the sequence, including the overall null hypothesis of homogeneity. Using a counting process formulation, the conditional posterior distributions of the frailties and proportional hazards regression coefficients have simple forms. Posterior computation proceeds via a data augmentation Gibbs sampling algorithm, a single run of which can be used to obtain model-averaged estimates of the population parameters and posterior model probabilities for testing hypotheses about the heterogeneity structure. The methods are illustrated using data from a lung cancer trial conducted by the Eastern Cooperative Oncology Group.

e-mail: zchen@cceb.upenn.edu

52. Diagnostic and Screening Tests

EVALUATION OF METHODS FOR ASSESSING THE PERFORMANCE OF A NEW TEST WHEN THERE IS IMPERFECT GOLD STANDARD

H.Refik Burgut*, Cukurova University
Macit U. Sandikci, Cukurova University
Fatin Kocak, Cukurova University

Accuracy of a new diagnostic test must frequently be assessed when a perfect gold standard does not exist. Uses of imperfect gold standard result in biases on the accuracy of the new diagnostic test. Several methodologies for assessing the accuracy of a new diagnostic test when there is no gold standard to compare are proposed in the literature. This list includes using single imperfect standard, discrepant resolution, latent class analysis, and composite reference standard among many. For the diagnosis of Helicobacter Pylori (HP) infection, several diagnostic tests such as histological examination, culture of biopsy specimens, rapid urea's test, urea breath test and serology are available and none is assumed to be as perfect gold standard. During the last couple of years, direct enzyme immunoassay, which uses only stool samples, has become available to detect Helicobacter Pylori antigen. In this presentation, several approaches for assessing the accuracy of direct enzyme immunoassay on detecting HP will be evaluated and compared. Based on the finding, a recommendation will be made as to which approach should be taken in order to evaluate the accuracy of direct enzyme immunoassay.

e-mail: refik@cu.edu.tr

COMPARISON OF QT PROLONGATION USING RODENT MODELS

Kanaka D. Tatikola * , Johnson & Johnson Pharmaceutical Research and Development, L.L.C.;
Jaya Natarajan, Johnson & Johnson Pharmaceutical Research and Development, L.L.C.;
James J. Colaianne, Johnson & Johnson Pharmaceutical Research and Development, L.L.C.;
Quan Li, Johnson & Johnson Pharmaceutical Research and Development, L.L.C

Over the past few years, much discussion has centered on the evaluation of QT prolongation and proarrhythmic potential for non-antiarrhythmic drugs. The ability of non-clinical studies to predict possible QT prolongation in humans will have considerable value in saving resources of R&D. We report the results from a non-clinical study of 22 drugs (11 known to cause QT prolongation in humans and 11 that do not) in two rodent species. For each species and drug, 3 to 5 animals were dosed with increasing dosages every 20 minutes for a total of 3 to 5 doses. Two replicates of QT and HR measurements were taken every minute. A matching vehicle group was included in each experiment and followed the same dosing and ECG measurement patterns. Various end points and statistical approaches were used to investigate the correlation of the animal models findings with human results. Degree of concordance between animal and human findings depended upon choice of end points selected and the animal model used.

e-mail: ktatikol@yahoo.com

USE OF RECEIVER OPERATING CHARACTERISTIC CURVE IN MEDICAL DECISION MAKING

Jiping Wang*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh

Receiver operating characteristic (ROC) curve analysis has been widely used to evaluate the capability of algorithms based on statistical models to predict outcomes in the public health and medical areas. These ROC analyses procedures use the area under the ROC curve (AUC) as an index of predictive capability and apply various inferential procedures to test the usefulness of the model. We have investigated this application of ROC methodology to logistic regression models where the parameters were estimated either by maximum likelihood estimation or the linear discriminate function. We demonstrate that the conventional method of assessing the adequacy of a predictive model results in an estimate of the AUC that is biased upward and the often used test based on the normal distribution to test predictive capability has an elevated type I error. Thus the conventional method results in an overly optimistic assessment of the model. We propose a less biased estimator of AUC, a valid test for the hypothesis $AUC=A_0$ and a method of constructing confidence limits for AUC when the independent variables follow a multivariate normal distribution. Furthermore, the procedure appears robust for deviations from the multivariate normal distributions as long as at least one of the independent variables is normally distributed.

e-mail: wang@nsabp.pitt.edu

COMBINING CORRELATED DIAGNOSTIC TESTS APPLICATION TO NEUROPATHOLOGIC DIAGNOSIS OF ALZHEIMER'S DISEASE

Chengjie Xiong * , Washington University in St. Louis;
Daniel W. McKeel, Jr., Washington University in St. Louis;
J. Phillip Miller, Washington University in St. Louis;
John C. Morris, Washington University in St. Louis

This article studies the problem of combining correlated diagnostic tests to maximize the discriminating power between the diseased population and the healthy population. We consider all possible linear combinations of multiple diagnostic tests and search for the one that achieves the largest area under the receiver operating characteristic (ROC) curve. We discuss the statistical estimation of the optimum linear combination test and the associated maximum area under the ROC curve. Our approach is based on the assumption of multivariate normal distribution of the multiple diagnostic tests. We also present the application of the proposed techniques to the neuropathologic diagnosis of Alzheimer's disease based on brain lesions from five different brain locations using a data set from the Washington University Alzheimer's Disease Research Center.

e-mail: chengjie@wubios.wustl.edu

MULTI-READER, MULTI-CASE ROC ANALYSIS: AN EMPIRICAL COMPARISON OF FIVE METHODS

Nancy A. Obuchowski*, Cleveland Clinic Foundation
Sergey V. Beiden, U.S. Food and Drug Administration
Kevin S. Berbaum, University of Iowa
Stephen Hillis, Iowa City VA Medical Center
Hemant Ishwaran, Cleveland Clinic Foundation
Hae Hiang Song, Catholic University Medical College, Seoul, Korea
Robert F. Wagner, U.S. Food and Drug Administration

Several statistical methods have been developed for analyzing multi-reader, multi-case (MRMC) ROC studies, yet little work has been done in comparing them. We re-analyzed the data from three previously reported real-life studies using five MRMC methods. We compared the 95% CIs and p-values obtained with each method. Important differences in p-values and CIs were seen when using parametric vs. nonparametric estimates of accuracy, and there were the expected differences for random- vs. fixed-reader models. Controlling for these differences, the Dorfman-Berbaum-Metz (DBM) and Obuchowski-Rockette (OR) methods gave similar results. The Beiden-Wagner-Campbell (BWC) method and Ishwaran's hierarchical ROC (HROC) sometimes detected differences that did not reach significance with other methods. The multivariate WMW, described by Song, yielded results similar to the other fixed-reader nonparametric methods. The results of this work suggest that the common sample size of 5-6 readers for MRMC studies is too small. Furthermore, until more fundamental work is done, investigators should consider comparing several MRMC approaches in the analysis of their data.

e-mail: nobuchow@bio.ri.ccf.org

EQUIVALENT VARIATIONS OF THE DORFMAN-BERBAUM-METZ AND OBUCHOWSKI-ROCKETTE METHODS FOR ROC DATA

Stephen L. Hillis * , Iowa City VA Medical Center;
Nancy A. Obuchowski, Cleveland Clinic Foundation;
Kevin M. Schartz, University of Iowa;
Kevin S Berbaum, University of Iowa

There are several different statistical methods for analyzing multireader ROC data that treat both cases and readers as random effects, with the Dorfman-Berbaum-Metz (DBM) method used most frequently. Another method is the corrected F method proposed by Obuchowski and Rockette (OR). The DBM method consists of a conventional 3-way modality-by-reader-by-case ANOVA of pseudovalues, while the OR method consists of a 2-way modality-by-reader ANOVA of accuracy estimates with correlated errors. We discuss easily incorporated variations of each method with respect to accuracy measure, covariance estimation, and degrees of freedom. We then show that the two methods yield identical results when based on the same accuracy measure, covariance estimation, and degrees of freedom

e-mail: steve-hillis@uiowa.edu

SIMULTANEOUS INFERENCE FOR PREVALENCE USING POOLED ASSESSMENTS

Joshua M. Tebbs*, Kansas State University
Melinda H. McCann, Oklahoma State University

When estimating the prevalence of a rare trait such as human immunodeficiency virus (HIV), the use of pooled testing can confer substantial benefits when compared to individual testing. In addition to screening experiments for infectious diseases, pooled testing (also known as group testing) has also been used in other applications, including drug testing, multiple-vector transfer designs in plant pathology, and epidemiological applications involving animal disease. In this talk, we consider situations wherein different strata (or treatments) are to be compared, with the goals being (a) to assess practical differences between strata and (b) to rank strata in terms of prevalence. To achieve these goals, we derive two simultaneous interval estimation procedures for use with pooled data. Since our procedures rely on asymptotic results, we investigate small-sample behavior of the procedures and compare them in terms of simultaneous coverage and mean interval length. We also discuss how to determine optimal pool sizes that deliver an acceptable simultaneous coverage probability, even when taking the cost of testing and interval precision into account. We illustrate our methods using data from an observational HIV study involving heterosexual males who use intravenous drugs.

e-mail: tebbs@ksu.edu

SOME STATISTICAL ASPECTS OF THE ANALYSIS OF GENOMIC SEQUENCES

Lily Wang * , University of North Carolina at Chapel Hill;
Kumar Sen, University of North Carolina at Chapel Hill

During the course of evolution, DNA sequences undergo continuous changes through mutations such as substitution, insertion or deletion. Along the lengthy stretches of sequences, some regions that do not change as much as the rest are called the "functional domains". Their resistance to change often suggests these regions serve critical functions; therefore, similarities of the sequences often suggest likeness in structure and function, in addition to relationships in phylogeny. Profile alignment uses multiple alignment profile of a family of related sequences to search a database for more examples of the family. Given a multiple alignment, as we are searching along another long sequence, we construct a distance measure between the profile and the sequence called the profile score. The scores are a set of dependent stationary triangular arrays. We derive the asymptotic distribution of the profile scores and show it can be well modeled using normal mixture distribution. The closed form expression for the largest characteristic observation from normal mixture model will then be shown. Next, we discuss the asymptotic properties of the extreme value random variable for the profile scores when mean and variance are unknown. Finally, numerical studies will be demonstrated.

e-mail: lilywang@isis.unc.edu

MODELING NUCLEOTIDE SEQUENCE VARIATION IN A VIRAL QUASISPECIES

Philip M. Dixon* , Iowa State University
Ozlem Ilk, Iowa State University

A viral quasispecies is a collection of related but distinct genetic sequences. Viral mutation rates are high, so a population founded by a single genetic sequence rapidly diversifies. Some of this sequence variation codes for different amino acid sequences; most is neutral. Recently, various groups have described temporal changes in a viral quasispecies. These changes are usually summarized using a phylogenetic tree that ignores the temporal structure of the data. A better model for the sequence change in longitudinal data can be constructed as a Bayesian hierarchical model. This combines a model for the mutation process, a model for the abundance of each sequence over time, and a model for the sampling process. Such a model can account for unobserved mutations. Inference is by MCMC. The approach is illustrated using data from a long term study of the PRRS virus, a recent and devastating viral disease in pigs.

e-mail: pdixon@iastate.edu

COMPARISON AND VALIDATION OF PROTEIN IDENTIFICATION SEARCHING TOOLS USING MASS SPECTROMETRY DATA

JungBok Lee *, Korea University;

Jae Won Lee, Korea University;

Mira Park, Eulji University

Proteome analysis typically involves protein separation by 2D gel electrophoresis followed by protein identification based on mass spectrometry (MS) peptide mapping and genome database searching which is widely used and fundamental method for proteome research. Protein identification based on MS of a protein digest assumes that a pattern of proteolytic peptide masses provides a fingerprint of a particular protein and that the fingerprint can be recognized when searching a genome database. Several algorithms have been developed such as MOWSE, MS-FIT, MASCOT, ProFound, etc. It is critical to have objective method to assess the performance of several algorithms which depends on the number of database entries, input parameters (including filtering of target DB, characteristics of sample protein, etc). In this talk, we compare several searching algorithms and validation techniques of searching results in various situations.

e-mail: jungboky@korea.ac.kr

A FUNCTIONAL DATA APPROACH TO MALDI-TOF MS PROTEIN ANALYSIS

Dean Billheimer*, Vanderbilt University

Matrix-assisted laser desorption-ionization, time-of-flight (MALDI-TOF) mass spectrometry (MS) is emerging as a leading technology in the proteomics revolution. MALDI-TOF MS allows direct measurement of the protein "signature" of tissue, blood, or other biological samples, and holds tremendous potential for disease screening, diagnosis and treatment. Despite recent technical advances in signal generation, key challenges remain in signal normalization and quantitation. Methods of functional data analysis (FDA) provide a natural framework for evaluating these complicated biological signals. I provide a summary of the MALDI-TOF MS technology, describe statistical characteristics of these data, and introduce an FDA approach for their analysis. The methods are illustrated in an analysis of protein mass spectra from brain tumors (gliomas) and normal white matter tissue. The goal is to characterize protein expression differences in tumor progression. FDA methods are shown to identify characteristic differences between glioma and normal brain tissue. Further, these differences would be difficult, if not impossible, to discern without use of the FDA (or similar) approach.

e-mail: dean.billheimer@vanderbilt.edu

MODEL-BASED ASSIGNMENT AND INFERENCE OF PROTEIN BACKBONE NUCLEAR MAGNETIC RESONANCE

Olga Vitek * , Purdue University;
Jan Vitek, Purdue University;
Bruce A. Craig, Purdue University;
Chris Bailey-Kellogg, Purdue University

Nuclear Magnetic Resonance (NMR) spectroscopy is a key experimental technique used to study structure, dynamics and molecular interactions of proteins. Analysis of NMR spectra, however, is hampered by a substantial amount of noise, and no generally accepted measure of uncertainty associated with the resulting interpretations exists. In this paper, we focus on backbone resonance assignment, an essential step in the NMR process which is typically viewed as a deterministic optimization problem. We argue that NMR spectra are subject to random variation, and ignoring this stochasticity leads to false optimism and erroneous conclusions. We propose a Bayesian statistical model that accounts for various sources of uncertainty and provides an automatable framework for inference. The proposed approach quantifies the uncertainty associated with the spectra in several ways: 1) it provides a measure of the overall plausibility of assignments, 2) it assesses the information content in the data by means of posterior distribution of plausible assignments, and 3) it provides a quantification of uncertainty in the individually assigned resonances in terms of their posterior standard deviations. The approach can be used to annotate assignments deposited in a database according to their information content.

e-mail: ovitek@stat.purdue.edu

SYSTEMATIC VARIATION IN DIFFERENCE GEL ELECTROPHORESIS DATA

Kimberly F. Sellers* , Carnegie Mellon University
Jeffrey Miecznikowski, Carnegie Mellon University
Surya Viswanathan, Carnegie Mellon University
William F. Eddy, Carnegie Mellon University
Jonathan S. Minden, Carnegie Mellon University

Two-dimensional Difference Gel Electrophoresis (DIGE) circumvents many of the problems associated with gel comparison via the traditional approach, two-dimensional gel electrophoresis (2DE). DIGE's accuracy, however, can be further improved by removing the many associated sources of systematic variation. This talk will identify these sources (including those caused by the apparatus detection system for locating proteins) and illustrate the results using data from a collection of experiments.

e-mail: ksellers@stat.cmu.edu

WAVELET BASED ANALYSIS OF HOMOLOGOUS DNA SEQUENCES FOR FOUR MAMMALIAN SPECIES

Jean Roayaei * , U.S. Food and Drug Administration, CDRH;
Juan-Miguel Marin, Universidad Rey Juan Carlos Madrid, Spain

The DNA sequences can be recoded, with respect to purine nucleotides (A,G) versus pyrimidine nucleotides (C,T), as a binary sequence W in the alphabet $\{-1,1\}$, where 1 corresponds to A,G and -1 corresponds to C,T. With such convention, each genomic sequence results in a word W generating a random walk. Previous research concluded that trajectories of particular DNA-based random walks exhibit self-similarity properties. The self-similarity indices were subsequently used to distinguish between the exons and introns. In this paper we explore the local and global scaling exponents or, equivalently Hurst exponents, for DNA random walks of several species. We take complete sequences and find the global scaling exponent by using orthogonal wavelet transformations.

e-mail: jbr@cdrh.fda.gov

5 4. Sequential Trials, Interim Decision Making, and Multiplicity

THE DESIGN AND IMPLEMENTATION OF GROUP SEQUENTIAL TRIALS FOR LONGITUDINAL ENDPOINTS

John M. Kittelson*, University of Colorado
Katrina Sharples, University of OtagoDunedin, New Zealand
Scott S. Emerson, University of Washington Seattle

Many clinical trials include repeated measurements of outcome over time. In such situations, the time-evolution of treatment effects may lead to bias and incorrect decisions at the interim analyses. Methods that allow correct inference at the interim analyses have been developed, but there are several practical issues in the implementation and application of these methods. The purpose of this work is to describe these issues and present potential solutions. We describe how to specify information-time at interim analyses when analyzing longitudinal endpoints. We demonstrate software to incorporate longitudinal endpoints into commercial group sequential design packages (PEST, EaST, Splus SeqTrial). We discuss how to estimate the information at interim analyses and thereby incorporate flexibility in the number and timing of interim analyses with longitudinal endpoints. We also discuss approaches to minimizing the potential for bias that is introduced when the actual measurement times differ from the pre-trial plan. We illustrate the application of these methods in the context of a group sequential design for a clinical trial of treatment for the symptoms of peripheral arterial disease.

e-mail: john.kittelson@uchsc.edu

CONFIDENCE INTERVAL ESTIMATION OF THE SUCCESS RATE IN EXACT GROUP SEQUENTIAL DESIGNS

Jihnhee Yu * , Roswell Park Cancer Institute;
James L. Kepner, SUNY at Buffalo

The sample space for an exact group sequential design can differ from that of the UMP one-stage design because of the chance of early stopping. Based on this, we wrote computer programs to generate the pmf for the success rate in each of the 3 types of exact 2-stage designs, so called Type I, Type II, and Type III designs. Using these programs, exact confidence intervals for the success rate are obtained. In this presentation, the distribution characteristics of the estimators are described and exact confidence intervals will be presented for all three types of 2-stage group sequential designs. Each pmf will be defined. We will show the success rate estimators are MLEs. Using the asymptotic normality of the MLEs, confidence intervals similar to the Wald confidence interval can be obtained. We also obtain confidence intervals analogous to the score confidence interval that is known for good coverage for a binomial proportion. Coverage probabilities between these different methods for each type of 2-stage design are compared.

e-mail: yuj@roswellpark.org

SOFTWARE FOR DESIGN AND ANALYSIS OF GROUP SEQUENTIAL CLINICAL TRIALS WITH MULTIPLE PRIMARY ENDPOINTS

Shuangge Ma*, University of Wisconsin
Michael R. Kosorok, University of Wisconsin
Thomas D. Cook, University of Wisconsin

In many phase III clinical trials, it is desirable to separately assess the treatment effects on more than one primary endpoint. C++ programs are provided for the computation of sample size, critical boundaries and test statistics for group sequential clinical trials with one or two primary endpoints. Our programs cover both the design and analysis phases. The computations are proper for any trial based on normally distributed test statistics, including those in which patients give a continuous response and in survival studies. All computations are based on the methodology proposed by Kosorok et al (2002). This approach resolves the problems of how to stop early when the treatment effect is clear in all endpoints and how to control the many possible error rates for concluding wrong hypotheses. We apply our software to the design and interim analysis of the Copernicus trial, which tests the effect of the beta-blocker carvedilol on the survival of patients with severe heart failure. It is shown that with certain marginal α and β spending functions, we could conclude a statistically significant beneficial treatment effect in both endpoints at the fourth interim analysis.

e-mail: shuangge@stat.wisc.edu

PROGRAMMING STRATEGIES FOR MODERATELY COMPLICATED CLINICAL RESEARCH STUDIES

Martin C. Weinrich * , University of Louisville

In much academic research collaboration, organizing data for analysis is quite simple. In large clinical trials, preparing the data is a complex task involving an entire team that follows a formal protocol. But certain moderately complex studies involve data management challenges for the statistician whose complexity far exceeds that of the analysis. This paper draws on examples from a multi-center clinical trial of congestive heart failure that used MS Access to acquire data and SAS for statistical analyses, and presents suggestions for dealing with such data: - Modularize the programming. Specifically, separate steps that build the dataset(s) for analysis from steps that perform statistical analyses. - Use the %INCLUDE statement to help combine modules as needed. - Organize files and folders so that code segments used by several analyses are not stored as duplicates. - Use comments to document the source code. - Use headers and footers that include the filenames of the source code and input dataset, and the run date. - Use macro variables to control printed output and program branching. - Use macros to simplify coding and make the program easier to read. - Save bad observations in separate datasets instead of discarding them or printing notes in the program log.

e-mail: martin.weinrich@louisville.edu

SIMULTANEOUS GROUP SEQUENTIAL ANALYSIS OF RANK-BASED AND WEIGHTED KAPLAN-MEIER TESTS FOR PAIRED CENSORED SURVIVAL DATAAdin-Cristian Andrei* , University of Michigan
Susan Murray, University of Michigan

This research develops strategies for sequentially monitoring differences in survivorship between two paired groups using a new class of non-parametric tests based on functionals of standardized paired weighted log-rank (\$PWLR\$) and standardized paired weighted Kaplan-Meier (\$PWKM\$) tests. Since during the trial these two tests may alternately assume the position of the most extreme statistic, by monitoring \$PEMAX\$, which is the maximum between the absolute values of the standardized \$PWLR\$ and \$PWKM\$, one may combine the advantages offered by rank-based and non rank-based paired testing paradigms. Simulations indicate that when monitoring treatment differences using \$PEMAX\$, the type I error is properly maintained and the testing procedure is nearly as powerful as the more advantageous of the two tests, in proportional hazards (PH) as well as non-PH situations. Hence, \$PEMAX\$ is more robust at preserving power and type I error than individually monitored \$PWLR\$ and \$PWKM\$, while maintaining a reasonably simple approach to design and analysis of results. An example from the Early Treatment Diabetic Retinopathy Study is given.

e-mail: andreaia@umich.edu

TIMING OF FUTILITY ANALYSES FOR ‘PROOF OF CONCEPT’ TRIALS

A. Lawrence Gould * , Merck Research Laboratories

Timing Of Futility Analyses For ‘Proof Of Concept’ Trials

“Proof of Concept” trials often are carried out to determine if a treatment is biologically active. With such trials, it may be useful to examine information accumulated at some point during the trial before completion to estimate the probability that the null hypothesis of no treatment effect will be rejected on completion of the trial. If this probability is low, the trial could be terminated and resources redirected more productively. If the probability is high, some time might be gained by undertaking the design of pivotal trials to demonstrate the benefits of a new treatment. When should the interim evaluation occur? This investigation evaluates how the timing of the interim evaluation affects the ability to reach a decision to stop or continue as a function of various assumptions about the true parameters and about the interim outcomes. The key finding is that there is no point to carrying out an evaluation before accumulating about 40% of the planned observations. Consequently, if trial costs are mostly at startup or for recruitment of subjects for extended periods of observation, it is unlikely that much benefit will be realized by considering the possibility of early termination for futility.

e-mail: goullda@merck.com

PERFORMANCE OF SOME MULTIPLE TESTING PROCEDURES TO COMPARE THREE DOSES OF A TEST DRUG AND PLACEBO

Lan Kong*, University of Pittsburgh
Gary G. Koch, University of North Carolina

A study design with multiple doses of a test drug and placebo is frequently used in clinical drug development. Multiplicity issues arise from performing multiple comparisons of treatment groups with placebo and among multiple treatment groups. With an increasing number of comparisons, the global type I error can be inflated when multiplicity is ignored; or the power will be reduced when adjustments for multiplicity are too conservative. In this paper, we propose some alternative strategies to manage the multiplicity and compare the proposed procedures with some traditional testing procedures through the simulated data representing various patterns of treatment differences. The purpose is to identify which methods perform better than the others and under what conditions. We focus on a study design with a placebo and three doses of a test drug (high, medium and low dose). In addition, we assume that the outcome of interest follows a normal distribution. Simulation results indicate that the knowledge of the dose response relationship is crucial to determine which method to be used for better performance. Some methods are more robust than the others with respect to various patterns of the alternative hypothesis.

e-mail: lkong@pitt.edu

VISION SYSTEMS ENGINEERING AND ANALYSIS – A UNIFIED PERSPECTIVE

Visvanathan Ramesh*, Siemens Corporate Research

An important factor for a successful deployment of image analysis systems is a good understanding of the limits of the algorithms involved and how it affects overall system performance. This talk will give an overview of the current status in statistical modeling and performance analysis of video analysis systems. The systems engineering methodology described in the talk involves two main steps: statistical modeling or performance characterization of component algorithms (component identification) and application domain characterization. Component identification involves the derivation of the deterministic and stochastic behavior of each module. This entails the specification of the ideal model and an error model in the input and relating their parameters to the output ideal model and error model parameters. The essence of the methodology is that each sub-step used in a vision system is treated as an estimator and therefore the estimator's behavior has to be characterized in terms of its distribution as a function of the input samples and error distribution parameters. Application domain characterization is a learning or estimation step wherein the restrictions on the application data relevant to the task at hand are specified in terms of prior distributions of parameters relevant to the algorithm/system representation chosen. These prior distributions can be viewed as specifying the range of possible images for the given application. The average or worst case performance of the system can be determined by combining the Component Identification steps and the Application domain modeling steps.

EMPIRICAL STATISTICAL MODELS FOR OBJECT RECOGNITION IN PHOTOGRAPHS

Henry Schneidermann*, Carnegie Mellon University

Automatically finding various types of objects (e.g., faces, automobiles) in photographic images is a challenging task. It requires accurate statistical models of visual appearance for each type of object and for the world at large. This talk will introduce the challenges of this statistical modeling problem. We will then describe the motivation and construction of various statistical models that we have used in practice.

DETECTING UNUSUAL HUMAN BEHAVIOR IN VIDEO

Jianbo Shi*, University of Pennsylvania

Humans can extract essential characteristics of behavior in a video quickly and easily. However, humans are limited in their memory capacity and ability to stay fully focused over long periods of time. Imagine you are given a large set of video, possibly thousands of hours long. You are asked to analyze the video and to find unusual events. By definition unusual events are rare. They are difficult to describe, hard to predict and can be quite subtle. Any system that can detect such unusual events/behaviors must sift through extremely large amounts of minute statistical details to detect a few relevant bits. Furthermore, such a system must be able to deal with a bewilderingly large number of normal activities, which can significantly out-number that of unusual ones. We present an unsupervised machine learning technique for detecting unusual activity in a large video set, using multiple simple features. No complex activity models and no supervised feature selections are used. Computationally, a video is divided into equal length segments and simple prototype features are extracted from each frame. A prototype-segment co-occurrence matrix is computed. Motivated by a similar problem in document-keyword analysis, we seek a *correspondence* relationship between prototypes and video segments, which satisfies the transitive closure constraint. We show that an important sub-family of correspondence functions can be reduced to co-embedding prototypes and segments to N-D Euclidean space. We prove that an efficient, globally optimal algorithm exists for the co-embedding problem. Experiments on a variety of complex real life video have validated our approach.

56. New Developments on Estimation Equations

DISTRIBUTION-BASED MARGINAL REGRESSION MODELS FOR LONGITUDINAL DATA

Jianhua Z. Huang, University of Pennsylvania;
Colin O. Wu * , National Heart, Lung and Blood Institute;
Lan Zhou, University of Pennsylvania

The widely used generalized estimating equations (GEE) are indispensable tools for estimating conditional mean-based regression models. Although successful in many applications, the conditional mean-based regression approach is potentially inadequate when the conditional mean is an inappropriate measure for the scientific question being investigated. Such situations may arise when (a) the outcome variable has a highly skewed or non-Gaussian conditional distribution whose characteristics can not be well captured by the mean, or (b) the outcome variable has ordinal scales or a mixed distribution, so that its mean does not have a meaningful interpretation. In this talk, we will discuss a class of marginal regression models for longitudinal data based on conditional distributions, which provides an alternative to the traditional conditional mean-based regression. Estimating equations based on U-statistics are used for estimation of the proposed models.

e-mail: wuc@nhlbi.nih.gov

ESTIMATING EQUATIONS FOR LONGITUDINAL DATA WITH TIME-DEPENDENT COVARIATES

Dylan S. Small*, University of Pennsylvania
Tze L. Lai, Stanford University

An appealing feature of the generalized estimating equation (GEE) approach to the analysis of longitudinal data is its use of a ‘working hypothesis’ about the parts of the model which are considered nuisance parameters. If the working hypothesis is correct, GEE will produce efficient estimates; if the working hypothesis is incorrect, GEE will still produce consistent estimates for the parameters of interest. In the presence of time- dependent covariates, it has been noted that only certain types of working hypotheses (in particular a working hypothesis of independence between all observations) can be used if the researcher wants to maintain consistency in the presence of misspecification. We develop an approach that allows the use of more flexible and more realistic working hypotheses in analyzing longitudinal data with time-dependent covariates via estimating equations. Our approach is shown to facilitate model selection. We demonstrate the advantages of our approach through simulation studies and theoretical analysis and present an empirical application to a longitudinal study of children’s health.

e-mail: dsmall@wharton.upenn.edu

IMPLEMENTATION OF A COMPLEX CORRELATION STRUCTURE APPROPRIATE FOR REPEATED BOUTS OF MEASUREMENTS IN A LONGITUDINAL STUDY

Justine N. Shults * , University of Pennsylvania;
Carissa A. Mazurick, Merck & Co, Inc;
Richard Landis, University of Pennsylvania

GEE (Liang and Zeger, 1986) has been widely applied for several simple correlation structures. However, one limitation of this approach is that it can be difficult to implement for complex structures. In this talk we will demonstrate this difficulty in a study of Interstitial Cystitis in women, for which the study investigators were forced to adopt an ad-hoc approach, in order to implement the correlation structure most appropriate for their data. We will discuss their structure, which is appropriate when several bouts of measurements are collected on study subjects; for example, when measurements are collected on three successive days in two bouts that are each separated by six months in time. We will then describe its implementation using the method of quasi-least squares (QLS). QLS is based on GEE, but uses a different approach for estimation of the correlation parameters, which allows for easier implementation of some complex structures, including the structure we consider in this presentation.

e-mail: jshults@cceb.upenn.edu

CONSISTENCY AND INCONSISTENCY OF MAXIMUM QUASI LIKELIHOOD ESTIMATORS

Bing Li*, Pennsylvania State University

It has long been speculated that, if a parametric class of estimating equations forms a conservative vector field, then, under some conditions, the maximum point of the potential function should be a consistent estimator of the parameter. This is part of the reason for preferring a maximum quasi likelihood estimator to other solutions of the quasi-likelihood equation. However, such sufficient conditions have not been established except in special cases. In this talk I will discuss two sets of reasonably general sufficient conditions for a maximum quasi likelihood estimator to be consistent. I will also demonstrate that, if these conditions are violated then, it is possible for a maximum quasi likelihood estimator to be inconsistent. These results will then be applied to study nonconservative estimating equations and generalized estimating equations. In particular, I will discuss what types of path integral of a nonconservative estimating equation will give rise to consistent maximum.

e-mail: bing@stat.psu.edu

57. Bayesian Methods in Genomics

VARIABLE SELECTION FOR GENE EXPRESSION DATA WITH TIME-TO-EVENT RESPONSES

Naijun Sha*, The University of Texas at El Paso;
Marina Vannucci, Texas A&M University;
Mahlet Tadesse, Texas A&M University

We investigate variable selection methods for parametric survival models, with emphasis on accelerated failure time models, where a linearized form can be obtained by taking a log transformation to the survival time. We develop selection method that allow for censored data by introducing a latent ‘complete’ failure time. The methods lead to estimates of the survival as well as to the identification of the variables that affect the survival outcome.

e-mail: naijun@math.utep.edu

NEW CLASSES OF MIXTURE MODELS FOR ANALYZING DNA MICROARRAY DATA

Joseph G. Ibrahim*, University of North Carolina
Ming-Hui Chen, University of Connecticut

In DNA microarray analysis, there is often interest in isolating a few genes whose expression best discriminate between tissue types. This is especially important in cancer, where different clinicopathologic groups are known to vary in their outcomes and response to therapy. The identification of a small subset of gene expression patterns distinctive for tumor subtypes can help design treatment strategies and improve diagnosis. Toward this goal, we propose a methodology for the analysis of high-density oligonucleotide arrays. An important issue that is often disregarded in the analysis of microarray data is the quantification limits of the technology. We propose a new class of mixture models that accommodate undetected or unreliable transcripts. The model is formulated in a hierarchical Bayesian framework, which in addition to making the fit of the model straightforward and computationally efficient, allows us to borrow strength across genes. Prior distributions are formulated for this class of models. A Microarray dataset is analyzed to demonstrate the new model.

e-mail: ibrahim@bios.unc.edu

A BAYESIAN MIXTURE MODEL FOR DIFFERENTIAL GENE EXPRESSION

Kim-Anh Do, M.D. Anderson Cancer Center;
Peter Mueller*, M.D. Anderson Cancer Center;
Feng Tang, M.D. Anderson Cancer Center

We review methods and models used in nonparametric Bayesian inference, using as a motivating example the analysis of microarray data for differential gene expression. The discussion will highlight advantages and limitations of inference based on full (posterior) probability models for the relevant unknown distributions. We will compare resulting inference with ad-hoc exploratory data analysis, with a similar parametric empirical Bayes approach proposed in Efron et al. (2001), and with fully parametric finite mixture models. We will discuss problems of the empirical Bayes inference that are mitigated by a nonparametric Bayesian approach. As a specific example of nonparametric Bayesian modeling we will focus on an extension of Dirichlet process mixtures of normal distributions. As in most nonparametric Bayesian inference, efficient computation is key to a successful implementation. We will review computational problems and strategies that are relevant for a wide class of nonparametric Bayesian models. Computation in the proposed nonparametric model is not as straightforward as in the empirical Bayes scheme. But we argue that inference is no more difficult than posterior simulation in similarly flexible fully parametric mixture models.

e-mail: pm@odin.mdacc.tmc.edu

A BAYESIAN METHOD FOR SIMULTANEOUS CLASS DISCOVERY AND GENE SELECTION

Mahlet G. Tadesse*, Texas A&M University

The analysis of the high-dimensional data generated by DNA microarrays poses challenge to standard statistical methods. This has revived a strong interest in clustering algorithms. A typical goal in these analyses is the discovery of new classes of disease and the identification of relevant genes. We propose a Bayesian method for simultaneously identifying the number of clusters in the data and selecting genes that best discriminate the different groups. We illustrate the methodology with a microarray data from an endometrial cancer study.

e-mail: mtadesse@stat.tamu.edu

58. Nonparametric Function Estimation

FUNCTIONAL ADAPTIVE MODEL ESTIMATION

Gareth M. James*, Univeristy of Southern California
Bernard W. Silverman, Oxford University

In this talk I am interested in modeling the relationship between a scalar, Y , and a functional predictor, $X(t)$. I introduce a highly flexible approach called 'Functional Adaptive Model Estimation' (FAME) which extends generalized linear models (GLM), generalized additive models (GAM) and projection pursuit regression (PPR) to handle functional predictors. The FAME approach can model any of the standard exponential family of response distributions that are assumed for GLM or GAM while maintaining the flexibility of PPR. For example standard linear or logistic regression with functional predictors, as well as far more complicated models, can easily be applied using this approach. A functional principal components decomposition of the predictor functions is used to aid visualization of the relationship between $X(t)$ and Y . The FAME approach will be illustrated on the prediction of five year survival for a patient given observations of blood chemistry levels over time and the prediction of arthritis based on bone shape. I will end with a discussion of the relationships between standard regression approaches, their extensions to functional data and FAME.

e-mail: gareth@usc.edu

HIGH DIMENSION LOW SAMPLE SIZE SURVIVAL ANALYSIS

Brent Johnson, University of North Carolina at Chapel Hill;
Danyu Lin, University of North Carolina at Chapel Hill;
J. S. Marron * , University of North Carolina at Chapel Hill

Distance Weighted Discrimination is a useful method for classification in High Dimension Low Sample Size contexts. An important application is micro-array data. Some appropriately weighted versions of Distance Weighted Discrimination are developed for analyzing survival related outcomes in a breast cancer data set.

e-mail: marron@email.unc.edu

RECONSTRUCTION, ALIGNMENT, AND COMPARISON OF CONTINUOUS TIME COURSE EXPRESSION PROFILES

Ziv Bar-Joseph, Carnegie Mellon University
Georg Gerber, Massachusetts Institute of Technology
Itamar Simon, Hebrew University
David Gifford, Massachusetts Institute of Technology
Tommi Jaakkola* , Massachusetts Institute of Technology

Many biological processes need to be understood as dynamical systems. Time course expression profiles represent one prominent source of information about unfolding biological processes such as the cell cycle. The available time course profiles are, however, sparse, noisy, irregularly sampled, and typically without repetitions. The profiles and the biological processes are affected by variations in the experimental conditions as well as the protocols involved. Moreover, few if any parametric assumptions are available for constraining the continuous curves that underlie the time course measurements. I will discuss a) spline reconstruction of continuous curves underlying the expression profiles, b) parametric alignment of the curves across different datasets and conditions, and c) the significance of deviations between the aligned profiles.

e-mail: tommi@ai.mit.edu

MODEL SELECTION IN IRREGULAR PROBLEMS: APPLICATION TO MAPPING QTL

David O. Siegmund*, Stanford University

Two methods of model selection are discussed for change-point like problems, especially those arising in genetic linkage analysis. The first is a method that selects the model with the smallest p-value, while the second is a modification of the Bayes Information Criterion (BIC). The methods are compared theoretically and on examples from the literature. For these examples, they are roughly comparable although the p-value based method is somewhat more liberal in selecting a high dimensional model.

e-mail: dos@stat.stanford.edu

SPATIAL AND SPATIO-TEMPORAL HIERARCHICAL MODELING FOR THE ANALYSIS OF SPATIAL DATA

Alan E. Gelfand*, Duke University

Currently, monitoring sites around the world gather enormous amounts of temporal data recording levels for a variety of environmental pollutants. Besides temporal dependence, it is expected that such data will exhibit spatial dependence whose nature will depend upon the spatial scale being used. Spatio-temporal modeling attempts to describe the overall nature of this dependence. The literature here is by now considerable. Indeed, even the literature employing a Bayesian modeling perspective is growing rapidly. We attempt a brief review of this latter literature in the context of a fairly general development of spatial and spatio-temporal hierarchical modeling. We then turn to two useful tools, coregionalization and directional derivative processes, which enable investigation of multiple pollution surfaces and gradients to these surface at particular locations and in particular directions. We illustrate with several pollution datasets.

e-mail: alan@stat.duke.edu

NORMALIZATION AND SIGNIFICANT ANALYSIS OF CDNA MICRO-ARRAYS

Jianqing Fan * , Princeton University

The quantitative comparison of two or more microarrays can reveal, for example, the distinct patterns of gene expression that define different cellular phenotypes or the genes that are induced in the cellular response to certain stimulations. Normalization of the measured intensities is a prerequisite of such comparisons. However, a fundamental problem in cDNA microarray analysis is the lack of a common standard to compare the expression levels of different samples. Several normalization protocols have been proposed to overcome the variabilities inherent in this technology. We have developed a normalization procedure based on within-array replications via a Semi-Linear In-slide Model (SLIM), which adjusts objectively experimental variations without making critical biological assumptions. The significant analysis of gene expressions is based on a newly developed weighted t- statistic, which accounts for the heteroscedasticity of the observed log-ratios of expressions, and a balanced sign permutation test. We illustrated the use of the newly developed techniques in a comparison of the expression profiles of neuroblastoma cells that were stimulated with a growth factor, macrophage migration inhibitory factor.

e-mail: jqfan@princeton.edu

BAYESIAN ANALYSIS OF SINGLE MOLECULE EXPERIMENTS

Samuel Kou * , Harvard University;

Sunney Xie, Harvard University;

Jun Liu, Harvard University Harvard School of Public Health

Recent technological advances allow scientists for the first time to follow a biochemical process on a single molecule basis, which, unlike traditional macroscopic experiments, raises many challenging data-analysis problems and calls for a sophisticated statistical modeling and inference effort. This paper provides the first likelihood-based analysis of the single-molecule fluorescence lifetime experiment, in which the conformational dynamics of a single DNA hairpin molecule is of interest. The conformational change is modeled as a continuous-time two-state Markov chain, which is not directly observable and has to be inferred from changes in photon emissions from a dye attached to the DNA hairpin molecule. In addition to the hidden Markov structure, the presence of molecular Brownian diffusion further complicates the matter. We show that closed form likelihood function can be obtained and a Metropolis-Hastings algorithm can be applied to compute the posterior distribution of the parameters of interest. The data augmentation technique is utilized to handle both the Brownian diffusion and the issue of model discrimination. Our results increase the estimating resolution by several folds. The success of this analysis indicates there is an urgent need to bring modern statistical techniques to the analysis of data produced by modern technologies.

e-mail: kou@stat.harvard.edu

EMPIRICAL LIKELIHOOD RATIO WITH ARBITRARY CENSORED/TRUNCATED DATA BY CONSTRAINED EM ALGORITHM

Min Chen * , University of Kentucky;
Jingyu Luan, University of Kentucky;
Mai Zhou, University of Kentucky

Empirical likelihood ratio method (Thomas and Grunkmier 1975, Owen 1988) is a general nonparametric inference procedure with many nice properties. But when we apply this method to arbitrary censored/truncated data with various parameters, the computation of empirical likelihood ratio is non-trivial. Turnbull (1976) proposed an EM type method for nonparametric estimation of the distribution function when the data are arbitrary censored/truncated. The EM algorithm was later modified by Frydman (1991) and Alioum (1996). Zhou (2002) proposed a modified self-consistency/EM algorithm to compute a class of Empirical Likelihood ratio with arbitrary censored/truncated data and constrains on the mean of the unknown distribution. We present in this paper a self-consistency/EM algorithm to compute, in the one sample case, maximum empirical likelihood with a type of constrains on the hazard of the unknown distribution. We also extend the algorithm to the proportional hazard regression analysis. This extension of EM algorithm allows for calculating the maximum empirical likelihood under the hypothesis about regression coefficients and/or under hypothesis about the baseline hazard of the unknown distribution. Thus the empirical likelihood ratios for hypothesis testing can be numerically found. Several examples will be presented.

e-mail: minchen@ms.uky.edu

ON THE ANALYSIS OF INFORMATIVELY INTERVAL-CENSORED DISCRETE FAILURE-TIME DATA: A SENSITIVITY ANALYSIS APPROACH

Michelle D. Shardell*, Johns Hopkins University
Daniel O. Scharfstein, Johns Hopkins University
Noya Galai, Johns Hopkins University
David Vlahov, Johns Hopkins University

In this talk, we discuss estimation of the effect of a covariate on the hazard of a discrete failure-time random variable, where the data are informatively interval censored. Sun (1997) developed a method for estimating the hazard ratio in the continuation ratio model, assuming non-informative interval censoring. We extend his EM algorithm method to handle informative censoring by eliciting information from subject-matter experts about the relationship between the censoring process and the failure time. Since this relationship is non-identifiable, our method is repeated over a range of assumptions to assess the sensitivity of hazard ratio estimates. Our method is applied to data from the AIDS Link to Intravenous Experience (ALIVE) study to estimate the hazard ratio of HIV infection comparing needle sharers to non-sharers.

e-mail: mshardel@jhsph.edu

PARAMETRIC ESTIMATION FOR ADDITIVE RATE REGRESSION MODELS

M. Brent McHenry * , University of Pittsburgh; Stuart
R. Lipsitz, Medical University of South Carolina

For failure time outcomes, modeling the hazard rate as an exponential function of covariates is by far the most popular. However, in the last few decades, additive hazard rate regression models have received some attention, in which the hazard rate is modeled as a linear function of the covariates. Popular fully parametric distributions include the exponential and piecewise exponential. In this paper, for an additive rate regression model in which the distribution of the failure time is exponential or piecewise exponential, we show that the maximum likelihood estimates (MLE) can be obtained using a Poisson linear model, without any additional programming or iteration loops. As a result, the MLEs can be obtained in any generalized linear models program. We apply the method to two data sets, a study of lupus nephritis and a study of leukemia, in which the additive hazard rate regression model appears more appropriate.

e-mail: mbmst29@pitt.edu

EMPIRICAL LIKELIHOOD RATIO WITH RIGHT-CENSORING AND LEFT-TRUNCATION DATA

Jingyu Luan* , University of Kentucky
Min Chen, University of Kentucky
Mai Zhou, University of Kentucky

Empirical likelihood ratio (ELR) method (Owen 1988, 2001) is a general non-parametric inference procedure with many nice properties. In the literature, there are results of empirical likelihood with censored data, and there are also results with purely truncated data (Owen 2001, Li 1995). But the empirical likelihood theorem for right censoring AND left truncation data is still not available. This paper will illustrate that $-2\log ELR$ with hazard type constraint follows an asymptotic chi-square distribution for right censoring and left truncation data in one sample situation. Examples and simulations will also be presented.

e-mail: luanjy@ms.uky.edu

SURVIVAL ANALYSIS USING AUXILIARY VARIABLES VIA NONPARAMETRIC MULTIPLE IMPUTATION

Chiu-Hsieh Hsu^{*}, University of Arizona;
Jeremy M.G. Taylor, University of Michigan;
Susan Murray, University of Michigan

We develop an approach, based on multiple imputation, that estimates the marginal survival distribution in survival analysis using auxiliary variables to recover information for censored observations. To conduct the imputation, we use two working proportional hazards models to define an imputing risk set. One model is for the event times and the other for the censoring times. Based on the imputing risk set, two nonparametric multiple imputation methods are considered: a risk set imputation, and a Kaplan-Meier imputation. For both methods a future event or censoring time is imputed for each censored observation. In a situation with a categorical auxiliary variable, we show that with a large number of imputes the estimates from the Kaplan-Meier imputation method correspond to the weighted Kaplan-Meier estimator. We also show that the Kaplan-Meier imputation-based method is robust to misspecification of either one of the two working models. In a simulation study with time independent and time dependent auxiliary variables, we show that the use of the multiple imputation methods can improve the efficiency of estimators and reduce bias due to dependent censoring. The Kaplan-Meier imputation method is shown to outperform the risk-set imputation approach. We apply the approach to AIDS clinical trial data comparing ZDV and placebo, in which CD4 count is the time-dependent auxiliary variable.

e-mail: phsu@azcc.arizona.edu

REDUCING BIAS IN PARAMETER ESTIMATES FROM STEPWISE REGRESSION IN PROPORTIONAL HAZARDS REGRESSION WITH RIGHT-CENSORED DATA

Chang-Heok Soh^{*}, Dana-Farber Cancer Institute and Harvard School of Public Health
David P. Harrington, Dana-Farber Cancer Institute and Harvard School of Public Health

When variable selection and model fitting are conducted on the same dataset, competition for inclusion in the model biases coefficient estimators away from zero. In proportional hazards regression with right-censored data, competition bias inflates the magnitude of parameter estimate of selected parameters, while omission of other variables may shrink coefficients toward zero. This paper explores the extent of bias in parameter estimates from stepwise regression in proportional hazards regression and proposes a bootstrap method, similar to those proposed by Miller (1984) for linear regression, to move the conditional mean of a parameter estimate, given that a variable is selected, closer to the population value of the parameter. We also use bootstrap methods to estimate the standard error of the adjusted estimators. Preliminary simulation results show that, for binary covariates, the bias in estimated coefficients could exceed 300% and the bootstrap method corrects for 18-62% of the bias. For normal covariates, the bias is smaller and the correction is less reliable. We also explore the effect of the adjustment on estimates of log relative risk, given the values of the covariates in a selected model. The proposed methods are compared with other shrinkage methods, such as the approximate method proposed by Harrell.

e-mail: csoh@hsph.harvard.edu

ON SEMIPARAMETRIC TRANSFORMATION CURE MODELS

Wenbin Lu * , North Carolina State University;
Zhiliang Ying, Columbia University

A general class of semiparametric transformation cure models is studied for the analysis of survival data with long-term survivors. It combines a logistic regression for the probability of event occurrence with the class of transformation models for the time of occurrence. Included as special cases are the proportional hazards cure model (Farewell, 1982; Kuk and Chen, 1992; Sy and Taylor, 2000; Peng and Dear, 2000) and the proportional odds cure model. Generalised estimating equations are proposed for parameter estimation. It is shown that the resulting estimators are asymptotically normal, with variance-covariance matrix that has a closed form and can be consistently estimated by the usual plug-in method. Simulation studies show that the proposed approach is appropriate for practical use. An application to data from a breast cancer study is given to illustrate the methodology.

e-mail: lu@stat.ncsu.edu

61. Experimental Design

TWO AT A TIME? OR ALL AT ONCE?

Donald J. Schuirmann*, U.S. Food and Drug Administration

Suppose we have a bioequivalence study with three treatments – A, B, and C – and the objective of the study is to make pairwise comparisons among the treatments. Suppose further that treatment C is different in kind from A and B, so that the assumption of homogeneous variance among the three treatments is questionable. One way to do the analyses, under normality assumptions, is Two at a Time – e.g., to test hypotheses about A and B, use only the data from A and B. Another way is All at Once – include the data from all three treatments in a single analysis, making pairwise comparisons within this analysis. If the assumption of homogeneous variance is correct, the All at Once approach will provide more d.f. for estimating the common variance, resulting in increased power. If the variance of C differs from that of A and B, the All at Once approach may have reduced power or an inflated type I error rate, depending on the direction of the difference in variances. I will attempt to quantify the difference between the two approaches for both the comparison of A to B and the comparison of A or B to C. Both parallel and crossover designs will be considered.

e-mail: schuirmann@cder.fda.gov

ESTIMATING THE VARIANCE COMPONENTS IN A REPLICATED CROSSOVER STUDY

Guoyong Jiang^{*}, Cephalon, Inc.;
Jianrong Wu, St. Jude Children's Research Hospital

For a uniform within sequence replicated crossover experiment, Chinchilli and Esinhart (1996) derived closed-form expressions for the maximum likelihood (ML) and restricted maximum likelihood (REML) estimators of the variance components. However, their ML and REML estimators are not straightforward. These estimators are based on transformed variables, but not on elements of the sample covariance matrix calculated from the original observations. In this article, we exploit the characteristics of the covariance matrix for the vector variable of measurements on each subject, and derive simple and intuitive expressions for the ML and REML estimators of the variance components. These estimators are expressed as known linear combinations of elements of the sample covariance matrix from the original observations. They are easy to understand and simple to calculate. We illustrate the new method by applying it to an example.

e-mail: jjiang@cephalon.com

ON THE USE OF A CROSSOVER DESIGN IN VACCINE RESEARCH

Jonathan B. Skinner^{*}, Ph.D., Wyeth Vaccines Research

On the Use of a Crossover Design in Vaccine Research The Balaam design is an extension of the two-sequence, two-period, two-treatment design (sequences: AB, BA) to four sequences: AB, BA, AA, BB. The two sequences AA and BB are usually added in order to study the direct treatment effect with adjustment for period and carryover effects on a within-subject basis. In vaccine research the Balaam design can be used to study carryover effects as the primary interest. This presentation will explain the analysis of a study of vaccine efficacy that used the Balaam design by comparing it with the textbook analysis of the Balaam design.

e-mail: skinnej@wyeth.com

CARRYOVER EFFECTS EXAMINED IN COLLECTIONS OF CROSSOVER STUDIES

Mary E Putt * , University of Pennsylvania

The possibility of carryover is an important consideration in the design of any crossover study, and can cause an investigator to abandon the design altogether. In crossover studies, carryover is defined as the lingering effect of one treatment into the subsequent period. It is well known that the test for carryover in any individual study has low power. However examining a collection of studies for evidence of carryover should improve the power for detecting this important effect. Here we develop a general approach for estimating the power to detect carryover effects in collections of crossover studies. For studies designed to detect a treatment effect, the power to detect clinically relevant levels of carryover was typically less than 15%, and the number of studies needed to detect an effect of this size was prohibitive. For two-treatment two-period bioequivalence studies, the power under certain conditions was sufficient for carryover to be detected with a moderate number of studies. However these studies would only have reasonable sample size if the intraclass correlation was high and the coefficient of variation was small. Unequivocal conclusions about the presence or absence of carryover from collections of studies appear unlikely.

e-mail: mputt@cceb.upenn.edu

A BETA-BINOMIAL MODEL FOR CROSS-OVER STUDIES OF CORRELATED BERNOULLI OUTCOMES, WITH AN APPLICATION TO CONDOM FAILURE DATA

Douglas J. Taylor*, Family Health International
Natalie W. Cheung, Family Health International
Mark A. Weaver, Family Health International

Evaluating the effect of treatment in a crossover study with correlated Bernoulli outcomes is a common research objective. In studies of condom functionality, for example, an indicator of device failure may be reported for each act of intercourse using a standard or experimental condom type. Enrolled couples typically use more than one condom of each type, leading to correlated outcomes. Two popular analysis methods for these types of data are Generalized Estimating Equations with logit-link, in which the odds ratio is a population average measure, and the logit-normal random effects model, in which the odds ratio has a couple-specific interpretation. If the effect on the odds of failure associated with a couple switching from one treatment to another is of most interest, then a random effects model is preferable. An alternative random effects model, the beta-binomial, has been used in various contexts. Assuming that the couple-specific probabilities of failure follow a beta distribution leads to a simple marginal distribution function when no within-couple effects are specified. In this paper we extend the beta-binomial model to account for a within-couple treatment effect. Results of fitting the new model to a set of condom failure dataset are presented and compared with those from the more traditional logit-normal model.

e-mail: dtaylor@fhi.org

OPTIMAL DESIGNS FOR ORDINAL OUTCOMES USING BETA REGRESSION MODELS

Yuehui Wu * , University of Pennsylvania;
Valeri V. Fedorov, GlaxoSmithKline;
Kathleen J. Propert, University of Pennsylvania

There is much literature on optimal design for models for categorical outcomes. Although there is ample literature on models and analysis of ordered categorical data, there is limited information on optimal design for ordinal data. Also, some ordinal data models may have difficulties in either fitting or interpretation if there are a large number of categories. We first develop a parsimonious beta regression model for ordinal outcomes that assumes the responses are from a beta distribution with parameters α and β . These parameters are then related to covariates such as dose in a clinical trial. We then explore locally optimal designs for the beta regression model which minimize the variances of maximum likelihood estimators of the unknown parameters of interest. Optimal designs and sensitivity to misspecification of model parameters are examined using a candidate points searching algorithm. The proposed method is applied to data from a clinical trial to illustrate the methods.

e-mail: ywu@cceb.upenn.edu

CROSS TRIAL STATISTICAL INFERENCE: BRIDGING VS. NON-INFERIORITY SCENARIOS

Sue-Jane Wang*, U.S. Food and Drug Administration

According to ICH E-5, often one of the primary objectives of a clinical bridging study conducted in a new region after a pharmaceutical product has been approved or marketed in the original region is to demonstrate that the effect of the approved new treatment in the new region is similar or equivalent to that in the original region. This may be an important regulatory consideration for bridging study applications in the new region. To achieve this objective, the design and analysis of a clinical bridging study scenario entails cross-trial statistical inference. The cross trial inference feature of the bridging evaluation is not dissimilar to the non-inferiority cross trial inference in which the analysis is often to demonstrate that the experimental treatment preserves a specified fraction of the effect of the active control. However, the concept of statistical inference, its interpretation and the implication are distinct between the two scenarios. In this work I highlight the major differences in cross trial inference between the bridging study scenario and the non-inferiority testing scenario. The discussion will also consider innovative adaptation ideas in use of genetic characteristics to the bridging concept when special populations are not easy to recruit.

e-mail: wangs@cder.fda.gov

MAPPING QUANTITATIVE TRAIT LOCI WITH CENSORED
OBSERVATIONS

Guoqing Diao * , University of North Carolina at Chapel Hill;
Danyu Lin, University of North Carolina at Chapel Hill;
Fei Zou, University of North Carolina at Chapel Hill

The existing statistical methods for mapping quantitative trait loci (QTLs) assume that the phenotype follows a normal distribution and is fully observed. These assumptions may not be satisfied when the phenotype pertains to the survival time or failure time, which is commonly subject to censoring due to random loss of follow-up or limited duration of the experiment. In this paper, we propose an interval mapping approach for censored failure time phenotypes. We formulate the effects of QTLs on the failure time through proportional hazards models and develop efficient likelihood-based inference procedures. In addition, we show how to assess genome-wide statistical significance. The performance of the proposed methods is assessed via extensive simulation studies. An application to a mouse cross is provided.

e-mail: gdao@bios.unc.edu

BAYESIAN MODEL AND VARIABLE SELECTION APPROACHES TO MAPPING MULTIPLE QUANTITATIVE
TRAIT LOCI

Kyoungmi Kim* , University of Alabama at Birmingham
Nengjun Yi, University of Alabama at Birmingham
Alfred Bartolucci, University of Alabama at Birmingham
David B. Allison, University of Alabama at Birmingham

In this study, we discuss Bayesian approaches to mapping multiple quantitative trait loci (QTL) by analyzing all markers of the whole genome. When the markers are evenly spaced, the locations of QTL can be reasonably detected when we assume that at most one QTL is located in the marker interval. However, the marker density is usually irregularly spaced. In this situation, we propose to substitute markers regularly by positions in the marker intervals if we assume that at most one QTL is on any marker interval. Our method is based on Bayesian variable selection methodology adapted for multiple regression in a hierarchical normal mixture model, where latent indicators are used to identify the multiple QTL. In this framework the promising QTL with significant effects can be identified with high posterior probability included in the model. The Markov Chain Monte Carlo algorithms are employed to simulate the posterior samples of all unknowns, including the number of QTL, the locations and genetic effects for all identified QTL, as well as other model parameters. The proposed method is demonstrated using both simulated and real data.

e-mail: kkim@ms.soph.uab.edu

AN EMPIRICAL BAYES APPROACH FOR EXPRESSION TRAIT LOCI (ETL) MAPPING

Christina Kendzior^{*}, University of Wisconsin

The development of statistical methods for mapping quantitative traits has received considerable attention. Effective methods now exist to account for different types of crosses or family structures, different kinds of phenotype, the presence of multiple genes affecting the trait and their genetic interactions, and the multiple testing issues that arise from tests at many markers. A number of groups have recently applied these QTL methods to the problem of mapping mRNA abundance measurements generated from microarrays by considering each individual transcript as a quantitative trait. However, most QTL mapping methods were developed to address the case where a small number of traits (oftentimes, just one) are being mapped. In expression trait loci (ETL) mapping, thousands of traits are considered simultaneously and the repeated application of individual tests is not the most efficient strategy. I will present an empirical Bayes modeling approach to enable ETL mapping. The inefficiency of the single trait method and the utility of the proposed method are demonstrated using microarray and genotype data from an F2 mouse cross in a study of diabetes. This is joint work with Meng Chen, Hong Lan, and Alan Attie.

e-mail: kendzior@biostat.wisc.edu

A NONPARAMETRIC TEST AGAINST STOCHASTIC ORDERING WITH APPLICATIONS IN QTL MAPPING

Fei Zou^{*}, University of North Carolina at Chapel Hill
Lei Nie, UMBC

In practice, such as in many linkage studies, when comparing distributions of a gene controlling the trait of interests, we may reasonably assume the distributions are stochastically ordered. In this paper, a distribution-free test for stochastic ordering in genetic mapping has been proposed, with the underlying distributions of quantitative trait loci (QTL) assumed to be stochastically ordered and otherwise unspecified. Least square approaches are proposed to test the existence of the QTL and estimate its effects simultaneously. As a byproduct, the confidence interval of the putative QTL locus can be obtained.

e-mail: fzou@bios.unc.edu

A MODEL SELECTION BASED INTERVAL MAPPING METHOD FOR AUTOPOLYPLOIDS

Dachuang Cao *, Purdue University;
Bruce A. Craig, Purdue University;
Rebecca W. Doerge, Purdue University

While extensive progress has been made in quantitative trait locus (QTL) mapping for diploid species, progress in QTL mapping of polyploids has been limited due to the polyploid's complex genetic architecture. To date, QTL mapping in polyploids has focused mainly on tetraploids with dominant and/or co-dominant markers. In this paper, we have extended this to include any even ploidy level based on dominant marker systems. Our approach first selects the most likely chromosomal marker configurations using a Bayesian selection criterion, and then fits an interval mapping model to the candidate models. The likelihood function and parameters including marker dosages, loci linkage phase, QTL location, and QTL effect, are estimated via the EM algorithm. Although presented via pseudo-doubled backcross experiments, our approach can be readily extended to other breeding systems. Our method is applied to single-dose restriction fragment autotetraploid alfalfa data. Several putative QTL associated with winter hardiness traits are identified. The performance of our method is then investigated through simulation studies.

e-mail: caod@stat.purdue.edu

A UNIFIED SEMIPARAMETRIC FRAMEWORK FOR QTL ANALYSES, WITH APPLICATION TO 'SPIKE' PHENOTYPES

Chunfang Jin*, University of Wisconsin-Madison

This paper proposes a general semiparametric model for multiple QTL analyses of complex phenotypes in backcross and intercross designs. The model provides tests about genetic hypotheses, such as additivity, dominance and epistasis, that do not require specifying the form of the phenotypic distribution. This contrasts with previous approaches based on transformations to normality and generalized linear models, which require careful consideration of the phenotypic distribution. Inferences involve extensions of the partial and conditional likelihoods of Zou et al. (2002) and Zou and Fine (2002) from single QTL backcross analyses. To facilitate genome screens, a novel resampling method is proposed which is similar in spirit to the popular permutation tests (Churchill and Doerge 1994). Its main advantages are that it is broadly applicable to multiple QTLs with non-normal phenotypes and achieves a substantial reduction in computational burden. A case-study of "spike" data on the genetic influences to recovery from listeria infection illustrates the practical utility of the proposed methods when compared with existing interval mapping methods using parametric models.

e-mail: cjin@wisc.edu

A TWO-STAGE HIERARCHICAL MODEL FOR DETECTING THE EPISTATIC CONTROL OF MATERNAL-OFFSPRING QTL OVER ENDOSPERM DEVELOPMENT

Yuehua Cui * , University of Florida;
George Casella, University of Florida;
Rongling Wu, University of Florida

Endosperm, a result of double fertilization in flowering plants, plays a vital role in the early stage of embryo development. The expression of most quantitative traits in endosperm results from direct (offspring) and indirect (maternal) genetic effects and involves complex interactions between quantitative trait loci (QTL) from the maternal and offspring genomes. To study the effects of maternal and endosperm QTL interaction over seed development, a two-stage hierarchical statistical model implemented with the maximum likelihood and EM algorithm is derived. The model incorporating both the maternal and offspring marker information can improve the accuracy and precision of genetic mapping. Extensive simulations under different experimental designs, heritability levels, and sample sizes were performed to investigate the statistical properties of the model. The QTL location and parameters are better estimated when two QTL are located at different intervals than when they are located at the same interval. Also, the additive effect of the offspring QTL is better estimated than the additive effect of the maternal QTL. Our model provides a powerful tool to study the genetic epistatic effects of the maternal and offspring genomes on endosperm traits in agricultural crops and can also be used to study the genetic significance of double fertilization in the evolution of higher plants.

e-mail: cuiyh@ufl.edu

63. Health Service and Epidemiologic Methods

ASSESSING COST-EFFECTIVENESS OF TREATMENTS USING PROPENSITY SCORES IN THE PRESENCE OF MISSING DATA

Alka Indurkha*, Harvard University School of Public Health
Nandita Mitra, Memorial Sloan-Kettering Cancer Center

Health summary measures are commonly used by policy makers to help make decisions on allocation of societal resources for competing medical treatments. The net monetary benefit is a health summary measure that overcomes the statistical limitations of a popular measure namely, the cost-effectiveness ratio. Propensity score adjustment of the net monetary benefit has been recently shown to provide less biased estimates in the presence of significant differences in baseline measures and demographic characteristics between treatment groups in quasi-randomized or observational studies. We conducted Monte-carlo simulation studies to better understand the utility of propensity score adjusted estimates of net monetary benefits when (a) important covariates are unobserved, and (b) significant covariates have patterns of missingness. We conclude that the net monetary benefit is to be preferred as a summary measure over the cost-effectiveness ratio as the estimates can be easily adjusted using propensity scores in quasi randomized studies in the event of unobserved covariates, and patterns of missing data. The methods are illustrated using data from SEER-Medicare for the treatment of bladder cancer.

e-mail: aindurkh@hsph.harvard.edu

INFERENCES ON THE MEANS FOR LOGNORMAL DATA CONTAINING ZEROS BASED ON GENERALIZED VARIABLE METHOD

Lili Tian * , University of Florida

This talk presents procedures for hypothesis testing and interval estimation of a single lognormal mean containing zeros based on the concepts of generalized variables. Similar procedures are also developed for obtaining confidence intervals and tests for the ratio (or the difference) of two independent lognormal mean containing zeros. The coverage probabilities and type I error control are empirically studied.

e-mail: ltian@biostat.ufl.edu

APPLYING MARKOV MIXED-EFFECTS REGRESSION MODELS IN ANALYSIS OF RISK FACTORS FOR TRANSITIONS OF ICU PATIENTS' COGNITIVE STATUS

Ayumi K. Shintani*, Vanderbilt University
E. Wesley Ely, Vanderbilt University
Theodore Speroff, Vanderbilt University

Delirium is a form of cognitive impairment highly prevalent among patients in the Intensive Care Unit (ICU), which has recently been shown to be associated with higher mortality and longer hospital stay. Most studies evaluating risk factors for delirium use logistic regression models with development of delirium as an outcome variable. These models ignore daily fluctuations of patients' cognitive status and may mask important risk factors. In this study we propose a first-order Markov mixed effects model to describe the probability of changes in patients' cognitive status over time by implementing a SAS macro for a generalized linear mixed model, GLIMMIX (Wolfinger and O'Connell, 1993). The traditional logistic regression approach did not indicate an association between sedative/analgesic medications and delirium. Using the Markov model approach, these medications were found to be significantly associated with transitions to delirium. Such findings have implications for the design of future studies in which investigators choose modifiable risk factors to reduce delirium. This investigation points out a potential hazard in using a traditional approach in comparison with the novel Markov approach when analyzing risk factors for clinical transitions in the complex ICU patient population.

e-mail: ayumi.shintani@vanderbilt.edu

A NEW METHOD OF PREDICTING U.S. AND STATE-LEVEL CANCER MORTALITY COUNTS FOR THE CURRENT CALENDAR YEAR

Ram C Tiwari * , National Cancer Institute;
Kaushik Ghosh, George Washington University and National Cancer Institute;
Ahmedin Jemal, American Cancer Society;
Mark Hachey, Information Management Services, Inc.;
Elizabeth Ward, American Cancer Society;
Michael J. Thun, American Cancer Society;
Eric J. Feuer, National Cancer Institute

Every January for more than 40 years, the American Cancer Society (ACS) has estimated the total number of cancer deaths that are expected to occur in the United States and individual states in the upcoming year. In a collaborative effort to improve the accuracy of the predictions, investigators from the National Cancer Institute (NCI) and the ACS have developed and tested a new prediction method. The new method was used to create the mortality predictions for the first time in Cancer Statistics, 2004 and Cancer Facts & Figures, 2004. Here, the authors present a conceptual overview of the previous ACS method and the new state space method (SSM), and the results of rigorous testing to determine which method provides more accurate predictions of the observed number of cancer deaths in the years 1997-1999.

e-mail: tiwarir@mail.nih.gov

A BIAS-CORRECTED WEIGHTED AVERAGE METHOD FOR MEASUREMENTS FROM HETEROGENEOUS INSTRUMENTS WITH SYSTEMATIC BIASES

Sharon X. Xie* , University of Pennsylvania
Vernon M. Chinchilli, Pennsylvania State University

Considering a situation in which observations are made by several instruments for a group of subjects, researchers are interested in estimating the true biological quantity associated with each subject. One approach to estimate the true quantity is simply averaging the observations from those instruments for each subject. However, since the instruments may have systematic biases and the variability of each instrument may be very different, the naive average can be improved by a bias-corrected weighted average with the systematic biases estimated using a gold standard among the instruments and with the weights estimated from the measurements of all the subjects. Such a bias-corrected weighted average estimator is studied in this paper. Asymptotic properties are developed, which include that the proposed estimator can do better than the gold standard. Small sample properties are studied using computer simulations and an illustration is provided.

e-mail: sxie@cceb.upenn.edu

SIZE DOES MATTER: A CASE FOR THE .04 SIGNIFICANCE LEVEL

Adrian E. Levin^{*}, Apex College, Nepal;
Michael Atwatter, Apex College, Nepal

In biomedical studies, it is common practice to use a .05 significance level for conducting frequentist hypothesis tests, such as trend tests or group comparisons. While this choice has some advantages—reasonably high power under the alternative hypothesis, for example—it is also somewhat arbitrary and can result in high false discovery rates, especially when compared to smaller significance levels. In many instances, researchers may want to reduce the probability of rejection under the null. For this reason, we propose a new approach to hypothesis testing based on a .04 significance level. We show analytically that, while there is some loss of power under this approach, it also substantially lowers the probability of rejection under the null. We conduct a series of simulation studies to demonstrate that this result holds under various models and sample sizes. The method is illustrated using data from a recent epidemiology study.

e-mail: biosboy5@hotmail.com

64. Applications of Random Effects Models

AN EFFICIENT MCEM ALGORITHM FOR FITTING GENERALIZED LINEAR MIXED MODELS FOR CORRELATED BINARY DATA

Ming Tan, University of Maryland Greenebaum Cancer Center
Guoliang Tian^{*}, University of Maryland Greenebaum Cancer Center
Hongbin Fang, University of Maryland Greenebaum Cancer Center

Generalized linear mixed models have been used widely in the analysis of correlated binary data arisen in many research areas. Maximum likelihood fitting of these models remains to be a challenge because of the complexity of the likelihood function. Current approaches are primarily to either approximate the likelihood or use a sampling method to find the exact likelihood solution. The former results in biased estimates, and the latter uses Monte Carlo EM (MCEM) methods with a Markov chain Monte Carlo algorithm in each E-step, leading to problems of convergence and slow convergence. This paper develops a new MCEM algorithm to maximize the likelihood for generalized linear mixed probit-normal models for correlated binary data. At each E-step, we utilize a non-iterative conditional sampling approach to generate independent samples from a truncated multivariate normal distribution, thus eliminating problems of convergence and slow convergence. Two real datasets from a three-period crossover trial and a children's wheeze study in six cities are analyzed to illustrate the proposed method and to compare it with existing methods. The results show that the proposed MCEM algorithm outperformed that of McCulloch (1994) substantially.

e-mail: gtian2@umm.edu

CHALLENGES IN THE ANALYSIS OF OBSERVATIONAL LONGITUDINAL AGING STUDIES

Geert Verbeke * , Katholieke Universiteit Leuven, Belgium;
Steffen Fieuws, Katholieke Universiteit Leuven, Belgium;
Bart Spiessens, Katholieke Universiteit Leuven, Belgium;
Emmanuel Lesaffre, Katholieke Universiteit Leuven, Belgium

Observational longitudinal data are often highly unbalanced in the sense that not a fixed number of measurements is available for all subjects, and measurements are taken at arbitrary time points. The analysis requires flexible models, such as linear mixed models in case of continuous data. We will show that, for observational data, many aspects of the model formulation deserve special attention. First, the actual parameterization of the model requires special attention. Indeed linear mixed models imply specific mean and covariance structures, which depending on the parameterization chosen, may or may not be realistic and/or meaningful for the data set at hand. Second, it is not always clear to what extent corrections are needed for baseline differences between study participant, nor is it often clear how such corrections will affect inferences of interest. Finally, the concepts of missing data and dropout are much less clearly defined than for balanced designed experiments. These issues will be discussed in the context of the Baltimore Longitudinal Study of Aging, an ongoing multidisciplinary observational study, with the study of normal human aging as primary objective. Keywords: Linear mixed model, Longitudinal data, observational data, Baseline characteristic, Parameterization, Unbalanced data

e-mail: geert.verbeke@med.kuleuven.ac.be

A MIXED-EFFECTS REGRESSION MODEL FOR LONGITUDINAL MULTIVARIATE ORDINAL DATA

Li C. Liu*, University of Illinois at Chicago
Donald Hedeker, University of Illinois at Chicago

A mixed-effects item response theory (IRT) model which allows for three-level multivariate ordinal outcomes and accommodates multiple random subject effects is proposed for analysis of multivariate ordinal outcomes in longitudinal studies. This model allows for the estimation of different item factor loadings (item discrimination parameters) for the multiple outcomes and the covariates can be at any level. Assuming either a probit or logistic response function, maximum marginal likelihood estimation (MMLE) is proposed utilizing multidimensional Gauss-Hermite quadrature for integration of the random effects. An iterative Fisher-scoring solution, which provides standard errors for all model parameters and generally converges faster than the EM algorithm, is used. Simulation results are presented to show the performance of the model. An analysis of a longitudinal substance use data set, where four ordinal items of substance use behavior (cigarette, alcohol, marijuana use, and getting drunk or high) are repeatedly measured over time, is used to illustrate application of the proposed model.

e-mail: lqi1@uic.edu

A MULTIVARIATE RANDOM EFFECTS MODEL WITH CLOSED-FORM LIKELIHOOD FOR CLUSTERED BINARY DATA

Brent A. Coull *, Harvard School of Public Health;
Rebecca A. Betensky, Harvard School of Public Health

We consider a multivariate random effects model for clustered binary data that is useful when interest focuses on the association structure among clustered observations. Based on a vector of gamma random effects and a complementary log-log link function, the model yields a likelihood having closed-form, making model fitting straightforward. Because the model completely specifies the joint distribution of responses for a cluster, it accommodates non-randomly sampled data, such as case-control data often encountered in familial aggregation studies. We use the methods to analyze data from a study investigating the familial aggregation of breast and ovarian cancer.

e-mail: bcoull@hsph.harvard.edu

GLMMS WITH CORRELATED RANDOM EFFECTS TO MODEL DISCRETE TIME SERIES

Bernhard Klingenberg*, University of Florida
Alan Agresti, University of Florida
James Booth, University of Florida

Univariate random effects in generalized linear models induce an exchangeable correlation structure. In contrast, time series or spatial measurements often show correlations decaying with increasing distance between measurements. We propose correlated random effects to model these inherent properties, paying special attention to unequally spaced discrete time series. The challenge for discrete data lies in the fact that the observed likelihood is not analytically tractable, and maximization of it is more involved. Most approaches in the literature adopt a Bayesian view or estimate parameters via generalized estimating equations. We propose maximum likelihood estimation via a Monte Carlo EM algorithm where both E- and M-step require numerical approximation. The approach is illustrated with the analysis of a cross sectional study over time and a long binary time series.

e-mail: bklingen@stat.ufl.edu

AN ADAPTIVE METHOD FOR CHECKING CONVERGENCE OF THE MONTE CARLO EM ALGORITHM

Jens C. Eickhoff^{*}, University of Wisconsin;
Jun Zhu, University of Wisconsin; Yasuo Amemiya, IBM
Thomas J. Watson Research Center

The EM algorithm has been a popular tool for obtaining the maximum likelihood estimates for generalized linear mixed effect or latent variable models. When the conditional expectation of a complete-data likelihood in an EM algorithm is analytically intractable, Monte Carlo integration is typically used to approximate the E-step. While the resulting Monte Carlo EM algorithm (MCEM) is flexible, assessing convergence of the algorithm is a more difficult task than in the original EM algorithm, because of the uncertainty involved in the Monte Carlo approximation. In this talk, we propose a convergence criterion using a likelihood-based distance. In our approach, the Monte Carlo sample size is selected adaptively at each step of the MCEM algorithm. We implement the proposed convergence criterion along with the simulation size selection in a one-way random effects model. The result shows that our MCEM iterations match the exact EM iterations closely over iterations.

e-mail: eickhoff@biostat.wisc.edu

MODELING AND ESTIMATION IN TWO-STAGE MODELS WITH APPLICATIONS

John P. Buonaccorsi^{*}, University of Massachusetts at Amherst

Many experiments proceed in two-stages in which at the first stage a random sample of “units” is chosen and then some quantity associated with the selected unit is estimated via second stage ‘sampling’. Conditionally, heteroscedasticity in the estimates over different units can arise for a number of reasons, including ‘inherent variation’ associated with the units and changing sampling effort. We address the problem of estimating the population mean and variance associated with the unobserved true values associated with the first stage of sampling, in a manner which is robust to the nature of the heteroscedasticity. It is shown that under very weak assumptions an unbiased estimate of the mean and the variance of the estimated mean is obtained by running a standard analysis on the estimates as if they were true unobserved values from the first stage. Further, this is the optimal thing to do under a variety of mechanisms that might create heteroscedasticity. An unbiased estimate of the population variance is obtained as is a robust estimate of the variance-covariance matrix of the estimated mean and variance. Applications include one-way random effects models with replication, meta-analysis models, random coefficients model and two (or multi)-stage sampling from finite populations.

e-mail: johnpb@math.umass.edu

SAMPLE SIZE CALCULATION AND CONCORDANCE ASSESSMENT FOR AN AGREEMENT STUDY

Jason Liao * , Merck Research Laboratories

It is often necessary to compare two measurement methods in medicine and other experimental sciences. This problem covers a broad range of data. Many authors have explored ways of assessing the agreement of two sets of measurements in terms of an index. However, there has been little attention to the problem of determining sample size for designing an agreement study. In this paper, a method is proposed to calculate sample size and assess concordance in conducting an agreement study. The sample size calculation and concordance evaluation are based on two rates: the discordance rate and tolerance probability. These two rates play similar roles as the significant level and power in the traditional sense. This method is especially appealing if there is no information about the variability of each individual measurement available before the experiment or if the agreement assessment is over a range instead of a fixed value, which is often the case in practice. In addition, a method is proposed to define the agreement interval for each individual paired observation in assessing the overall concordance. A real data set is used to demonstrate the sample size calculation and concordance assessment.

e-mail: jason_liao@merck.com

STATISTICAL ISSUES IN CALIBRATION OF SERUM CREATININE TESTING IN THE CHRONIC RENAL INSUFFICIENCY COHORT STUDY

Christina Gaughan*, University of Pennsylvania
 Marshall M. Joffe, University of Pennsylvania
 J. Richard Landis, University of Pennsylvania
 Harold I. Feldman, University of Pennsylvania

The Chronic Renal Insufficiency Cohort (CRIC) Study is a multicenter study. Although most laboratory testing is to be performed at a single central laboratory, testing serum creatinine to determine eligibility for the study must be performed at each of the centers. To ensure a uniform entry criteria, the serum creatinine obtained at each site must be calibrated to a reference laboratory. The serum creatinine measures used in the calibration study were performed on replicates of the same samples. We consider here three issues: 1) determining what is the correct calibration equation to use when the regression equations differ with the range of serum creatinines in the sample; 2) determining whether calibration equations from different samples reflect the same underlying relationships; and 3) determining the correct calibration equation when, by design, the dependent variable is the same for several independent predictors.

e-mail: mjoffe@cceb.upenn.edu

A THRESHOLD DOSE-RESPONSE MODEL WITH RANDOM EFFECTS IN TERATOLOGICAL EXPERIMENTS

Daniel Hunt, St. Jude Children's Research Hospital;
Shesh N. Rai *, St. Jude Children's Research Hospital

Teratological experiments are controlled dose-response studies in which impregnated animals are randomly assigned to various exposure levels of a toxic substance. Subsequently, both continuous and discrete responses are recorded on the litters of fetuses that these animals produce. This clustered binary data usually exhibits over-dispersion. To model the correlation and/or variation, the beta-binomial distribution has been assumed for the number of positive fetal responses within a litter. Although the mean of the beta-binomial model has been linked to dose-response functions, in terms of measuring over-dispersion, it may be a restrictive method in modeling data from teratological studies. Also for certain toxins, a threshold effect has been observed in the dose-response pattern of the data. We propose to incorporate a random effect into a general threshold dose-response model to account for the variation in responses, while at the same time estimating the threshold effect. We fit this model to a well-known data set in the field of teratology. Simulation studies are performed to assess the validity of the random effects threshold model in these types of studies.

e-mail: Shesh.Rai@stjude.org

SOME STATE SPACE MODELS OF CARCINOGENESIS AND ESTIMATION OF PARAMETERS

Wai-Yuan Tan, University of Memphis
Lijun Zhang*, University of Memphis

In this paper we have developed a state space model for carcinogenesis. By using this state space model we have also developed statistical procedures to estimate the unknown parameters via multi-level Gibbs sampling method. We have applied this model and the methods to the British physician data on lung cancer with smoking. Our results indicate that the tobacco nicotine is an initiator. If $t > 60$ years old, then the tobacco nicotine is also a promoter.

e-mail: lzhang1@memphis.edu

SUMMARIZING FLARE ASSAY IMAGES IN COLON CARCINOGENESIS

Malgorzata Leyk Williams * , Texas A&M University;
Raymond J. Carroll, Texas A&M University

Measurement of the amount of oxidative damage to DNA is one tool that can be used to estimate the beneficial effect of diet on the prevention of colon carcinogenesis. The FLARE assay is a modification of the single-cell gel electrophoresis (Comet) assay, and provides a measure of the 8OHdG adduct in the cells. Here, each data point is a function describing the amount of damage to the cell's DNA. We describe ways of analyzing this data, including simple summary statistics as well as fitting functions to it, to describe differences between the amount of damage in cell's DNA as a results of exposure to corn or fish oil diets.

e-mail: gosia@tamu.edu

CORRELATION COEFFICIENT INFERENCE ON CENSORED BIOASSAY DATA

Liang Li*, Cleveland Clinic Foundation
William Wang, Merck Research Laboratories
Ivan Chan, Merck Research Laboratories

In vaccine clinical trials, immunologic responses sometimes can not be accurately measured by bioassays. For example, a serial dilution assay usually reports the range of the response instead of the exact value. In some other assays, the measurement is not available if the response is lower than the assay's detection limit. In both cases, the measurements are censored. We are interested in computing the confidence interval for the correlation coefficient of two assay measurements that are subject to censoring. We propose using the maximum likelihood method to estimate the correlation coefficient, and constructing its confidence interval based on the second-order Taylor's expansion of the Fisher's Z transformation. The method can be viewed as an extension of the Fisher's Z transformation to the case of censored data. Extensive simulations show that the proposed method is computationally simple and can provide satisfactory coverage probabilities even with small sample sizes. The proposed method works with many types of censored data in a similar way.

e-mail: lli@bio.ri.ccf.org

A STOCHASTIC MODEL TO ANALYZE CLONAL DATA ON MULTI-TYPE CELL POPULATIONS

Ollivier Hyrien * , University of Rochester Medical Center

A stochastic model designed to analyze experimental data on the development of cell clones composed of two (or more) distinct types of cells is presented. The proposed model is an extension of the traditional multi-type Bellman-Harris branching stochastic process allowing for non-identical time-to-transformation distributions defined for different cell types. A simulation-based procedure has been developed for parametric inference from experimental data on cell clones under the proposed model. The model and associated methods of parametric inference have been applied to the analysis of proliferation and differentiation of cultured O-2A progenitor cells that play a key role in the development of the central nervous system. It follows from this analysis that the time to division of the progenitor cell and the time to its differentiation (into an oligodendrocyte) are not identically distributed, with the difference being more pronounced between the variances than between the mean values.

e-mail: ollivier_hyrien@urmc.rochester.edu

66. Recent Advances in Tumor Growth Modeling

LONGITUDINAL MODELS WITH CONSTRAINED PARAMETERS FOR TUMOR XENOGRAFT EXPERIMENTS

Mnig T. Tan* , University of Maryland
Hongbin Fang, University of Maryland
Guo Tian, University of Maryland

In cancer drug development, demonstrated activity in xenograft models, where mice are grafted with human cancer cells, is an important step to bring a promising compound to human. A key outcome variable in these experiments is tumor volumes measured serially over a period of time, while mice are treated with an anti-cancer agent following certain schedules. However, multiple tumors are grafted in a mouse and some of them may fail to grow the tumor and a mouse may die during the follow-up or may be sacrificed when its tumor volume quadruples. Incomplete observations also arise if the tumor shrinks to an undetectable level at certain time points during the follow-up. In addition, if no treatment were given to the tumor-bearing mice, the tumors would keep growing until the mice die or are sacrificed. This intrinsic growth of tumor in the absence of treatment constrains the parameters should also be accounted for in the statistical modeling and causes further difficulties for analysis. We develop a multivariate mixed-effects model with constrained parameters to model the complexities of the dose-response relationship. The method is illustrated with a xenograft study evaluating the anti-tumor effects of exemestane, tamoxifen and letrozole for postmenopausal breast cancer.

e-mail: mtan@umm.edu

MODELING TUMOR GROWTH WITH RANDOM ONSET

Paul. S Albert * , National Cancer Institute;
Joanna H. Shih, National Cancer Institute

The longitudinal assessment of tumor volume is commonly used as an endpoint in small animal studies in cancer research. Groups of genetically identical mice are injected with mutant cells from clones developed with different mutations. The interest is on comparing tumor onset (i.e., the time of tumor detection) and tumor growth after onset, between mutation groups. This article proposes a class of linear and nonlinear growth models for jointly modeling tumor onset and growth in this situation. Our approach allows for interval censored time of onset and missing at random dropout due to early sacrifice, which are common situations in animal research. We show that our approach has good small-sample properties for testing and is robust to some unverifiable modeling assumptions. We illustrate this methodology with an application examining the effect of different mutations on tumorigenesis.

e-mail: albertp@ctep.nci.nih.gov

QUANTAL RESPONSE MODELS OF CANCER DETECTION IN THE PRESENCE OF SCREENING

Andrei Y. Yakovlev*, University of Rochester

The natural history of cancer includes the initiation, promotion and progression stages of tumor development, with the structure of the latter stage being dependent on a specific mechanism of cancer detection. This paper discusses methods for estimating unobservable characteristics of tumor latency from multivariate data on those quantities that are observed at the time of tumor detection. More specifically, we consider the utility of a mechanistic model of breast cancer screening in the analysis of bivariate data on age and tumor size at detection. The model assumes that the rate of detection is proportional to tumor size. The model is proven to be identifiable. Biologically meaningful parameters of the model are estimated by the method of maximum likelihood from the data on age and tumor size at detection resulted from two randomized trials. When properly calibrated, the model provides a good description of the U.S. national trends in breast cancer incidence and mortality. In particular, the model provides an excellent prediction of the stage-specific age-adjusted incidence of invasive breast cancer as a function of calendar time for the period 1975-1999. Predictive properties of the model are also illustrated with an application to the dynamics of age-specific incidence over the period 1975-1999 and size-specific age-adjusted incidence over the period 1988-1999.

e-mail: andrei_yakovlev@urmc.rochester.edu

NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION FOR CLUSTERED/LONGITUDINAL DATA USING KERNEL AND SPLINE METHODS

Xihong Lin*, University of Michigan
Naisyin Wang, Texas A&M University
Alan Welsh, Southampton University
Raymond J. Carroll, Texas A&M University

We consider nonparametric and semiparametric regression estimation for clustered/longitudinal data using kernel and (smoothing and regression) spline methods. A key feature of the smoothing spline method is that it can be fit using mixed models. However, its properties have not been well understood. Unlike independent data, it is found recently that the same is not true for clustered/longitudinal data. conventional kernel methods fail to account for the within-cluster correlation and are local, while spline methods are able to account for this correlation and are nonlocal. We show that a smoothing spline estimator is asymptotically equivalent to a newly proposed seemingly unrelated (SUR) kernel estimator. The most efficient spline and SUR kernel estimators are obtained by accounting for the within-cluster correlation and are nonlocal, but have asymptotically negligible bias. Our results justify the use of efficient, non-local estimators such as smoothing clustered/longitudinal data. We extend the results to semiparametric regression models, where some covariate effects are modeled parametrically, while others are modeled nonparametrically. We derive the semiparametric efficient score and show the profile/kernel or spline estimator is semiparametric efficient. The results are illustrated using simulation studies and data examples.

e-mail: xlin@umich.edu

NONPARAMETRIC MIXED-EFFECT MODELS

Chong Gu * , Purdue University

Mixed-effect models are widely used for the analysis of correlated data such as longitudinal data and clustered observations. In this talk, I will present some recent results on the nonparametric estimation of the fixed effects in such models. Using Henderson's likelihood, the 'variance components' can be turned into 'mean components,' and computation and cross-validation strategies developed for independent data can be used to handle correlated data. The optimality of cross-validation in the setting is discussed and simulations are conducted to study the empirical performance. Real-data examples are shown to illustrate the applications of the methodology; open-source R code is demonstrated.

chong@stat.purdue.edu

BAYESIAN INFERENCE FOR WAVELET-BASED MODELING OF FUNCTIONAL DATA

Marina Vannucci * , Texas A&M University

In this talk I will describe Bayesian methodologies for wavelet-based modeling of functional data extended beyond the single curve case. I will first consider the choice of explanatory variables in multivariate linear regression models with predictors arising as curves. I will use wavelet transforms to represent the curves through wavelet coefficient sets describing local features in a parsimonious way. I will then use mixing priors and MCMC methods to select wavelet coefficients that best predict a multivariate response. I will show applications to infrared spectroscopy and also discuss extensions of the methodologies to classification models. Next, I will address the problem of mean function estimation for functional data that have a nested hierarchical structure. I will use experimental data arising from carcinogen-induced colon cancer in rodent models. Here multiple shrinkage priors on the wavelet coefficients at the top of the hierarchy will result in adaptive regularization of the function estimates. The method will lead to estimates and posterior credible intervals for the mean function and random effects functions, as well as the variance components of the model.

e-mail: mvannucci@stat.tamu.edu

68. Bioinformatics in Genomics Studies

CLINICAL GENOMICS - PRINCIPLES AND APPLICATIONS

Chuanbo Xu * , Genaissance Pharmaceuticals, Inc.;
 Krishnan Nandabalan, Genaissance Pharmaceuticals, Inc.;
 Madan Kumar, Genaissance Pharmaceuticals, Inc.;
 Ruhong Jiang, Genaissance Pharmaceuticals, Inc.;
 Richard Judson, Genaissance Pharmaceuticals, Inc.;
 Brad Dain, Genaissance Pharmaceuticals, Inc.; Carol
 Reed, Genaissance Pharmaceuticals, Inc

Genomics and genetics are making ways into the mainstream practices of the drug discovery and development and patient care in the clinics. We have built the technology platform leveraging the gene variation information of the human genome and we are applying the pharmacogenomics and pharmacogenetics knowledge to aid in the clinical trial design, to develop genetic diagnostics for multiple diseases and to identify genetic markers of drug responses in disease areas, including the cardiovascular and central nervous systems. This presentation will summarize our understanding and the progress made with our marker discovery and validation processes.

e-mail: c.xu@genaissance.com

AN INNOVATIVE APPROACH THAT SYNTHESIZES EXPRESSION PROFILES, GENOTYPES AND PHENOTYPES

Stephanie A. Monks*, University of Washington, Seattle and Rosetta Inpharmatics LLC
Haoyuan Zhu, Rosetta Inpharmatics LLC
Amy Leonardson, Rosetta Inpharmatics LLC
Eric Schadt, Rosetta Inpharmatics LLC

The number of disease genes that have been mapped in the human genome has grown exponentially over that past two decades. Nevertheless, gene mapping for diseases and their associated risk traits that are of public health interest has yielded few successes. With the advent of technology to measure changes in gene expression, it should be possible to unravel some of the complexity existing for these common diseases by incorporating genetic variation, patterns of genetic inheritance and gene expression components. Here, we provide two strategies for combining such data. The first analyzes the genetics of each gene expression trait individually and relies on clustering of loci that influence the phenotypic variation of multiple gene expression traits. This is contrasted with multi- dimension reduction techniques that create a smaller set of uncorrelated expression-based traits. These new traits, or “super-genes,” are computed to explain the maximal amount of variation in the expression data under different constraints. The super-genes are then analyzed to determine the loci influencing their phenotypic variation. These strategies will be compared in order to determine what type of data set will be most appropriate for each technique.

e-mail: monks@u.washington.edu

POOLING INFORMATION ACROSS DIFFERENT STUDIES AND AFFYMETRIX CHIP TYPES TO IDENTIFY PROGNOSTIC GENES FOR LUNG CANCER

Jeffrey S. Morris*, M.D. Anderson Cancer Center
Li Zhang, M.D. Anderson Cancer Center
Guosheng Yin, M.D. Anderson Cancer Center
Keith Baggerly, M.D. Anderson Cancer Center
Chunlei Wu, M.D. Anderson Cancer Center

Our goal in this work is to pool information across studies conducted at different institutions using two different versions of Affymetrix chips to identify genes offering information on lung cancer patients; survival above and beyond the information provided by readily available clinical covariates. We combine information across chip types by identifying matching probes present on both chips, then combining them into new probesets using Unigene clusters. This method yields comparable expression level quantifications across chips without sacrificing much precision or significantly altering the relative ordering of the samples. We fit a series of multivariable Cox models containing clinical covariates and genes and identified 26 genes that provide information on survival after adjusting for the clinical covariates, while controlling the false discovery rate at 0.20 using the Beta-Uniform mixture method. Many of these genes appear to be biologically interesting, and worthy of future investigation. Only one gene in our list has been mentioned in previously published analysis of these data. It appears that the increased statistical power provided by the pooling was key in finding these new genes, since only 9 out of the 26 genes were detected when applying our methods to the two data sets separately, i.e. without pooling.

e-mail: jeffmo@odin.mdacc.tmc.edu

OPTIMAL SURVEILLANCE OF HEALTH EVENTS

Marianne E. Frisen*, Statistical Research Unit Göteborg University

Statistical surveillance has the goal of detecting an important change in the underlying process as soon as possible after it has occurred. Optimality of surveillance will be discussed in the context of sequences of health events. Timeliness and the control of false alarms are important issues. An overview of methods and optimality criteria will be given.

e-mail: marianne.frisen@statistics.gu.se

SIMULATING ANTHRAX ATTACKS AND EVALUATING DETECTION TECHNIQUES

Ken Kleinman*, Harvard Medical School;
 Allyson Abrams, Harvard Medical School;
 Richard Platt, Harvard Medical School;
 Martin Kulldorff, Harvard Medical School

We describe simulating anthrax attacks in eastern Massachusetts and show results of applying various detection methods to the simulations. Anthrax attacks are simulated based on random release points or lines, with a variety of shapes of spore distribution and number and aggressiveness of spores. Strata include shapes (based on atmospheric conditions) day of week and time point in the year, and density of population. We also simulate the time from distribution to first symptoms among those affected. After simulating attacks, we evaluate the test characteristics of several competing statistical surveillance tools that are currently being applied in the real data to which the simulated events are added. These include simple and time-weighted time-series models, small area regression tools, and unadjusted and adjusted scan statistics. We conclude that the best single tool in such conditions is the adjusted scan statistic, but that a smorgasbord of tools provides superior sensitivity.

e-mail: ken_kleinman@harvardpilgrim.org

SURVEILLANCE OF SPATIAL MAXIMA

Peter A. Rogerson*, University at Buffalo

Often there is interest in the statistical significance of the peak value on a possibly smoothed map of regional disease rates. In addition, there may be a series of such maps observed over time, and interest then lies in detecting quickly any increase in the distribution of the maximum value. In this paper, a cumulative scheme is derived for the quick detection of changes in the Type I Gumbel distribution, which is often used to model maximum values. The approach is optimal in the sense that it minimizes the time taken to detect a change of prespecified size in the distribution's parameter. The critical threshold for the cumulative sum, above which an alarm is signaled, may be determined by either Monte Carlo simulation, or an approximation based upon the previous work of Siegmund. A related problem is to monitor the region in which the maximum value occurs. This is associated with recent developments in nonparametric cusum method; this approach will also be described and illustrated. Following the development of the approach, the methods are illustrated using both simulated and actual data.

e-mail: rogerson@buffalo.edu

MODELING SPATIALLY CORRELATED SURVIVAL DATA WITH MULTIPLE CANCERS

Sudipto Banerjee * , University of Minnesota

The last decade has witnessed major advances in medical science and health care that have prolonged survival of patients suffering from potentially fatal diseases. Consequently, medical databases today offer much more information than were available to researchers some years ago. Survival data are available on very sick patients suffering from multiple cancers, who are still undergoing treatment to fight disease progression. Analysis of such data help in assessing treatment and therapeutic strategies for dealing with diseases at advanced stages, for assessing possible associations between the diseases and detecting their progression patterns in the patients. Also available in most databases today are geographical information on patients. Used conjunctively with statistical software, sophisticated programs known as GIS (Geographical Information Systems) enable easy production of maps for raw data as well as estimates of other interesting quantities. Through proper use, such maps facilitate discerning spatial patterns in diseases. We address the development of a modelling framework that will enable the analysis of such complex databases as above to answer interesting questions pertaining to the health sciences. Our methods focus upon some underdeveloped aspects of spatial and survival data analysis tools, seeking to account for the multivariate nature of the data.

e-mail: sudiptob@biostat.umn.edu

THE ANALYSIS OF SPATIALLY-REFERENCED RECURRENT EVENT HEALTH DATA WITH COMORBIDITIES

Andrew B. Lawson * , University of South Carolina;
Shae Sutton, University of South Carolina;
Charity Moore, University of South Carolina

The analysis of geo-referenced health data is now well developed. Less well developed is the analysis of survival and longitudinal data which is geo-referenced. In this talk we will focus on the modeling of recurrent event data based on the observations of repeated visits to health providers. The dataset used is a Medicaid data set of asthma, congestive heart failure and diabetes diagnosed patients and their progression through disease states. The data is from rural South Carolina. We examine the use of a Weibull recurrence model within a Bayesian framework and examine its extension to multiple type stages and comorbidities.

e-mail: alawson@gwm.sc.edu

70. Analysis of Intensively Collected Data

SCIENTIFIC ISSUES DRIVING MODEL DEVELOPMENT FOR INTENSIVE LONGITUDINAL DATA

Theodore A. Walls* , Penn State University

Data collected using real-time approaches such as handheld computers, satellites, cell phones, web interfaces, and other technological devices are changing the face of scientific inquiry in the health and social SCIENCES. These data are appropriately termed ‘intensive longitudinal data’ (ILD) because there are typically many occasions of measurement, many subjects and many variables. In the last decade, numerous studies using handheld computers paired with extensive and sophisticated self-report protocols have produced intensive longitudinal data. This presentation provides an overview of several of these studies, summarizes the scientific questions that tend to arise, outlines extensive statistical challenges inherent to ILD, and enumerates several promising modeling approaches. These approaches include functional data analysis, dynamical systems models, autoregressive modeling, state space and structural equation models, and point process models. These approaches are to be treated in Walls, T.A. & Schafer, J.S. (forthcoming in 2005) *Models for Intensive Longitudinal Data*. New York: Oxford University Press.

e-mail: taw12@psu.edu

BACKFITTING FOR RANDOM VARYING-COEFFICIENT MODELS

Hua Liang*, St. Jude Children's Research Hospital
Wu Hulin, University of Rochester

In this paper we propose a random varying-coefficient model for longitudinal data. This model is different from the standard varying-coefficient model in the sense that the time-varying coefficients are assumed to be subject-specific, and can be considered as realizations of stochastic processes. This modeling strategy allows us to employ powerful mixed-effects modeling techniques to efficiently incorporate the within-subject and between-subject variations in the estimators of time-varying coefficients. Thus, the subject-specific feature of longitudinal data is effectively considered in the proposed model. A backfitting algorithm is proposed to estimate the coefficient functions. Simulation studies show that the proposed estimation methods are more efficient in finite-sample performance compared to the standard local least squares method. An application to an AIDS clinical study is presented to illustrate the proposed methodologies.

e-mail: hua.liang@stjude.org

SEMI-NONPARAMETRIC MODELS AND INFERENCE FUNCTIONS FOR HIGH DIMENSIONAL DATA

Annie Qu *, Oregon State University

A nonparametric model applying spline or kernel methods which allows for richer and more flexible model structures is appealing for prediction, but it is also highly computationally intensive and introduces high dimensional nuisance parameters. This makes it difficult to incorporate dependence structure into the model. In addition, there are limited methods for deriving explicit inference functions for testing goodness-of-fit for time varying coefficients models in the current statistical literature. We are interested in problems arising from cell cycle microarray data where gene expressions are measured over time. There are two kinds of correlations in cell cycle data. Measurements are certainly correlated within a gene where it is measured over cycles, and measurements could also be correlated between genes since some genes are biologically related and regulate the same phenotypical characteristics. Notice that there are many gene-specific parameters involved in the model, but often there are no replications of measurements. We are interested in developing a semi-nonparametric model which reduces the dimension of the nuisance parameters and still addresses the correlated nature of the measurements.

e-mail: qu@stat.orst.edu

MODELS AND METHODS FOR ANALYSIS OF INTENSIVELY COLLECTED DATA

Runze Li *, Penn State University; Tammy L. Root, Penn State University

With modern data collection devices, such as hand-held computers, and huge data storage space, researchers may easily and intensively collect data. Analysis of intensively measured data poses great challenges for biostatisticians, statisticians and researchers in other fields. This talk will present some statistical issues closely related to analysis of intensively collected data.

e-mail: rli@stat.psu.edu

71. Methods for Missing or Unbalanced Data

EFFECTS OF TRUNCATION ON EVALUATION OF RISK FACTORS, WITH CASE-STUDY IN ANTHRAX

Sarah JE Barry*, Johns Hopkins School of Public Health
Ron Brookmeyer, Johns Hopkins School of Public Health

Right truncation occurs in retrospective studies when only deaths are observed and those with the disease who did not die within the study period are missing from the data, resulting in people with longer times to death being missed. Here we consider truncation of incubation periods and the effects on potentially predictive risk factors. Our motivation is work on anthrax; of public health concern because of its potential use as a biological weapon. In this talk we consider the 1979 anthrax outbreak in Sverdlovsk, Russia, the largest documented in modern times, in which 70 people died. In this outbreak, antibiotics were distributed approximately 15 days after exposure. This may have caused a selection of people with incubation periods of longer than 15 days to be unobserved because they received antibiotics that prevented their deaths, thus biasing the data towards short incubation periods. We performed simulations and sensitivity analyses to evaluate the effects of ignoring truncation on risk factors such as age, gender and directional distance from source. We found that if the truncation probability did not depend on a covariate there was attenuation of effects, with severe attenuation only when this probability was high. If truncation probabilities varied by the covariate of interest (differential truncation), there was modification but no reversal of covariate effects.

e-mail: sbarry@jhsph.edu

AN EMPIRICAL COMPARISON OF IMPUTATION METHODS FOR USE WITH PUBLIC HEALTH DATA

Stephanie T. Broyles * , Tulane University

Multiple imputation is a popular method to address missing data and is currently supported in standard statistical software. This method, however, is not a practical way to deal with missingness in open data collection systems, such as disease surveillance systems, because multiple imputation requires maintaining and analyzing multiple datasets. In addition, disease surveillance data are most often used in resource and program planning, which require simple data summaries rather than hypothesis testing (i.e., proportions rather than tests of effects). This research will compare several alternative methods to multiple imputation designed for use in an open dataset with categorical variables. The accuracy of the different methods in predicting proportions will be considered, as well as each method's sensitivity to assumptions about the mechanism of missingness. Comparisons will be made to both complete-case analysis, which is commonly used with open datasets, and to multiple imputation.

e-mail: sbroyl@lsuhsc.edu

A SIMPLE SENSITIVITY ANALYSIS TOOL FOR NONIGNORABLE COARSENING: APPLICATION TO COMPETING RISK

Jiameng Zhang* , University of Pennsylvania
Daniel F. Heitjan, University of Pennsylvania

Missing, censoring and grouping are common types of coarsening. Standard statistical method ignoring the coarsening mechanism might lead to incorrect inferences. We extend a simple sensitivity analysis tool, the index of sensitivity to nonignorability (Troxel et al. 2003), to the evaluation of the nonignorability of the coarsening process. Using this index and available information from the coarse data, one can easily assess the sensitivity of each key inference to the nonignorable coarsening. We apply this sensitivity analysis method to the dependent competing risk problem. A simulation study shows that the method is appropriate for practical use. We apply the method to evaluate sensitivity to nonignorable censoring in the Stanford Heart Transplant Data.

e-mail: jzhang@cceb.upenn.edu

COMPARING THE ANALYSES OF UNBALANCED SPLIT-PLOT EXPERIMENTS USING MIXED MODELS WITH REML ESTIMATES OF THE VARIANCE COMPONENTS TO OTHER INFERENCE PROCEDURES PREVIOUSLY RECOMMENDED

Christina Smith * , Kansas State University;
Dallas Johnson, Kansas State University

Several procedures for constructing confidence intervals and testing hypotheses about fixed effects in unbalanced split-plot experiments have previously been presented and discussed by Remmenga and Johnson. A few of the procedures considered have been recommended as useful and reliable procedures. Since the advent of the SAS® MIXED procedure, mixed model analyses with REML estimates of the variance components is easily accessible to researchers. This report will compare the analysis of unbalanced split-plot experiments using mixed model procedures with REML estimates of the variance components to the previously established procedures by means of additional simulation studies.

e-mail: cdsmith@ksu.edu

ANALYSIS OF LONGITUDINAL DATA WITH DROPOUTS AND INTERMITTENTLY MISSING VALUES

Gong Tang*, University of Pittsburgh

In many longitudinal clinical trials, the repeatedly measured outcome is subject to missingness which consists of two general patterns: dropout and intermittent missingness. Recent practice promotes collecting the reasons that result in missing values. In many trials the dropouts are strongly related to the outcome and the intermittent missingness are due to reasons, for example, missing a scheduled visit because of previous performance or missing data form, which are suspected to be unrelated to the underlying missing values. A general framework that incorporates the dropout process and the random intermittently missing-data process is proposed. It is shown that when the parameters for these processes and the complete-data model are distinct, the intermittently missing-data process can be ignored in likelihood inference. A robust procedure based on a pseudo-likelihood approach is also proposed and compared to the maximum likelihood.

e-mail: got1@pitt.edu

ANALYSIS OF INCOMPLETE HRQOL DATA IN THE REMATCH TRIAL

Huiling Li *, Columbia University; Daniel F. Heitjan, University of Pennsylvania

A pattern-mixture model is applied for modeling the joint distribution of incomplete repeated measurements of quality of life and right censored survival times in the evaluation of how a treatment affects an individual's well-being over time. The statistical model assumes that the survival times follow a multinomial distribution and that quality of life outcomes follow a multivariate normal distribution conditional on the survival time. The model is estimated using a Bayesian approach by importance sampling, and then used to create multiple imputation of the missing outcomes. The methods are illustrated through the analysis of the data from a randomized clinical trial of a treatment for cardiovascular disease.

e-mail: HL2034@columbia.edu

LARGE SAMPLE BIAS IN THE LAST OBSERVATION CARRIED FORWARD APPROACH UNDER INFORMATIVE DROPOUT

Chandan Saha *, Indiana University
Michael P. Jones, University of Iowa

In clinical trials, subjects are usually followed over a fixed period of time to analyze their response to the treatments randomly assigned to them. Missing data are an inescapable problem in such clinical trials. When the main interest is the outcome at endpoint of the study, the last observation carried forward (LOCF) is the most frequently used approach for dealing with missing values in clinical trials with continuous variables. However, there are several criticisms of using this approach because of possible bias in estimating the treatment effects. Some researchers emphasized that it would be interesting to have a theoretical quantification of the magnitude of bias caused by dropouts inherent to various analysis. This paper establishes a theoretical results on quantifying the large sample bias in LOCF approach under informative dropout. We consider a longitudinal study with two groups: treatment and control, and model the dropout mechanism as a function of subject specific random intercept, slope and group membership. Several case studies are included to show the magnitude of bias in the estimators for the treatment effects and variance.

e-mail: cksaha@iupui.edu

ANALYSIS OF ALLELIC LOSS DATA USING MARGINAL MODELS

Yan Li *, Ohio State University; Lei Shen, Ohio State University

Allelic loss (or loss of heterozygosity) data have been widely used to identify putative tumor-suppressor genes that play an important role in the process of tumorigenesis. Such data are binary indicators at multiple marker positions for a number of subjects. Some of the special features of allelic loss data are high rates of missing (noninformative) values and positive correlations between markers on the same chromosome. Newton et al. (1998) and Newton and Lee (2000) developed a parametric instability-selection model that captures the essential features of the data. We use data from a study employing a genome-wide scan to study the plausibility of the assumptions in their model, and propose a method to analyze allelic loss data using marginal models and generalized estimating equations for correlated data. This model requires fewer assumptions than the Newton model, and allows covariate information, such as tumor grade, to be readily incorporated. In addition, we compare the theoretical properties and practical performance of our method to those of a non-parametric model proposed by Miller et al. (2003) that is useful in exploring the data without assuming a specific model.

e-mail: li.560@osu.edu

INFERENCE ON THE LOCATION OF A TUMOR SUPPRESSOR GENE BY MODELING FREQUENCY OF ALLELIC LOSS

Andrew Sterrett*, University of North Carolina at Chapel Hill

Fred A. Wright, University of North Carolina at Chapel Hill

Allelic loss is often part of the pathway leading to tumorigenesis. Analysis of markers along the genome highlights regions of elevated occurrence of allelic loss, which in turn suggests the presence of an important growth-suppressing gene in the vicinity. Often the pattern of loss for each tumor, or allelotype, is reported, and a nonhomogeneous Markov chain model can be used to analyze correlated binary data of this type. Difficulties may arise, however, if a researcher wants to pool several published analyses. While many studies publish allelotypes, some studies provide only the proportion of tumors showing allelic loss among informative tumors at each marker. We describe an extension of the allelotype Markov chain model to handle these frequencies of allelic loss (FAL). The model accommodates missing data across markers through a type of importance sampling known as sequential imputation. Hypothesis testing and estimation of gene location will be discussed. We compare the efficiency of FAL analyses to analyses of allelotypes. Finally, we describe a Gaussian approximation to the likelihood which also accommodates missing data.

e-mail: asterret@bios.unc.edu

LOCATION-SPREAD DISPLAYS OF MANY POINT ESTIMATES AND THEIR ASSOCIATED MEASURES OF DISPERSION

Mario Peruggia * , Ohio State University; Jason C. Hsu, Ohio State University

In many statistical applications, inferential conclusions are summarized numerically by a collection of point estimates. Along with each estimate, a measure of accuracy or dispersion is typically reported and often the two are combined to form some type of interval estimate. Perhaps most commonly, this circumstance occurs when sample means and standard errors are computed for several groups of observations. Because of clutter, traditional graphical summaries become visually ineffective when more than a few pairs of point estimates and associated dispersions have to be represented. The location-spread display introduced in this talk overcomes the limitations stemming from visual clutter and can represent effectively very many pairs. The use of the location-spread representation is illustrated with two examples. In the first example, the representation is used to summarize aspects of the posterior distributions of numerous parameter contrasts in a hierarchical Bayes model. In the second example, the representation is used to summarize thousands of gene expression estimates arising from an analysis of microarray data.

e-mail: peruggia@stat.ohio-state.edu

A GRAPHICAL APPROACH FOR QUALITY CONTROL OF OLIGONUCLEOTIDE ARRAY DATA

Dung-Tsa Chen*, University of Alabama at Birmingham

Microarray technology have advanced genomic research. However, examination of data quality poses a unique challenge due to enormous volume of microarray data. Particularly, quality control of microarray data for array comparability becomes very important because data analysis including incomparable arrays is likely to generate invalid results. Unfortunately, issues of array comparability in microarray data quality control have not been addressed adequately, either in literature or in practice. In contrast to the existence of many analytical methods, no methods are currently available to screen out incomparable arrays. As a result, statistical analysis tends to be contaminated with poor quality data and the results become unreliable. In this study, we propose a graphical approach to address the issue. The proposed approach uses percentile methods to group data, and applies the 2D image plot to display the grouped data. Moreover, an invariant band is employed to quantify degrees of array comparability.

e-mail: dtchen@uab.edu

STATISTICAL DESIGN AND ANALYSIS OF POOLS USING OPTIMAL COVERAGE AND MINIMAL COLLISION

Katja S. Remlinger * , North Carolina State University;
Jacqueline M. Hughes-Oliver, North Carolina State
University; S. Stanley Young, National Institute of Statistical Sciences;
Raymond L. Lam, GlaxoSmithKline

Discovery of a new drug involves screening large chemical libraries to identify active compounds. Screening efficiency can be improved by testing compounds in pools. We consider two criteria to design pools: optimal coverage of the chemical space and minimal collision between compounds within a pool. Five pooling designs are applied to a public data set. We evaluate each method by determining how well the design criteria are met. One design, called MC, uniformly dominates all other designs, but all criteria-designed pools outperform randomly created pools. Furthermore, we discuss blocking and synergism between compounds as other effects that must be investigated when performing pooling experiments. These effects can be especially of interest in the light of combination therapies. A formal analysis of the MC design is performed using a model-based likelihood approach. The effects of blocking and synergism are included into the model.

e-mail: ksremlin@stat.ncsu.edu

PROBABILISTIC HIGH-DIMENSIONAL DISCRIMINATION FOR DRUG SCREENING

Alexander G. Gray*, Carnegie Mellon University

Automated high-throughput drug screening constitutes a critical emerging approach in modern pharmaceutical research. The statistical task of interest is that of discriminating active versus inactive molecules given a target molecule, in order to rank potential drug candidates for further testing. Because the core problem is one of ranking, our approach concentrates on accurate estimation of unknown class probabilities, in contrast to popular non-probabilistic methods which simply estimate decision boundaries. While this motivates nonparametric density estimation, we are faced with the fact that the molecular descriptors used in practice typically contain thousands of binary features. In this paper we attempt to improve the extent to which kernel density estimation can work well in high-dimensional discrimination settings. We present a synthesis of techniques (SLAMDUNK: Sphere, Learn A Metric, Discriminate Using Nonisotropic Kernels) which yields favorable performance in comparison to previous published approaches to drug screening, as tested on a large proprietary pharmaceutical dataset.

e-mail: agray@cs.cmu.edu

SHARED FRAILTY MODELS FOR RECURRENT EVENTS AND A TERMINAL EVENT

Lei Liu *, University of Michigan;
Robert A. Wolfe, University of Michigan;
Xuelin Huang, M. D., Anderson Cancer Center

There has been increasing interest in the analysis of recurrent event data (Cook and Lawless, 2002). In many situations, a terminating event like death can happen during the follow-up period to preclude further occurrence of the recurrent events. Furthermore, the death time may be dependent on the recurrent event history. In this paper we consider frailty models for the recurrent events data in the presence of such a terminal event. We define frailty proportional hazard models for the recurrent and terminal events. The dependence is modeled by conditioning on a shared frailty that is included in both hazard functions. Covariate effects can be taken into account in the model as well. Maximum likelihood estimation and inference are carried out through Monte Carlo EM algorithm with Metropolis-Hastings sampler in the E-step. An analysis of hospitalization and death data for dialysis patients is presented to illustrate the proposed methods. This model avoids the difficulties encountered in alternative approaches which attempt to specify a dependent joint distribution with marginal proportional hazards and yields an estimate of the degree of dependence.

e-mail: liulei@umich.edu

MODELING OF RECURRENT EVENT SURVIVAL DATA BASED ON EXTENSIONS OF THE ANDERSEN-GILL MODEL

Zekarias Berhane*, University of Pittsburgh
Lisa A. Weissfeld, University of Pittsburgh

There are many methods available for the modeling of recurrent time to event data. These methods are based primarily on the use of the Cox proportional hazards model either through a modification of the definition of time at risk, as in Andersen and Gill (1982), conditional modeling as in Prentice, Williams and Petersen (1981), or marginal modeling as in Wei, Lin and Weissfeld (1989). Recurrent event data present a unique modeling problem since the assumption of proportionality of the hazard function tends to break down with each successive recurrence. To address this issue, we propose an extension of the Andersen and Gill (1982) approach with Gray's (1992) spline-based survival model as the underlying model. We use both the additive and time-varying versions of Gray's (1992) model for this extension. With the use of the time-varying version of this model, the assumption of proportionality of the hazard function is relaxed, resulting in an improved model for the recurrent event setting. Use of the Gray's additive model allows the linearity restriction to be relaxed resulting in an improvement as well. We present simulation studies comparing these models and illustrate their use in the rhDNase example presented in Therneau and Hamilton (1997).

e-mail: ztbst1@pitt.edu

CONTROLLING BIAS IN INTERNAL PILOTS WITH THE UNIVARIATE APPROACH TO REPEATED MEASURES

Christopher S. Coffey * , University of Alabama at Birmingham;
Keith E. Muller, University of North Carolina at Chapel Hill;
David Sawrie, University of Alabama at Birmingham

Uncertainty surrounding the error covariance matrix often presents the biggest barrier to achieving accurate power analysis in the ‘univariate’ approach to repeated measures analysis of variance (UNIREP). Internal pilot designs were introduced to resolve such uncertainty about error variance for t-tests. Recently, Coffey and Muller (Statistics in Medicine, 2003) presented simulations in the UNIREP setting which demonstrated that internal pilots allow maintaining power or reducing expected sample size when the covariance matrix used for planning differs from the true value. However, they cautioned against the widespread use of internal pilots in such settings for two reasons. 1) Ignoring the randomness of final sample size can inflate test size, at least with small to moderate sample sizes. 2) The Muller-Barton approximation understates UNIREP power for some conditions, which can lead to over-corrected sample size. We solve the first problem by extending strategies for controlling test size in univariate models to the UNIREP setting. In turn, results illustrate that a new power approximation (Muller, Edwards, and Taylor, JSM Abstract, 2003) solves the second problem.

e-mail: ccoffey@uab.edu

MODELS AND ANALYSIS OF RECURRENT EVENT DATA

Edsel A. Pena*, University of South Carolina
Elizabeth H. Slate, Medical University of South Carolina
Juan R. Gonzalez, Catalan Institute of Oncology, Spain

Many biomedical or health studies involve the situation where the study subject experiences an event of interest in a recurrent fashion. Examples are repeated hospitalizations due to a chronic disease, the occurrence of migraine headaches, epileptic seizures, episodes of depression, and many others. In this talk I will describe a flexible and general class of statistical models that incorporates the impact of interventions after each event occurrence, the effect (possibly weakening or strengthening) of accumulating event occurrences on the subject, the effect of concomitant variables, and the effect of unobserved frailty variables which induce correlations among the inter-event times for a subject. I will describe a semi-parametric approach for estimating the parameters of this class of models, and present properties of the inferential procedures. These procedures will be demonstrated by analyzing some biomedical and public health data sets.

e-mail: pena@stat.sc.edu

REGRESSION ANALYSIS WITH MULTIVARIATE GROUPED AND CONTINUOUS SURVIVAL DATA

Andrea B. Troxel*, University of Pennsylvania School of Medicine
Denise A. Esserman, Columbia University

Multivariate survival data are common in clinical trials, especially when a secondary outcome, such as quality of life (QOL), is of considerable importance. Often, one outcome, such as survival, is measured continuously, while a second is subject to grouping. We present a method that synthesizes previous proposals to analyze multivariate continuous and grouped survival data. We use the continuous and grouped-data versions of the proportional hazards model, and leave the dependence among the multiple events unspecified. The resulting estimators are consistent and asymptotically normal; we provide a robust expression for the asymptotic covariance. Simulation results demonstrate that small-sample behavior is sufficient for use in practical settings, and an example is provided concerning quality of life in colorectal cancer patients.

e-mail: atroxel@cceb.upenn.edu

74. Categorical Data

MARGINAL MODELS FOR BINARY DATA: EXISTENCE, UNIQUENESS AND IMPLICATIONS

Bahjat F. Qaqish*, University of North Carolina at Chapel Hill;
Anastasia Ivanova, University of North Carolina at Chapel Hill

Marginal regression models for correlated binary responses are specified as models for marginal logistic contrasts, including log odds and log odds ratios. The marginal logistic contrasts are functions of cell probabilities obtained through the multivariate logistic transform. However, it is possible to specify numerical values of the marginal logistic contrasts that do not correspond to any valid distribution. Hence, inverting the multivariate logistic transform is of interest in computation, simulation and design of software for fitting marginal models to correlated binary data. We show that the inverse either exists uniquely or does not exist. We describe an algorithm for detecting whether a given set of marginal logistic contrasts corresponds to a valid joint distribution, and for computing that distribution if it exists.

e-mail: qaqish@bios.unc.edu

OPTIMAL NONRANDOMIZED TESTS FOR COMPARING TWO BINOMIAL PROBABILITIES

Weizhen Wang*, Wright State University

A method, different from the traditional p -value method, is proposed to generate tests to compare the probabilities of successes for two independent binomial random variables. This method uses a partial ordering on a small part of the sample space to obtain nonrandomized tests, which are of level- α , have rejection regions including that of Fisher's exact test (1935) as a subset, have sizes larger than those of existing level- α tests in general, and are easy to construct. Extensive numerical studies show that the rejection regions of the proposed tests cannot be enlarged any more with still level- α .

e-mail: wwang@math.wright.edu

THE EFFECT OF OBSERVATIONAL UNIT SIZE ON POISSON GOODNESS-OF-FIT TESTS

John W. Ely, University of Iowa; Jeffrey D. Dawson*, University of Iowa

Empirical evidence suggests that the significance of Poisson goodness-of-fit tests may depend upon the size of the observational units, which can be arbitrarily chosen or created by collapsing smaller units into larger ones. For example, when modeling emergency room admissions related to certain diagnoses, one could potentially analyze the data at the hour, shift, day, week, or month level, and some of these choices result in data that appear to be Poisson while others do not. Furthermore, the relationship between the statistical significance of goodness-of-fit tests and the degree of collapsing can be non-monotonic. To investigate this phenomenon, we simulated data from Poisson, auto-correlated Poisson, zero-inflated Poisson, and negative binomial distributions, at various degrees of data collapsing and at different distributional means. We found that an auto-correlated Poisson process is more likely other non-Poisson distributions to give rise to non-monotonic test statistics.

e-mail: jeffrey-dawson@uiowa.edu

EXACT BOOTSTRAP CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO PROBABILITIES USING DATA FROM TWO INDEPENDENT BINOMIAL SAMPLES WITH SMALL SAMPLE SIZES

Yan Lin*, Medical University of South Carolina
Stuart R. Lipsitz, Medical University of South Carolina

In studies in which subjects are randomized to two treatment groups and the response is binary (eg., success or failure), estimation of the difference in success probabilities is of interest. Estimation of the difference of two independent binomial probabilities is a common problem in medical and health care studies. Use of confidence intervals has been increasingly encouraged in addition to, even in place of, the tests of statistical significance. In this paper, we propose an estimate of the difference of two probabilities in small samples based on the median unbiased estimate of the two probabilities. We also propose the use of the bootstrap to form a confidence interval for the difference of two probabilities based on our estimate. Instead of a Monte Carlo bootstrap, one can easily calculate the 'exact' bootstrap distribution of our estimate of the difference, and use this distribution to calculate confidence intervals.

e-mail: liny@musc.edu

ESTIMATION OF PARAMETERS FOR BINARY OUTCOME DATA IN THE PRESENCE OF MISSING COVARIATE INFORMATION

Gina M. D'Angelo*, University of Pittsburgh;
Lisa Weissfeld, University of Pittsburgh

Case-control designs are frequently used for medical studies with designs ranging from a sampling of all cases to a response based sampling approach. Missing covariate data are inherent in these designs due to the expense of data collection and/or happenstance. Since complete case analysis with data missing at random (MAR) can produce inefficient and biased estimates it is necessary to employ other methods to address these statistical issues. We propose an inverse probability weighted estimating equation technique incorporating score operators to handle covariates MAR with a binary outcome in a case-control study. The benefit of this method over likelihood based methods is the fact that one need not specify the full likelihood. However, a difficulty with this estimating equation technique is specification of the missing data mechanism. In addition to developing this method for a binary outcome and discrete covariates we will compare two approaches used to specify the missing data mechanism. Simulation studies will be performed to validate the method and to examine the robustness of the method to misspecification of the missing data mechanism.

e-mail: dangelo@upci.pitt.edu

A BIAS ANALYSIS OF THE LOGISTIC NORMAL LIKELIHOOD RATIO TEST

Rafael E. Diaz*, Tulane University

The Logistic Normal Likelihood Ratio (LNLR) test has been proposed among the methodologies to compare two groups of proportions. However, despite its effectiveness to test the difference between two groups of proportions, there is a controversy concerning its ability to measure this difference. The model used by the LNLR test is a particular case of hierarchical generalized logistic models. The current interest in developing more accurate techniques to test and measure the parameters of this type of model has encourage the settling of this controversy. In this study a bias analysis is performed for the random parameter of the LNLR test using the recently proposed Laplace 6 approximation. The results of this analysis, for some scenarios of interest, show that this approximation is far superior to the commonly used Penalized Quasi - Likelihood approximation. The positive results of this analysis suggest that it is possible to construct an informative measure for the difference between two groups of proportions with the parameters of the LNLR test by means of prediction intervals. The study of the statistical properties of these intervals will follow as a consequence of the results in this analysis.

e-mail: rdiaze@tulane.edu

MODELING THE ON-OFF PROCESS TO ANALYZE THE POSTURAL INSTABILITY OF PD PATIENTS

Debajyoti Sinha * , Medical University of South
Carolina; Ming-Hui Chen, University of Connecticut;
Peng Huang, Medical University of South Carolina

We propose to study the binary postural instability on-off process that incorporates a Parkinson's disease patient's multiple onset times. A novel latent structure model is proposed for the binary on-off process with time dependent covariates. The proposed model does not require even-spaced visiting schedules and the correlation of the binary process is naturally built through the underlying latent process. Theoretical and computational properties of the proposed model are examined and a real dataset from a Parkinson clinical trial is analyzed to demonstrate the methodology.

e-mail sinhad@musc.edu

TESTING SIMILARITY BETWEEN TWO GROUPS OF CURVES

Yolanda Munoz Maldonado*, Texas A&M University

A test statistic is developed for assessing the similarity between two groups of curves in a functional data analysis setting. The proposed statistic is shown to have a U- statistic representation which is used to derive its large sample properties. Results from an empirical power study are reported, and the method is illustrated using data from the ganglioside concentration in the Locus Coeruleus region for old and young rats.

e-mail: ymunoz@stat.tamu.edu

BOOTSTRAP CONFIDENCE BANDS

Laura L. Johnson * , National Cancer Institute;
Paula Diehr, University of Washington

We proposed a way to jointly model longitudinal health status data and survival data while avoiding imputing data after death, a method for incorporating death into the statistical analysis of longitudinal health status data, and for allowing non-monotone data, the Probability of being Alive and Healthy (PAH). In order to compare these trajectories over time, Bootstrap Percentile Confidence Bands were used to allow comparisons between the two non-monotonic curves. Special emphasis is placed on sampling from a longitudinal dataset.

e-mail: johnslau@mail.nih.gov

DIRICHLET PROCESS MIXTURE MODELS FOR MARKOV PROCESSES

Yongqiang Tang*, North Carolina State University
Subhashis Ghosal, North Carolina State University

We consider the problem of the estimation of the transition density of a Markov process. We do not assume a parametric model and follow the Bayesian approach. We put a prior using a Dirichlet mixture of normal densities. Markov chain Monte Carlo methods are used for the computation of the posterior. Because of an inherent non-conjugacy in the model, usual Gibbs sampling procedure used for the density estimation problem is hard to implement. We propose using the recently proposed “no-gaps algorithm” to overcome this difficulty. The problem of the prediction of the next observation, which has a lot of practical interest, can be identified as the computation of the posterior expectation of the transition density. We show consistency of the Bayes procedures in appropriate topologies by constructing appropriate uniformly exponentially consistent tests and showing that neighborhoods of the true transition density receive positive prior probabilities in an appropriate sense. We extend our methodology to nonlinear time series for real data analysis. Numerical examples show excellent agreement between asymptotic theory and the finite sample behavior of the posterior distribution.

e-mail: ytang@unity.ncsu.edu

FUNCTIONAL MIXED EFFECTS MODELS WITH PRIOR INFORMATION

Li Qin *, University of Pennsylvania;
Wensheng Guo, University of Pennsylvania

In functional data analysis, the basic unit of the data analysis is a curve, so are the parameters. In many situations, some prior information about the curves is available, such as the curves are similar in shape to a family of parametric functions or they are subjected to some linear equality constraints. Such information needs to be accounted for in the model to improve the estimation and inference. In this paper, we propose a new class of functional mixed effects models that incorporate the prior information in both the functional fixed effects and functional random effects. The functional fixed effects and functional random effects are modelled in the same functional space, and therefore the population-average curves and subject-specific curves have the same property. These models can be rewritten in multivariate state space forms and estimated by an $O(N)$ modified Kalman filtering and smoothing algorithm. The proposed models are applied to a cortisol data set obtained from a study on fibromyalgia. The properness of the incorporated prior information is also assessed.

e-mail: lqin@cceb.upenn.edu

REFERENCE PRIOR IN BAYESIAN PENALIZED SPLINE ANALYSIS

Xin Zhao*, Cornell University
Martin T. Wells, Cornell University

Nonparametric (or semiparametric) regression using penalized splines is known to be equivalent to a linear mixed model, $Y = XB + Zu + e$, where XB is the fixed-effects polynomial component of the spline and Zu is the random-effects component of spline basis functions. Further, this model assumes that both u and e have a single variance component and zero covariances. We employ bayesian methods to fit the model, using a reference prior that depends on the eigenvalues of ZZ^T . Based on simulation studies and real data analyses, we found that this prior can automatically adapt to changes in the number of knots to get an appropriately smooth spline estimate. Other commonly used reference priors cannot do this. We recommend the use of this reference prior in bayesian penalized spline analyses. First, it is an objective prior, which is appropriate since we don't want strong prior restrictions on the variance components. Second, it leads to a shrinkage estimator that makes use of the structure of ZZ^T . Third, the automatic adaptation to changes in the number of knots makes it attractive in practice. Finally, our results show that the reference prior also has good frequentist properties.

e-mail: xz45@cornell.edu

76. Surrogate Endpoints, Screening and Assessing Agreement

A CRITICAL REVIEW OF STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN CONTINUOUS MEASUREMENTS

Joseph L. Hagan*, University of Louisville;
Stephen W. Looney, Louisiana State University

Statistical methods that are commonly used to assess agreement between continuous measurements are critically reviewed. A survey of the clinical research literature yielded 138 articles published in 2001 that included some type of method comparison study. These articles were perused in order to determine the type of statistical analyses the investigators used to assess agreement between the two methods being compared. In the 138 articles, 189 method comparison studies were described; the most commonly used statistical technique was the intra-class correlation coefficient (ICC) (118 studies, 62.4%), followed by the Pearson correlation coefficient (PCC) (53 studies, 28.0%), and the Bland-Altman method (25 studies, 13.2%). These results indicate that the PCC is still commonly used for method comparisons, despite the fact that it has been known to be inappropriate for this purpose for over 30 years. The most commonly used method, the ICC, is also known to have shortcomings as a measure of agreement, and has only limited applicability in method comparison studies. Given the current state of the clinical literature with regard to statistical analyses that are used in method comparison studies, statisticians should be more proactive in promoting the use of appropriate statistical techniques for assessing agreement between continuous measurements.

e-mail: sloon1@lsuhsc.edu

STATISTICAL METHODS FOR THE ANALYSIS OF DISTORTION PRODUCT OTOACOUSTIC EMISSIONS

Peter F. Craigmile*, The Ohio State University
Wayne M. King, The Ohio State University

The deleterious effects of undiagnosed and unremediated hearing loss are well documented and mandate that reliable nonbehavioral measures of hearing sensitivity are available to health care professionals. In many patient populations, reliable behavioral data may be difficult or impossible to obtain. Thus, nonbehavioral audiologic measures play a critical role in the assessment of the auditory system and the diagnosis of auditory dysfunction. Distortion product otoacoustic emissions (DPOAEs) are an important nonbehavioral measure of cochlear function, providing a close analogue of the behavioral pure-tone audiogram. DPOAEs are sinusoidal distortion products produced by nonlinearities in the healthy cochlea. The detection of these distortion products is typically accomplished using Fourier based spectral estimates. In this talk we demonstrate how current estimation methods used in the literature can be extended to test for DPOAEs in the presence of correlated background noise. Through models based on normal adult real-ear noise measurements, we demonstrate the distribution of statistics for DPOAEs detection under the null hypothesis, and illustrate the power of such tests under fixed alternatives by simulation.

e-mail: pfc@stat.ohio-state.edu

STATISTICAL MODELS FOR VACCINE CORRELATES OF PROTECTION

Andrew J. Dunning*, Wyeth Vaccines Research

Many diseases that vaccines prevent are severe or life-threatening – diphtheria, smallpox and typhoid are examples. Thus, in the search for improved vaccines, placebo-controlled trials are often considered unethical. These diseases also tend to be rare, so active comparator trials are expensive. Therefore, there is considerable interest in surrogate endpoints. Vaccines work by simulating the immune system in a way similar to the disease. If an immunological assay can be found which indicates whether a person is protected against disease, threshold levels of the assay may be considered as correlates of protection against the disease. Existing models for vaccine correlates of protection have difficulty distinguishing between absence of disease as a result of immunologic protection and absence of disease resulting merely from not being exposed. We present statistical models that separate these effects, thus providing improved methods for estimating the efficacy of vaccines based on immunological assays.

e-mail: adunning2@earthlink.net

LEFT, RIGHT AND INTERVAL BIVARIATE CENSORED DATA: EVALUATING SCREENING MAMMOGRAPHY IN THE PRESENCE OF LEAD-TIME BIAS, LENGTH BIAS AND OVER-DETECTION

Jonathan D. Mahnken*, University of Texas
Wenyaw Chan, University of Texas
Jean L. Freeman, University of Texas

Of the large clinical trials evaluating screening mammography efficacy, none included women ages 75 and older. Recommendations on an upper age limit at which to discontinue screening are based on indirect evidence and are not consistent. Screening mammography will be evaluated using observational data from the SEER-Medicare linked database. Measuring the benefit of screening mammography is difficult due to the impact of lead-time bias, length bias and over-detection. The underlying conceptual model divides the disease into two stages; pre-clinical (T0) and symptomatic (T1) breast cancer. Treating the time in these phases as a pair of dependent bivariate observations, (t_0, t_1) , estimates will be derived to describe the distribution of this random vector. To quantify the effect of screening mammography, statistical inference will be made about the mammography parameters that correspond to the marginal distribution of the symptomatic phase duration (T1).

e-mail: jdmahnke@utmb.edu

THE ANALYSIS OF CLUSTERED MATCHED-PAIR DATA

Valerie L. Durkalski*, The Clinical Innovation Group;
Yuko Y. Palesch, The Clinical Innovation Group and
The Medical University of South Carolina; Stuart
Lipsitz, The Medical University of South Carolina;
Phil Rust, The Medical University of South Carolina

Evaluation of the performance of a new diagnostic procedure with respect to a standard procedure arises frequently in practice. When the response of interest is in a dichotomous form, McNemar's test for matched-pair samples is used to determine if the proportion of correctly identified disease states is the same for both procedures. The main assumption of this test is independent paired responses; however, when more than one outcome from an individual is measured by each procedure, the data are clustered. Examples of such cases can be seen in colon cancer screening studies. Variance adjustment methods for the analysis of clustered matched-pair data have been proposed; however, because of unequal cluster sizes, variability of correlation structures within a cluster (within paired responses in a cluster as well as between paired responses in a cluster), and unequal success probabilities among the clusters, the performances of some available methods are not consistent. This research proposes the use of the method of moments to calculate a consistent variance estimator. Using Monte Carlo simulation, the size and power of the proposed test are compared to those of two currently available methods.

e-mail: durkalsv@musc.edu

ISOTONIC LOGISTIC DISCRIMINATION

Sungyoung Auh*, University of Pittsburgh
Allan R. Sampson, University of Pittsburgh

Linear logistic discrimination has been a standard approach in discrimination applications. However, the underlying linear constraint is restrictive, and hence relaxing the linearity to monotonicity is desirable. In this presentation, we propose an isotonic logistic discrimination procedure which generalizes linear logistic discrimination by allowing linear boundaries to be more flexibly shaped as monotone functions of the discriminant variables under each of the three standardly considered sampling schemes, (i) prospective (ii) mixture and (iii) retrospective, used for obtaining a training data set. Likelihood-based inference is provided for each of the three sampling schemes. The proposed method is illustrated through an example taken from a breast cancer study and compared with two recent approaches to monotone discrimination in application.

e-mail: suast1@pitt.edu

MODELING THE AGREEMENT OF DISCRETE BIVARIATE SURVIVAL TIMES USING KAPPA COEFFICIENT

Ying Guo * ; Emory University;
Amita K. Manatunga, Emory University;
Nengjun Yi * , The University of Alabama at Birmingham

Estimation of agreement is important to assess the interchangeability of different measurements or the acceptability of a relatively simple method as a replacement for a gold standard. In this paper, we develop a local kappa coefficient to assess the agreement between two discrete time to events measurements. We model the marginal distributions for the two event times and the local kappa coefficient in terms of covariates. An estimating equation is used for modeling the marginal distributions and a pseudo-likelihood procedure (Shih, 1998) is used to estimate the agreement parameters in the kappa model. The performance of the estimation procedure is examined through simulations. The proposed method can be extended to multivariate discrete survival distributions.

e-mail: yguo2@sph.emory.edu

CURRENT DEVELOPMENTS AND STATISTICAL CHALLENGES IN DEVELOPING SYNDROMIC SURVEILLANCE SYSTEMS FOR ANTI-BIOTERRORISM

Lori Hutwagner * , Centers for Disease Control

The intentional release of *Bacillus anthracis* by mail in the fall of 2001, and recent outbreaks of SARS and moneypox stress the need to improve public health's ability to detect outbreaks of infectious diseases, whether naturally occurring or caused by bioterrorism. Consequently, there is considerable interest in innovative surveillance systems that use clinical, pre-diagnostic data sources, such as visits to emergency departments (ED), for rapid detection and public health response. The analysis of specific signs and symptoms prior to diagnosis is referred to as syndromic surveillance.

Syndromic surveillance encompasses data from emergency departments, emergency call systems, physician office data, and over-the-counter drug sales. The field of syndromic surveillance is rapidly expanding, as systems are being implemented by numerous state and county health departments. However, current empirical evaluation and validations syndromic surveillance are lacking.

Syndromic surveillance faces epidemiological issues and statistical challenges. Not all syndromic surveillance systems are created equal. A critical issue regarding generalizability across different populations has not yet been appropriately evaluated – some data sources have limited applicability to local or national target populations. Locations using syndromic surveillance with small population census have small number problems. Small numbers are also an issue when the data is stratified by area of interest, for example by hospitals. Small numbers also play a role when looking for syndromes that represent rare disease. Additionally, there is the statistical issue of what is the best statistic for comparing the results. Evaluation and validation studies have reported sensitivity, specificity, time to detection, Kappa, ROC curves, and AMOC curves to name a few. The statistical challenges also extend into multiple data sources. These issues range from the combination of these data sources to the issues of which data provide a better interpretation of what is happening with the health of the community. The evaluation of syndromic surveillance is ongoing and is currently based on the identification of naturally occurring outbreaks with assumptions that a bioterrorism event will have similar resulting patterns. As these evaluations unfold, more statistical issues will be raised.

SYNDROME SURVEILLANCE SYSTEM FOR ANTI-BIOTERRORISM: STATISTICAL ISSUES IN TREND ANALYSIS AND OUTBREAK DETECTION

Wei Wang * , University of South Florida; Yiliang Zhu, University of South Florida

With the very real threat of bioterrorism, the critical need for timely detection of an outbreak has accelerated the time frame for major enhancements to the public health infrastructure. Medical syndrome surveillance system is one of the earliest developments. A surveillance system monitors healthcare utilization patterns in real time for the first signs of a covert germ warfare attack, which may appear either as a single observation or clusters of infected victims seeking health care. Among all statistical detection techniques currently employed, most utilize two major components, an empirical time series model for baseline trend in the absence of aberrations, and algorithms for detecting deviation from the normal levels. Despite a large number of such algorithms being used in many systems, relatively limited has been done to evaluate the performance of these techniques. We will review both statistical techniques for baseline trend analysis and statistical algorithms for aberration detection in terms of sensitivity and specificity, and compare several algorithms commonly deployed. We illustrate these results through theoretical verification, case analyses, and simulations.

e-mail: wwang5@hotmail.com

MISSING DATA IN DAILY SYNDROMIC SURVEILLANCE SYSTEMS

Martin Kulldorff*, Harvard Medical School
Farzad Mostashari, New York City Department of Health and Mental Hygiene
Rick Heffernan, New York City Department of Health and Mental Hygiene
Jessica Hartman, New York Academy of Medicine

Syndromic surveillance systems for the early detection of disease outbreaks are based on daily analyses of health data from for example hospital emergency rooms, pharmacy sales or ambulance dispatches. Since data typically arrives from multiple data providers, there is missing data whenever a provider is unable to transfer the data within a day of it being collected, and that may happen for a variety of reasons. The prospective space-time scan statistic is commonly used for disease outbreak detection. First we show that if ignored, missing data can create strong false disease outbreak detection signals. We then show how the prospective space-time scan statistic can be adjusted for missing data. We illustrate both the problem and solution using hospital emergency department surveillance data from New York City.

e-mail: martin_kulldorff@hms.harvard.edu

78. Novel Statistical Approaches to High Dimensional Analyses

ANALYSIS OF GENOME-WIDE EPISTASIS VIA BAYESIAN MODEL SELECTION

Nengjun Yi*, The University of Alabama at Birmingham

Epistasis, the interaction among genes, is speculated to be ubiquitous in the genetic control of most common human diseases, e.g., obesity, hypertension, and cancer. Analysis of genome-wide epistasis must accommodate a very large number of potential genetic effects, which presents a major challenge to determining the genetic model with respect to the number of QTL, their positions and their genetic effects. In this study, we use the methodology of Bayesian model and variable selection to develop strategies for identifying multiple QTL with complex epistatic patterns in experimental designs. We treat the number of QTL as random variable. Under the current number of QTL, we introduce indicator variables to represent the number of effects and the actual effects presented in the model. We use the methodology of Bayesian model and variable selection to develop strategies for searching multiple QTL with complex epistatic patterns and estimating the associated parameters simultaneously. Utility and flexibility of the method are demonstrated using both simulated and real data sets. Sensitivity of posterior inference to prior specifications of the number and genetic effects of QTL are investigated.

e-mail: nyi@uab.edu

ANALYSIS OF GENOME-WIDE EPISTASIS VIA BAYESIAN MODEL SELECTION

Nengjun Yi *, The University of Alabama at Birmingham

Epistasis, the interaction among genes, is speculated to be ubiquitous in the genetic control of most common human diseases, e.g., obesity, hypertension, and cancer. Analysis of genome-wide epistasis must accommodate a very large number of potential genetic effects, which presents a major challenge to determining the genetic model with respect to the number of QTL, their positions and their genetic effects. In this study, we use the methodology of Bayesian model and variable selection to develop strategies for identifying multiple QTL with complex epistatic patterns in experimental designs. We treat the number of QTL as random variable. Under the current number of QTL, we introduce indicator variables to represent the number of effects and the actual effects presented in the model. We use the methodology of Bayesian model and variable selection to develop strategies for searching multiple QTL with complex epistatic patterns and estimating the associated parameters simultaneously. Utility and flexibility of the method are demonstrated using both simulated and real data sets. Sensitivity of posterior inference to prior specifications of the number and genetic effects of QTL are investigated.

e-mail: nyi@uab.edu

USE AND ESTIMATION OF FALSE DISCOVERY RATE IN GENETICS

Chiara Sabatti *, UCLA

This talk will focus on the uses of FDR in the analysis of genetic data and on the methodological developments that these problems have motivated. The first part of the talk will offer a selective review of the applications of FDR procedure in gene-mapping and gene expression arrays analysis, paying particular attention to which interpretation of FDR they offer. In the second part of the talk I will present a hierarchical Bayesian model for the analysis of gene expression arrays that leads to results comparable to the ones of FDR thresholding, while allowing for flexible assumption on gene-specific variances.

e-mail: csabatti@mednet.ucla.edu

POWER & SAMPLE SIZE ESTIMATION WHEN MANY SIMULTANEOUS HYPOTHESIS TESTS ARE PLANNED

Grier P. Page*, University of Alabama at Birmingham
Gary Gadbury, University of Missouri – Rolla
Jode W. Edwards, University of Alabama at Birmingham
David B. Allison, University of Alabama at Birmingham

High dimensional biological studies often involve testing thousands of hypotheses in a single experiment. In the case of microarray studies as many as 65,000 tests may be conducted for each experimental condition. It is important in the context of these studies to determine the sample size that will allow meaningful interpretations of the data. Traditional power analysis methods may not be well suited to this task when thousands of hypotheses are tested in discovery-oriented basic research. We introduce the concept of expected discovery rate (EDR), which is proportion of genes which are truly differentially expressed in the experiment that will be found to be significantly differentially expressed, with an approach that combines parametric mixture modeling of the p-value distribution of pilot data with parametric bootstrapping to estimate the sample size needed for a desired accuracy of results as measured by a combination of the significance, EDR, and the posterior probabilities of false positive and negative results.

e-mail: gpage@uab.edu

79. Non-Inferiority/Equivalence Trials

DELTA WHAT? CHOICE OF OUTCOME SCALE IN NON-INFERIORITY TRIALS

Rick Chappell*, University of Wisconsin – Madison

Equivalence trials are experiments which attempt to show that one intervention is not too much inferior to another on some quantitative scale. The cutoff value is commonly denoted as Delta. For example, one might wish to show that the hazard ratio of nonfatal cardiovascular event or death among patients given an experimental blood-thinning drug versus a currently approved drug thinning drug is 1.3 or less. Naturally, a lot of attention is given to choice of Delta. In addition to this, I assert that even more than in superiority clinical trials the *_scale_* of Delta in equivalence trials must be carefully chosen. Since null hypotheses in superiority studies generally imply no effect, they are often equivalent or at least compatible when formulated on different scales. However, nonzero Deltas on one scale usually conflict with those on another. For example, the four hypotheses of arithmetic or multiplicative differences of either survival or hazard in general all mean different things unless Delta = 0. This can lead to difficulties in interpretation when the clinically natural scale is not a statistically convenient one. Difficulties and solutions are illustrated using a trial of equivalence in Stroke prevention (Darius, 2002) and proposed designs for crossover equivalence trials suggested by Tang, Tang and Chan (2003) and Tango (1998).

e-mail: chappell@stat.wisc.edu

CHOICE OF NON-INFERIORITY MARGIN

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

Choice of non-inferiority margin depends very closely on the objective of the non-inferiority analysis. If the analysis aims at demonstrating that the experimental treatment under study is not much worse (traditional notion of non-inferiority) than the selected positive control, then a relevant indifference zone for the treatment difference would have to be clearly defined in selection of the non-inferiority margin. On the other hand, if a part of the trial objective is the efficacy of the experimental treatment, then a major consideration in the margin selection is whether the non-inferiority analysis can lead us to conclude that the experimental treatment would have beaten placebo had the placebo been studied. Another concept relating to these different objectives concerns retention of a certain proportion of the control effect. A critical issue in the margin selection pertains to the statistical errors associated with the non-inferiority analysis in relation to the trial objective. In this presentation, we shall revisit these issues and evaluate some methods of margin selection.

e-mail: hung@cder.fda.gov

NON-INFERIORITY AND EQUIVALENCE MARGINS IN CLINICAL PHARMACOLOGY

Scott D. Patterson*, GlaxoSmithKline Pharmaceuticals

Equivalence trials have a long history in clinical pharmacology. Following review of the development of standards for average bioequivalence testing used for comparing formulations of the same drug substance, we will consider the implications of regulatory acceptance standards proposed recently for population and for individual bioequivalence. We will conclude with an overview of the application of average equivalence testing in other areas of clinical pharmacology drug development and discuss the choice of regulatory standards for acceptance testing.

e-mail: scott.d.patterson@gsk.com

CONTINUOUS TOXICITY MONITORING IN PHASE II TRIALS IN ONCOLOGY

Anastasia Ivanova * , University of North Carolina at Chapel Hill;
Bahjat F. Qaqish, University of North Carolina at Chapel Hill;
Michael Schell, University of North Carolina at Chapel Hill

The goal of a phase II trial in oncology is to evaluate the efficacy of a new therapy. The dose investigated in a phase II trial is usually an estimate of a maximum tolerated dose (MTD) obtained in a preceding phase I trial. Since this estimate might not be precise, stopping rules for toxicity are routinely used in Phase II trials. We give recommendations on how to construct such rules. Estimation of the probability of toxicity and response is also discussed.

e-mail: aivanova@bios.unc.edu

EFFECTS OF AIDS VACCINE ON SUB-POPULATIONS OF CD4 T CELLS, CD8 T CELLS AND B CELLS UNDER HIV INFECTION

Wai-Yuan Tan, The University of Memphis
Ping Zhang*, The University of Memphis
Xiaoping Xiong, St. Jude Children's Research Hospital

AIDS vaccine can be considered as a pathogen without infectivity and hence no harm. It will activate the CD4 T cells which in turn will secrete cytokines to activate CD8 T cells and B cells. The activated CD8 T cells would differentiate to yield primed CD8 T cells which will kill the productively HIV-infected CD4 T cells (cytotoxic effect) whereas the activated B cells will differentiate to yield plasma B cells which will generate HIV antibodies to neutralize or block HIV (humoral effects). Furthermore the activated CD4 T cells would retire to become memory CD4 T cells which will be reactivated to become activated CD4 T cells upon future infection. To assess effects of AIDS vaccine, in this paper we will proceed to develop a stochastic model for the subset of CD4 T cells (Virgin CD4 T cells, activated CD4 T cells, effector CD4 T cells and memory CD4 T cells), the subset of CD8 T cells (Virgin CD8 T cells, activated CD8 T cells, primed CD8 T cells and memory CD8 T cells) and the subset of B cells (Virgin B T cells, activated B cells, plasma B cells and memory B cells). We will develop stochastic differential equations for these subsets. By using these stochastic equations, through computer simulation we will illustrate how the AIDS vaccine elicits its effects through cytotoxic and humoral avenues.

e-mail: pzhang1@memphis.edu

A CRITICAL HISTORY OF INDIVIDUAL AND COLLECTIVE ETHICS

Charles M. Heilig*, Centers for Disease Control and Prevention
Charles Weijer, Dalhousie University

Much ethical discourse in trial methodology turns on the purported tension between the needs of the individual patient in a trial and the needs of all future patients whose treatment stands to be informed by the results of that trial. These ideas were explicitly rendered as individual ethics and collective ethics in 1971 in order to motivate a mathematical solution to the posed ethical dilemma. A critical historical and thematic analysis of individual and collective ethics yields five key points: (1) The normative force of these concepts has never been adequately argued. (2) Individual and collective ethics do not solve the problem of ethically leveraging accumulating data. (3) The notions of the “individual” and the “collective” are too vague to prompt clear moral imperatives. (4) These concepts have never successfully been linked to a standard ethical framework. (5) And statistics and statisticians have a key role to play in concert with research ethicists in devising and justifying ethical trial methodology. Finally, the presentation offers some constructive directions for either rehabilitating individual and collective ethics or advancing an alternative system to meet the needs targeted by individual and collective ethics.

e-mail: cheilig@cdc.gov

EVALUATION OF POWER CHANGE BASED ON CORRECTNESS OF MODEL, SAMPLE SIZE, AND DETECTABLE DOSAGE IN FIRST HUMAN DOSE STUDIES

Adeline Yeo * , Eli Lilly and Company

Pre-clinical data is key to powering first human dose (FHD) studies in the pharmaceutical industry. Nonlinear dose-response relationships, starting with a linear ascend and ending with a plateau, are often observed in pre-clinical studies. However, the nonlinear models are not easy to implement in analyses. Hence, the use of linear models is generally preferred over nonlinear models. The goal of this presentation is to perform simulations of FHD data based on the distribution of pre-clinical data to assess the power loss in fitting a linear regression model and an ANOVA model with Dunnett adjustments. The choice of model fit provides a balance between power and robustness. The overall results showed that different level of power loss occurs with different sample sizes for each dosing group. In addition, while the linear regression model offers minimal power loss when the linear dose proportionality assumption is satisfied, it is a less robust approach compared to the ANOVA model in failing to meet the assumption.

e-mail: yeoaa@lilly.com

ESTIMATING CORRECT-USE CONDOM FAILURE PROBABILITIES

Mark A. Weaver*, Family Health International
Douglas J. Taylor, Family Health International
Rosalie C. Dominik, Family Health International

In a clinical study comparing the failure probabilities of two condom types, the sample of all reported acts of intercourse in which a study condom was used by a randomized participant is typically defined to be the primary analysis sample. However, it may also be desirable to make comparisons among only those acts in which the participants correctly followed all condom use instructions before, during, and after the act of intercourse. The timing associated with the definition of correct use creates a dilemma in that an act cannot be classified as a 'correct-use act' until after the completion of both intercourse and withdrawal; if a condom fails during intercourse then the couple has no chance at correct use during withdrawal. As a result of the conditional nature of this problem, it is not a simple matter to specify a 'correct-use' subset of the primary analysis sample. We develop estimators for the correct-use failure probabilities, the corresponding standard errors, and we develop test statistics for comparing the correct-use failure probabilities between condom groups. We demonstrate the utility of the proposed methods by applying them to data from a clinical study of condom contraceptive effectiveness.

e-mail: mweaver@fhi.org

81. Gene Expression Analysis

STATISTICAL METHODS FOR ANALYZING TISSUE MICROARRAY DATA

Xueli Liu *, UCLA;
Vladimir Minin, UCLA;
Yunda Huang, UCLA;
David B. Seligson, UCLA;
Steve Horvath, UCLA

Tissue microarrays (TMAs) are a new high-throughput tool for the study of protein expression patterns in tissues and are increasingly used to evaluate the diagnostic and prognostic importance of biomarkers. TMA data are rather challenging to analyze. Covariates are highly skewed, non-normal, and may be highly correlated. We present statistical methods for relating TMA data to censored time-to-event data. We review methods for evaluating the predictive power of Cox regression models and show how to test whether biomarker data contain predictive information above and beyond standard pathology covariates. We use bootstrap methods to validate model fitting indices such as the C-index. We also present data mining methods for characterizing high risk patients with simple biomarker rules. Since researchers in the TMA community routinely dichotomize biomarker expression values, survival trees are a natural choice. We also use bump hunting (patient rule induction method), which we adapt to the use with survival data. The proposed methods are applied to a kidney cancer tissue microarray data set.

e-mail: xueli@mednet.ucla.edu

ASSOCIATIONS BETWEEN GENE EXPRESSIONS AND SURVIVAL TIMES IN THE PRESENCE OF COVARIATES

Taesung Park*, Seoul National University Seoul, Korea

Sin-Ho Jung, Duke University

Seung Yeoun Lee, Sejong University Seoul, Korea

Microarray technology makes it possible to monitor expression levels of thousands of genes simultaneously, while making it possible to understand genome-wide gene regulations and interactions. In the area of cancer clinical trials, the gene expression data are commonly observed along with clinical information such as survival times and covariates. Cox models are most commonly used to identify genes providing prognostic information on patient survival. In this study, we propose a testing procedure to identify prognostic genes in the presence of other covariates. The proposed method uses a permutation test based on the martingale residuals from Cox models. The proposed test is illustrated using microarrays data.

e-mail: tspark@stats.snu.ac.kr

INFERENCE IN HIGH DIMENSION LOW SAMPLE SIZE SETTINGS WITH CENSORED OUTCOMES

Brent A. Johnson*, University of North Carolina

The analysis of high dimension low sample size data, such as micro-arrays or image data, is commonplace in biostatistics today. Much of the recent work has focused on outcomes which are known or assumed to be known, such as tumor type or status. However, outcomes which may be censored, e.g. survival or relapse time, are also common in clinical data and of great value to clinicians. An inferential method based on a penalized likelihood is discussed and illustrated in a breast cancer dataset.

e-mail: bjohnson@bios.unc.edu

MODELING THE RELATIONSHIP BETWEEN LVAD SUPPORT TIME AND GENE EXPRESSION CHANGES IN THE HUMAN HEART BY PENALIZED PARTIAL LEAST SQUARES

Xiaohong Huang*, University of Minnesota
Wei Pan, University of Minnesota

Heart failure affects more than 20 million people in the world. Heart transplantation is the most effective therapy, but the number of eligible patients far outweighs the number of available donor hearts. The left mechanical ventricular assist device (LVAD) has been developed as a successful substitution therapy that aids the failing ventricle while a patient is waiting for the donor heart. We obtained genomics data from paired human heart samples harvested at the time of LVAD implant and explant. The heart failure patients in our study were supported by the LVAD for various periods of time. The goal of this study is to model the relationship between the time of LVAD support and gene expression changes. To serve the purpose, we propose a novel penalized partial least squares (PPLS) method to build a regression model. Compared with partial least squares (PLS) and Breiman's random forest method, PPLS gives the best prediction results for the LVAD data.

e-mail: xiaohong@biostat.umn.edu

HIDDEN MARKOV MODELING OF GENOMIC INTERACTIONS

Ernst C. Wit, University of Glasgow;
Nial Friel, University of Glasgow;
John D. McClure*, University of Glasgow

Microarray technology has made the simultaneous measurement of gene transcription a routine activity. Whereas gene transcription is only one stage in the complex genomic process of living organisms, it gives a fascinating insight in one aspect of this activity across the whole genome. Gene regulation is a complex biological process, which involves gene-gene and gene-protein interactions. An operator region, to which the enzyme polymerase can bind to start transcription, precedes the gene sequence. Such local features regulating transcription, pose the question whether there might be local spatial gene interactions. We define a Hidden Markov Model (HMM) to relate the observed expression levels to hidden states "Up", "Down" and "Same" for a time-series gene expression dataset. A Potts Model is identified to describe the interactions between neighboring states. A typical problem in these types of model is the estimation of the hidden parameters because of the intractability of the normalizing constant. Recent work by Pettitt et al (2002) provides a clue to avoid to use pseudolikelihood and to solve this issue for a wide class of HMMs.

e-mail: ernst@stats.gla.ac.uk

IDENTIFICATION OF GENOMIC ABERRATIONS FROM MICROARRAY DATA

Chiang-Ching Huang*, Northwestern University
Jeremy Taylor, University of Michigan

With the advent and recent proliferation of genomic technologies such as gene expression arrays, researchers are now able to explore gene expression patterns for the majority of the genes in the genome. One active research area using gene expression profiles is the study of the transcriptome map. Several studies from various genomes have revealed unexpected gene clusters within which the gene expression levels are highly correlated. To facilitate the search for such clusters in gene expression data, we propose a type of hidden Markov model (HMM) that will infer expression status for genes along the chromosomes. We construct a HMM in which gene expression can be categorized into five states: region of overexpression, singleton of overexpression, normal expression, region of underexpression, and singleton of underexpression. This model is applied to a lung cancer gene expression data set to search for abnormal expression regions. We also assess the global impact of those regions on the patients' clinical variables, such as tumor stage, tumor differentiation, and patient survival.

e-mail: huangcc@northwestern.edu

SEMIPARAMETRIC REGRESSION FOR MICROARRAY GENE EXPRESSION DATA: SUPPORT VECTOR MACHINES AND LINEAR MIXED MODELS

Dawei Liu*, University of Michigan;
Xihong Lin, University of Michigan;
Debashis Ghosh, University of Michigan

We propose a semiparametric regression model to relate a normal clinical outcome to clinical covariates and gene expressions, where the clinical covariate effects are modeled parametrically and gene expression effects are modelled nonparametrically using the technique of support vector machine. The nonparametric function of gene expressions takes into account the fact that the number of genes related to the clinical outcome is likely to be large and the genes are likely to interact with each other. The maximum penalized likelihood method is used to estimate the regression coefficients and nonparametric function. We show that the dual problem of the primal support vector machine problem can be formulated using a linear mixed effects model. Estimation hence can proceed within the linear mixed model framework using standard mixed model software. Both the regression coefficients of the clinical covariate effects and the support vector estimator of the nonparametric gene expression function can be obtained using the Best Linear Unbiased Predictor in linear mixed models. The smoothing parameter can be estimated as a variance component in linear mixed models. A score test is developed to test for the significant gene expression effects. The methods are illustrated using a prostate cancer data set and evaluated using simulations.

e-mail: liudawei@umich.edu

GOODNESS-OF-FIT TEST FOR NONLINEAR MODELS VIA LOCAL LINEAR SMOOTHERS

Chin-Shang Li*, St. Jude Children's Research Hospital

I propose a data-driven test that assesses the lack of fit of nonlinear regression models. The comparison of local linear kernel and parametric fits is the basis of this test, and specific boundary-corrected kernels are not needed at the boundary when local linear fitting is used. Under the parametric null model, the asymptotically optimal bandwidth can be used for bandwidth selection. This selection method leads to the data-driven test that has a limiting normal distribution under the null hypothesis and is consistent against any fixed alternative. The finite-sample property of the proposed data-driven test is illustrated, and the power of the test is compared with those of some existing tests via simulations. We illustrate the practical use of the proposed test with a real-life dataset of ratios of weights to heights for boys.

e-mail: chinshang.li@stjude.org

CONDITIONALLY GENERALIZED LINEAR MODELS

Joan G. Staniswalis *, University of Texas at El Paso

This research involves the use of varying-coefficients within a likelihood framework using the exponential family of distributions. An algorithm is proposed for fitting the varying-coefficients in a regression model that is non-linear when the parameters are simultaneously unknown, but linear when the parameters are considered one-at-a-time. The approach combines backfitting with iteratively reweighted least squares, providing an alternative to linearization techniques used with non-linear models, or a full-scale profile likelihood approach. Regression diagnostics for detecting outliers and influential points are briefly considered. An application to renal failure data for Texas provides an estimate of disease prevalence by ethnicity and economic status using aggregate information available at the county level, not the subject level.

e-mail: joan@math.utep.edu

A UNIFIED APPROACH TO TESTING AND MEASURING ASSOCIATION THROUGH A SEMIPARAMETRIC ODDS RATIO MODEL

Hua Yun Chen*, University of Illinois at Chicago

Odds ratio and correlation coefficient are often used to measure the association between two discrete variables and two continuous variables respectively. When either variable takes a mixture of discrete and continuous values, it is not clear how the association between the two variables can be measured. We propose a semiparametric odds ratio model that unifies approaches for the discrete variables and for continuous variables, and thus can be applied to the latter case to measure the association. The score test generated from the proposed model encompasses many known tests for testing conditional independence in varying situations. When adapted to test no correlation among clustered observations, the proposed method directly tests the hypothesis of no correlation among clustered observations while the traditional score test based on variance components model usually test both no correlation and no overdispersion at the same time. In addition, the proposed method can be applied to the retrospective study such as the familial study in genetics. Estimation of the association parameter is also discussed and the proposed method is applied to studying the bacterial virgamosis in women with high risk of HIV infection.

e-mail: hychen@uic.edu

SEMIPARAMETRIC INFERENCES IN STABILITY STUDIES

Young Kyoung Min*, University of Florida;
Annpey Pong, Forest Laboratories, Inc.;
Ying Qing Chen, University of Florida and University of California at Berkeley

Semiparametric regression models have been proposed and used in stability analysis to estimate the shelf life of drug product. In this paper, a new estimation approach based on the comparability of random variation of continuous repeated outcomes is proposed. The new approach will be studied under the FDA's full sampling plan, and the reduced sampling plans such as matrixing and bracketing designs. The methodologies are demonstrated by simulation studies and applications in practical data analysis.

e-mail: ykymin@yahoo.com

A SEMIPARAMETRIC BAYESIAN MODEL FOR TIME SERIES DATA

Kaushik Ghosh*, National Cancer Institute
Ram C. Tiwari, National Cancer Institute

We model a time series with piecewise linear segments, where the slopes of the individual segments follow a nonparametric distribution G . G is assumed to have a Dirichlet process prior. A Markov Chain Monte Carlo scheme is developed to fit the model and project it into the future. This method is illustrated through an application to short-term prediction of cancer mortality in the U.S.

e-mail: ghoshk@mail.nih.gov

83. Spatial Modeling II

SPATIAL ANALYSIS OF ALCOHOL AVAILABILITY AND VIOLENT CRIME

Dennis Gorman, Texas A&M University;
Li Zhu *, Texas A&M University;
Scott Horel, Texas A&M University

Previous studies have shown a relationship between violent crime and alcohol outlet density using geospatial analysis. This study examined the relationship between neighborhood alcohol outlet densities, socio-structural characteristics and violent crime. Neighborhood socio-structural, alcohol density and violent crime data were collected from archival sources, and analyzed using geospatial analyses. The spatial analysis showed that total alcohol outlet densities contributed significantly to violent crime within target zip codes but not in adjacent zip codes, while population densities in adjacent zip codes contributed significantly to violent crime.

e-mail: lizhu@srph.tamushsc.edu

SECOND-ORDER SPATIAL ANALYSIS OF EPIDERMAL NERVE FIBERS

Lance A. Waller*, Emory University

Leong L Traci, Emory University

William R. Kennedy, University of Minnesota

Gwen Wendelschafer-Crabb, University of Minnesota

Aila Sarkka, Chalmers University of Technology

The density of epidermal nerve fibers (ENFs) in the skin negatively correlates with progression of diabetic neuropathy. Images of skin blisters suggest (qualitatively) that the spatial pattern of ENFs may also change as the density declines. The pair-correlation function defines the correlation between observed ENF locations as a function of distance. We explore the diagnostic capability of the pair-correlation function in identifying the suggested change in spatial pattern. In particular, we assess intra- and inter-individual variability in the pair-correlation function based on skin biopsies and explore approaches of statistical comparison between families of pair correlation function curves.

e-mail: lwaller@sph.emory.edu

BAYESIAN SPATIAL PREDICTION USING RJMCMC

Arin Chaudhuri*, North Carolina State University

Montserrat Fuentes, North Carolina State University

David Holland, U.S. Environmental Protection Agency

The spatial distribution of air pollutants like Ozone are usually non stationary in nature and hence there is a need to develop non-stationary models to capture various kinds of non-stationarity. In this paper we propose a hierarchical bayesian non-stationary model obtained by the convolution of stationary process. The dimensionality of the parameters is not fixed. We use Reverse Jump MCMC developed by Green to estimate the parameters.

e-mail: achaudh@ncsu.edu

ON THE STATISTICAL MODELING AND ANALYSIS OF ADVERSE HEALTH OUTCOMES DUE TO CHRONIC ARSENIC EXPOSURE

Munni Begum * , University of North Carolina at Chapel Hill;
Pranab K. Sen, University of North Carolina at Chapel Hill Chapel Hill

The statistical task in an environmental risk assessment (ERA) of toxicants is involved primarily in the assessment of exposure and in the formulation of concentration /dose – response model. These stressors often work in synergism and slow progression, also only a part of the ambient toxicity enters the human uptake process. It is important to model toxicokinetics of the chemical in the body system that would help better understanding of absorption, metabolism, distribution and elimination process of the chemical compound. The statistical modeling of toxicokinetics as well as the spatio-temporal distribution of the target toxic material in a target region would determine the prevalence level of the exposure. The mode of absorptions of the toxicants and their in vivo biological activity and reaction also deserve to be addressed properly. In this paper we focus on statistical modeling involved in risk assessment of Arsenic. A non homogeneous Poisson process model is proposed for modeling toxicokinetics while a multivariate log normal model taking the spatial pattern over time into account is proposed to assess spatio-temporal distribution. An appropriate stochastic model for formulating the exposure – health hazard risk association is also proposed. The plausibility of causation of arsenic with multiple health outcomes is explored.

e-mail: begum@email.unc.edu

MULTIVARIATE PARAMETRIC SPATIOTEMPORAL MODEL FOR COUNTY LEVEL BREAST CANCER SURVIVAL DATA

Xiaoping Jin*, University of Minnesota
Bradley P. Carlin, University of Minnesota

In clustered survival settings where clusters correspond to geographic regions, biostatisticians are increasingly turning to models with spatially distributed random effects. These models begin with spatially-oriented frailty terms, but may also include further region-level terms in the parametrization of baseline hazards or various covariate effects (as in a spatially-varying coefficients model). In this paper, we propose a multivariate conditionally autoregressive (MCAR) model as a mixing distribution for these random effects, as a way of capturing correlation across both the regions and the elements of the random effect vector for any particular region. We then extend this model to permit analysis of temporal cohort effects, and show how our spatiotemporal model may be efficiently fit in a hierarchical Bayesian framework implemented using MCMC computational techniques. We illustrate our approach in the context of county-level breast cancer data from 22 annual cohorts of women living in the state of Iowa, as recorded by the Surveillance, Epidemiology, and End Results (SEER) database. Hierarchical model comparison using the DIC statistics, as well as maps of the fitted county-level effects, reveal the benefit of our approach in modeling spatiotemporal breast cancer survival.

e-mail: xiaopinj@biostat.umn.edu

A RESIDUAL INFORMATION CRITERION FOR THE GENERAL
LINEAR MIXED MODEL

Lloyd J Edwards * , University of North Carolina at
Chapel Hill; Pranab K. Sen, University of North
Carolina at Chapel Hill; Bahjat F. Qaqish, University
of North Carolina at Chapel Hill; Matthew J. Gurka,
University of North Carolina at Chapel Hill

Shi and Tsai (2002) derived a residual information criterion, RIC, based on the residual log-likelihood to be used as a selection criterion for regression models, including classical regression models, Box-Cox transformation models, weighted regression models and regression models with autoregressive moving average errors. The RIC was shown to be a consistent criterion that in most cases performed better in model selection than widely used criteria such as the AIC, BIC, -square, Mallows , and the final prediction error. With a focus on longitudinal data, we provide an extension of the RIC to the general linear mixed model. We show that the extension of the RIC to the general linear mixed model, denoted RIC-MM, is also a consistent criterion. We use simulation studies to compare the RIC-MM to the AIC and BIC, two widely used model selection criteria in linear mixed model analysis.

e-mail: Lloyd_Edwards@unc.edu

INFERENCE TECHNIQUES FOR DATA TRANSFORMATIONS IN LINEAR MIXED EFFECTS MODELS FOR
LONGITUDINAL DATA

Matthew J. Gurka* , University of North Carolina at Chapel Hill
Lloyd J. Edwards, University of North Carolina at Chapel Hill

The general linear mixed model is a powerful and flexible tool for representing continuous longitudinal data and is consequently becoming a standard procedure in many areas of biometric research. As is the case for univariate linear models with independently and identically distributed errors, it is quite common in practice to transform the response in order to meet the assumptions of the mixed model, namely normality of the error components. The methodology of such a model with a parametric transformation has been studied extensively in the univariate case and has started to receive attention in linear mixed models. However, inference techniques for the transformation parameter developed for the univariate model have not been extended to the mixed model. A methodological treatment of a test for the Box-Cox transformation in the general linear mixed model will be beneficial due to its use in practical applications. Various tests of the value of the transformation parameter are proposed with a focus on using the residual likelihood in estimating the parameters of the mixed model. Simulation studies allow assessment of the proposed inference tools.

e-mail: gurka@email.unc.edu

SYNTHESIZING CONDITIONAL AND MARGINAL INFERENCES IN COVARIATE-DEPENDENT NONLINEAR MIXED-EFFECTS MODELS

Zengri Wang *, 3M Company;
Thomas A. Louis, Johns Hopkins University

Marginal models and mixed-effects models are commonly used to model clustered data. When both marginal and conditional inferences are of interest, these models may not be adequate since they focus on either marginal or conditional inference, but not on both. Regression parameters in nonlinear mixed-effects models usually do not have an explicit marginal interpretation since the conditional functional shapes do not carry over to the marginal scale. When covariate-dependent random effects are present, regression inferences based on conditional models may be highly sensitive to random effects assumptions that are often difficult to verify in practice. We investigate an alternative parameterization of the generalized linear mixed models with a rescaling factor associated with a bridge distribution to accommodate both marginal and conditional inferences. The marginal mean of response distribution and conditional variance of random effects are modeled in an orthogonal fashion and multivariate or multilevel covariate-dependent random effects are modeled through the structured bridge distribution. The relationships between the new model and some existing models are discussed. Examples are given to illustrate the new approach.

e-mail: zwang@mmm.com

ROBUST R-ESTIMATION FOR UNBALANCED HETEROSCEDASTIC ONE-WAY RANDOM EFFECTS MODELS

Inkyung Jung*, University of North Carolina at Chapel Hill;
Pranab K. Sen, University of North Carolina at Chapel Hill

For estimating a common mean in the unbalanced heteroscedastic one-way random effects model, the weighted least square estimator (which is ML estimator under normality assumption) is often used. Two robust R-estimators derived from the signed rank statistics, assuming only symmetric but otherwise arbitrary continuous distributions for the random effect and random errors, are proposed here. These procedures eliminate the need of estimating other nuisance parameters, which is required for the weighted least square estimator. Besides robustness, proposed estimators can have asymptotic normality and better efficiency than other classical parametric competitors in various situations. Numerical evaluation for performance of proposed estimators from simulation study is also provided.

e-mail: ihnn@email.unc.edu

WHEN BLUP'S ARE BAD

Ed Stanek^{*}, University of Massachusetts – Amherst;
Julio M. Singer, University of Sao Paulo, Brazil;
George Reed, University of Massachusetts – Worcester;
Wenjun Li, Tufts-NEMC

When BLUPs are BAD Edward J. Stanek III Biostatistics and Epidemiology, UMass Amherst, MA USA Julio M. Singer Statistics, U of Sao Paulo, Brazil George Reed Prevention and Behavioral Medicine, UMass, Worcester, USA Wenjun Li Biostatistics Research Center, Tufts-NEMC, Boston, MA USA Best linear unbiased predictors (BLUPs) are commonly added to estimates of fixed effects to predict the expected response of sampled clusters, such as schools, clinics, or classrooms. The predictors arise from mixed models that represent sample clusters as random effects. We assume the clusters are of finite size, and are selected from a finite population. The ‘B’ in BLUP stands for minimum MSE, but we show in a simple example how in some settings, the simple cluster mean sample is better. In these contexts, the BLUP is bad. We characterize situations where BLUPs will have higher MSE. Finally, we contrast the usual mixed model BLUPs with predictors of random effects based on random permutation models, illustrating the superior performance of the latter.

e-mail: stanek@schoolph.umass.edu

DESIGN AND DATA ANALYSIS OPTIONS FOR COMPARISONS OF ASSISTED REPRODUCTIVE TECHNOLOGIES

James A. Hanley^{*}, McGill University
Nandini Dendukuri, McGill University
Robert Platt, McGill University
Marie-Helene Mayrand, McGill University

In the ‘alternating sequence design’ used to compare success rates with assisted reproductive technologies, women or couples are randomized to receive either the standard or experimental treatment in the first cycle, and — if they do not become pregnant— crossed between standard and experimental treatments after each successive cycle. Norman and Daya (Fertility and Sterility, 2000) have shown that, in the presence of heterogeneity of fertility, and an effective treatment, the overall efficacy of the experimental treatment is overestimated by this design. They advised that in order to achieve an accurate estimate of efficacy, the trial should be run for at least three cycles and all data from even-numbered cycles be excluded from the analysis, which should then be restricted only to odd-numbered cycles. We describe approaches that make use of the data from all cycles to produce estimates that are both less biased and more precise. The methods are generalizations of those applicable to the ‘constant sequence’ design, where naive methods that do not take account of the heterogeneity produce underestimates of treatment efficacy.

e-mail: James.Hanley@McGill.CA

MEASURING THE COMPLEXITY OF GENERALIZED LINEAR HIERARCHICAL MODELS

Haolan Lu*, University of Minnesota
 James S. Hodges, University of Minnesota
 Bradley P. Carlin, University of Minnesota

Hierarchical models usually feature multi-level structure: as the number of levels increases, so does the model's complexity and the difficulty in measuring it. The complexity of a hierarchical model is crucial for carrying out model diagnostics, selection and comparison. However, measuring hierarchical models' complexity remains unclear. This paper describes a measure of complexity for generalized linear hierarchical model based on linear model theory, and applies it to Poisson count data. We begin using simulation with a simple linear random-effects model to demonstrate the methodology. We then illustrate its application in a more complicated model by introducing correlation of the random-effects, specifically, spatial conditionally autoregressive (CAR) models, often used in disease mapping. Finally, we extend this measure of complexity to the case having additive spatial regional clustering effects and heterogeneity random effects. In the context of examples analyzing overall cancer mortality and breast cancer late detection risk in Minnesota counties, comparisons are made between our measure and a commonly used Bayesian measure of model complexity, p_D , the effective number of parameters (Spiegelhalter et al., 2002). Both measures are consistently close in all cases.

e-mail: haolanl@biostat.umn.edu

85. Markov Chain Models

HIDDEN MARKOV MODELS FOR MICROARRAY TIME COURSE DATA IN MULTIPLE BIOLOGICAL CONDITIONS

Ming Yuan*, University of Wisconsin at Madison;
 Christina M. Kendziorski, University of Wisconsin at Madison

Among the first microarray experiments were those measuring expression over time, and time course experiments remain common. Most methods to analyze time course data attempt to group genes sharing similar temporal profiles within a single biological condition. However, with time course data in multiple conditions, a main goal is to identify differential expression patterns over time. We present a Hidden Markov modeling approach designed specifically to address this question. Simulation studies show a substantial increase in sensitivity without an increase in the false discovery rate when compared to a marginal analysis at each time point. Results from three case studies are also discussed.

e-mail: yuanm@stat.wisc.edu

ESTIMATING SOJOURN TIME DISTRIBUTIONS WITH PERIODIC OBSERVATIONS UNDER A SEMI-MARKOV MODEL

Pai-Lien Chen^{*}, Family Health International
Hsiao-Chuan Tien, University of North Carolina at Chapel Hill

Semi-Markov models are useful to describe transitions in health or disease status in medical research. Estimating latent risk functions among disease statuses with data collected at periodic intervals can be difficult because of incomplete transition information. This study presents a pseudo likelihood approach to analyzing multistate data under a time homogeneous semi-Markov model when transition epochs are incomplete. We first estimate mean sojourn times for each observed time interval based on a competing risk model and its Markov process, and then use those estimated mean sojourn times as the length of transition epochs in observed likelihoods. We use an HIV cohort study to illustrate the method.

e-mail: pchen@fhi.org

MODELING THE NATURAL PROGRESSION OF A CHRONIC CONDITION FROM OBSERVATIONAL DATA WITH TIME-DEPENDENT COVARIATES

Yali Liu^{*}, Purdue University;
Bruce A. Craig, Purdue University

To assess the benefits of different treatment strategies, it is often necessary to estimate the natural progression of a disease or condition from observational data. Recently, Craig *et al.* (1999) proposed a discrete-time non-homogeneous Markov model for situations when the observations are at unequally spaced intervals and the time of intervention is interval censored. In this paper, we address the incorporation of time-dependent covariates, known only at the observation times, into this model and compare the performance of several different approaches. The model is then applied to the WESDR (Wisconsin Epidemiologic Study of Diabetic Retinopathy) data set.

e-mail: ylliu@stat.purdue.edu

BAYESIAN SEARCHING FOR MARKOV MODELS OF HIV MUTATION PROCESS

Jing Zhao*, University of Pennsylvania
Edward I. George, University of Pennsylvania Wharton School
Andrea S. Foulkes, University of Pennsylvania

A wide variety of mutations in the viral genome of Human Immunodeficiency Virus Type-1 (HIV-1) are associated with reduced susceptibility to Antiretroviral therapies (ARTs). Characterizing the process by which these mutations accumulate over time will have broad implications for clinical decision making and help further our knowledge of the biological mechanism of drug resistance. In this paper, we consider testing that the probabilities of mutating over time in response to treatment pressure are independent across sites on the viral genome. This presents an analytic challenge due to the high dimensionality of the viral genome (i.e. the large number of genetic loci and the complex, uncharacterized relationships among them). We begin by defining states based on the observed patterns of viral mutations and model the transitions between these states assuming a first order homogenous Markov process. We propose a Bayesian Model Selection framework that identifies the model with the highest posterior probability. This approach is well suited to a large model space. Markov Chain Monte Carlo (MCMC) simulations are implemented to identify the best model based on the highest visiting frequency and Bayes Factor calculation are presented to provide supporting evidence from the data for the selected model.

e-mail: jzhao@cceb.upenn.edu

FITTING A MIXED EFFECTS MARKOV MODEL FOR REPEATED
BINARY OUTCOMES WITH NONIGNORABLE DROPOUT

Robert J. Gallop *, West Chester University;
Thomas R. Ten Have, University of Pennsylvania

In many areas of research, repeated binary measures often represent a two-state stochastic process, where individuals can transition among two states. In a behavioral or physical disability setting, individuals can flow from susceptible or subthreshold state, to an infectious or symptomatic state, and back to a subthreshold state. Quite often the transition among the states happens in continuous time but is observed at discrete, irregularly spaced time points which may be unique to each individual. Methods for analyses are typically based on the Markov assumption. Cook (1999) introduced a Markov model that accommodates the subject to subject variation in the model parameters with random effects. We extend this model by adding a nonignorable dropout component to the model. Specification of the distribution of the random effects is made to guarantee a closed form expression of the marginal likelihood. This methodology is illustrated by applications to a data set from a parasitic field infection survey, a data set from a cocaine treatment study, and a data set from an aging study. Simulations suggest that the shared parameter model is robust with respect to at least one alternative non-ignorable model.

e-mail: rgallop@wcupa.edu

CHARACTERIZING THE PROGRESSION OF GENOMIC MUTATIONS IN HIV: A MARKOV PROCESS APPROACH

David Loecke*, Harvard School Of Public Health
Victor DeGruttola, Harvard School of Public Health

Because of the rates of genetic mutation and of viral replication observed in HIV infection, development of resistance to antiretroviral drugs is a common occurrence. Characterizing the progression of the HIV genomic mutations over time, especially along pathways which lead to high level drug resistance, is useful because this progression affects the drug options that will remain to a patient after failing a given treatment regimen. Following the work of Foulkes & DeGruttola (2002) we assume the viral genotypes can be characterized into a finite number of states which are defined by patterns of genomic mutations. We also consider a hidden Markov process to model the transition rates between states and extend their approach to incorporate covariates which are possibly time varying. Important covariates include the amount of diversity in the viral genome which may impact on the type of mutations that develop. We apply our methods to 150 patients who participated in two phase II clinical studies of efavirenz combination therapy. The methods were motivated by the availability of data on multiple viral clones derived from plasma samples at each patient visit. The sequences can be found in the Stanford HIV RT and Protease Sequence Database.

e-mail: dloecke@hsph.harvard.edu

86. Assessing Agreement Measurements: Current Issues and New Developments

A REPEATED MEASURES CONCORDANCE CORRELATION COEFFICIENT FOR CLUSTERED PAIRED DATA

Tonya S. King*, Penn State University;
Vernon M Chinchilli, Penn State University

The concordance correlation coefficient is commonly used to assess agreement between two raters or two methods of measuring a response when the data are measured on a continuous scale. However, the situation may arise in which repeated measurements are taken for each rater or method, e.g. longitudinal studies in clinical trials or bioassay data with subsamples. This paper proposes a coefficient for measuring agreement between two raters or two methods of measuring a response in the presence of repeated measurements. We illustrate the methodology with examples comparing (1) two assays for measuring serum cholesterol, (2) two measurements of percentage body fat, from skinfold calipers and dual energy x-ray absorptiometry, and (3) 1-hour vs. 2-hour blood draws for measuring cortisol in an asthma clinical trial.

e-mail: tking@psu.edu

ASSAY VALIDATION WITH CENSORED DATA

Huiman X. Barnhart*, Duke University
Jingli Song, Eli Lilly and Company
Robert Lyles, Emory University

In assay validation, it is important to assess agreement between two assays, say new and standard assays, based on different techniques. Oftentimes, one or both assays have lower limit(s) of detection and thus measurements are left censored. For example, in HIV (Human Immunodeficiency Virus) study, the branched DNA (bDNA) assay was developed to quantify HIV-1 RNA concentration in plasma. Validation of the bDNA assay involves assessing agreement of measurements by bDNA assay and by the standard assay, RT-PCR (reverse transcriptase - polymerase chain reaction). Both bDNA and RT-PCR assays have lower limits of detection and thus new statistical method is needed for assessing agreement between two left-censored variables. In this talk, we present a generalized estimating equations approach to evaluate agreement between two assays that are subject to lower limits of detection. The concordance correlation coefficient is used as an agreement index. The methodology is evaluated via simulation studies and illustrated by an example from HIV study.

e-mail: huiman.barnhart@duke.edu

ON THE ROBUSTNESS OF THE CONCORDANCE CORRELATION COEFFICIENT ESTIMATED BY VARIANCE COMPONENTS

Josep L. Carrasco *, Universitat de Barcelona

On The Robustness Of The Concordance Correlation Coefficient Estimated By Variance Components

The intraclass correlation coefficient and the concordance correlation coefficient are two of the most popular aggregate procedures used to measure agreement when data are on a continuous scale. It has been demonstrated that the concordance correlation coefficient is a certain intraclass correlation coefficient, specifically the intraclass correlation coefficient where the observers are considered as a fixed effect, but measuring agreement among observers is desired. Then, the observer effect is included in the intraclass correlation coefficient expression as a sum of squares rather than a variance. The difference between the concordance correlation coefficient and the intraclass correlation coefficient is found in the estimation method. The intraclass correlation coefficient is estimated by variance components of a normal mixed model instead of the common moment method used to estimate the concordance correlation coefficient. Since the variance components approach is a likelihood-based procedure, it can be argued that the moment method is more robust where the data is non-normal distributed. This hypothesis is assessed through a simulation study.

e-mail: carrasco@medicina.ub.es

MARGINAL ANALYSIS OF CLUSTERED DATA WHEN CLUSTER SIZE IS INFORMATIVE

John M. Williamson*, Centers for Disease Control and Prevention
Glen A. Satten, Centers for Disease Control and Prevention
Somnath Datta, University of Georgia

We propose a new approach to fitting marginal models to clustered data when cluster size is informative. This approach uses a generalized estimating equation (GEE) that is weighted inversely with the cluster size. We show that our approach is asymptotically equivalent to within-cluster resampling (Hoffman, Sen, and Weinberg, 2001, *Biometrika*), a computationally intensive approach in which replicate data sets containing a randomly selected observation from each cluster are analyzed and the resulting estimates averaged. Using simulated data and an example involving dental health, we show the superior performance of our approach compared with unweighted GEE, the equivalence of our approach with WCR for large sample sizes, and the superior performance of our approach compared with WCR when sample sizes are small.

e-mail: jow5@cdc.gov

MAKING INFERENCE BY PERTURBING A NON-SMOOTH ESTIMATING FUNCTION

Lu Tian*, Harvard University
L.J. Wei, Harvard University

We will discuss some recent developments on resampling methods for estimating equations. More specifically, we are going to introduce resampling methods based on bootstrap-like perturbations or asymptotical pivotal quantities and illustrate the methods using some examples. In addition, we will present a novel resampling method utilizing updated Bayesian computational techniques for estimating equations that are difficult to solve numerically.

e-mail: ltian@hsph.harvard.edu

INFERENCE BASED ON ESTIMATING FUNCTIONS WITH U STATISTIC STRUCTURE

Jack Kalbfleisch*, University of Michigan
Wenyu Jiang, University of Waterloo

Suppose that inference about parameters of interest are to be based on unbiased estimating functions that are U- statistics of degrees 1 and 2. We define suitable studentized versions of such estimating functions and consider asymptotic approximations as well as bootstrap methods based on resampling the estimated terms in the estimating function. These methods lead directly to confidence intervals derived from the asymptotic pivotal. Particular examples in this class of statistics include L_p estimation as well as Wilcoxon rank regression methods and other related estimation problems. We provide some numerical methods to carry out the inference procedures. These methods are compared in examples and simulations with a related suggestion for inference for U statistics based on resampling proposed by Jin, Ying and Wei (Biometrika, 2001).

e-mail: jdkalbfl@umich.edu

88. Statistical Genetics - Modeling Interaction and Multilocus Analyses

DEFINITION, DETECTION AND INTERPRETATION OF GENE-GENE INTERACTION (EPISTASIS) IN COMPLEX DISEASE

Heather J. Cordell * , University of Cambridge

Detection and interpretation of gene-gene interactions (epistasis) has been considerably confused in the literature by differing (usually unstated) definitions and assumptions and by the use of the same terminology to apply to rather different statistical and biological concepts. Classical concepts of epistasis as applied to mutually exclusive phenotype events are not easily generalized to apply to dichotomous outcomes such as presence or absence of disease, particularly for complex diseases where we expect to see reduced penetrance and the presence of phenocopies. In this presentation I will survey some of the most commonly used statistical and biological definitions of epistasis and point out some problems with and differences between them. The degree to which (various definitions of) epistasis can be detected using data such as that obtained from inbred mouse strains and human affected sib pairs will be discussed. Biological interpretation of the resulting tests and the degree to which such statistical models can elucidate underlying disease mechanisms will be considered.

e-mail: heather.cordell@cimr.cam.ac.uk

TESTING FOR LOCUS-LOCUS INTERACTIONS USING AFFECTED RELATIVE PAIRS: MULTIPOINT ANALYSIS AND INCOMPLETE IBD

Peter A. Holmans*, University of Wales

There has recently been much interest in the development of methods for incorporating covariates into linkage analyses of complex traits. Holmans (2002) extended the logistic regression method for affected sib pairs of Rice (1997) to enable tests of locus-locus interaction to be performed. It was found that allowing for interaction between loci could substantially increase linkage evidence in some situations. However, these analyses assumed that the locations of the disease loci were known, requiring only a single test. Furthermore, the IBD states of the sib pairs were also assumed to be known with certainty. In practice, wide lod-score peaks from single-locus analyses will usually be observed. This raises the question of whether to restrict analyses to the maxima of these peaks, or to search over a wider area. The power of various search strategies to detect interactions, or linkage allowing for interactions, is investigated under a variety of two-locus genetic models. In addition, the effect of incomplete IBD information is investigated by varying the density of the marker grids on the two chromosomes. It is found that denser marker grids are necessary to detect interactions than to detect linkage at either locus individually.

e-mail: holmanspa@cardiff.ac.uk

JOINT ANALYSES OF LINKED DISEASE GENES: LOCATION ESTIMATION AND HYPOTHESIS TESTING METHODS

Joanna M. Biernacka*, University of Toronto
Shelley B. Bull, University of Toronto
Samuel Lunenfeld, Research Institute, Toronto
Lei Sun, University of Toronto

For diseases with complex genetic etiology, more than one susceptibility gene may exist in a single chromosomal region. Under explicit assumptions about the number of disease genes in a region, general estimating equations can be used to estimate the putative disease gene locations and expected identical-by-descent allele sharing in affected sib pairs at these genes. We propose methods to evaluate the evidence for two versus one disease loci in a region in a quasi-likelihood framework. We formulated tests based on wald, modified quasi-score and approximate quasi-likelihood ratio test statistics. A number of issues arise in this testing problem that affect the null distributions of the test statistics, including an identifiability problem. Because of the difficulties in determining the asymptotic null distributions of these statistics and the small sample sizes generally available in genetic studies, we assess significance empirically by simulation. We evaluated the accuracy and efficiency of each of the tests by simulation, and found that the approximate quasi-likelihood ratio tests and our modified quasi-score test perform better than the wald test. Power to detect the presence of two linked disease genes increases with the number of affected sib pairs, greater IBD sharing at the two loci, and larger distance between the two loci.

e-mail: biernac@utstat.utoronto.ca

MULTILOCUS MODELS FOR LINKAGE ANALYSIS WITH COVARIATES

Jane M. Olson, Case Western Reserve University

Complex genetic diseases are generally believed to be multifactorial in etiology, that is, due to multiple genetic and environmental factors. To cope with this complexity in the context of genetic linkage analysis, several authors have suggested methods for considering jointly the contributions of two or more genetic loci. In parallel, several authors have proposed methods for including phenotypic covariates in linkage studies, primarily to allow for locus heterogeneity or gene-environment interaction. Here, we will discuss multilocus models that allow inclusion of phenotypic covariates. In principal, these methods allow the user to study gene-gene interactions while at the same time either 1) characterizing phenotypically the linked families, and/or 2) including gene-environment interactions. The methods will be illustrated with examples.

e-mail: olson@darwin.cwru.edu

89. Bayesian Computing and Analysis

LOSS FUNCTION BASED RANKING IN TWO-STAGES HIERARCHICAL MODELS

Rongheng Lin * , Johns Hopkins University;
 Thomas A. Louis, Johns Hopkins University,
 Susan Paddock, Rand Statistics Group;

Greg Ridgeway, Rand Statistics Group Health, U.S. Food and Drug Administration

Several authors have studied the performance of optimal, squared error loss (SEL) estimated ranks. Though these are effective, in many applications interest focuses on identifying the relatively good (e.g., in the upper 10%) or relatively poor performers. We construct loss functions that address this goal and evaluate candidate rank estimates, some of which optimize specific loss functions. We study performance for a fully parametric hierarchical model with a Gaussian prior and Gaussian sampling distributions, evaluating performance for several loss functions. Results show that though SEL-optimal ranks and percentiles do not specifically focus on classifying with respect to a percentile cut point, they perform very well over a broad range of loss functions. We compare inferences produced by the candidate estimates using data from The Community Tracking Study.

e-mail: rlin@jhsph.edu

BAYESIAN FACTOR ANALYSIS MODELS FOR MENTAL DISORDERS: A CASE STUDY OF THE WHITELEY INDEX FOR HYPOCHONDRIASIS

David M. Shera*, The Children's Hospital of Philadelphia / University of Pennsylvania
Joseph G. Ibrahim, University of North Carolina

Hypochondriasis is a mental disorder where the patient believes he or she is physically ill when no disease is present, or at least no known disease can be identified. This article presents a case study of recent Whiteley Index data, which comes as two cohorts, combined with results from a much earlier study. The self-rated instrument has fourteen items and each outcome is from a discrete ordinal (Likert) scale from 1 to 5. Factor analysis models for ordinal outcomes are considered and informative prior elicitation techniques are demonstrated. Markov chain Monte Carlo (MCMC) methods are developed for sampling from the posterior distribution of the parameters of interest. The results are consistent with theory and provide evidence for the validity of the three hypothesized latent constructs.

e-mail: shera@email.chop.edu

BENEFITS OF BENCHMARK ESTIMATION

Steven N. MacEachern*, Ohio State University;
Subharup Guha, Ohio State University

Markov chain Monte Carlo simulation, applied to a Bayesian problem, yields a set of parameter values drawn from the posterior distribution. The most common implementation is to store a systematic subsample of the generated parameter values, and to use the stored points to estimate a variety of features of the posterior distribution. In passing from the full sample to the subsample, a considerable amount of information is discarded. Benchmark estimation (Guha, MacEachern and Peruggia, 2002) provides a means of improving simulation by recovering the discarded information or by incorporating information external to the simulation. The additional information may consist of better estimates (based on the full sample) or known values (in a Bayesian context, features of the posterior which may be evaluated analytically). The technique is easy to implement. A range of improvements can be obtained. In this talk, we illustrate a variety of improvements and discuss the issue of how to select the strata and benchmarks most effectively. We find that sound implementation of the technique can improve its performance markedly.

e-mail: browmac@sbcglobal.com

BAYESIAN MACHINE LEARNING FOR LARGE P SMALL N REGRESSION

Sounak Chakraborty*, University of Florida
Malay Ghosh, University of Florida
Bani K. Mallick, Texas A&M University

Statistical machine learning plays a very important role in different areas of science. It learns a task from a set of data called a ‘Training Set’, and applies to a data set called ‘Test Set’ for prediction. The Support Vector machine (SVM) has recently been introduced as a new technique for solving a variety of learning and function estimation problems. It expresses predictions in terms of a linear combination of kernel functions. In our paper, we introduce a full Bayesian support vector regression model with Vapnik’s ϵ -insensitive loss function, based on reproducing kernel Hilbert spaces (RKHS). This provides a full probabilistic description of SVM. We have also considered Relevance Vector Machine (RVM) introduced by Bishop, Tipping and others. Instead of the original treatment of the RVM relying on the use of type II maximum likelihood estimates of the hyper-parameters, we put a prior on the hyper-parameters and use Markov chain Monte Carlo technique for computation. We apply our model for prediction of blood glucose concentration in diabetics using florescence based optics. We also introduce full Bayesian support vector regression (SVR) and relevance vector regression (RVR) models when the response is multivariate. The multivariate version of the SVM and RVM is illustrated with a prediction problem in near-infrared (NIR) spectroscopy.

e-mail: schakrab@stat.ufl.edu

EVALUATION OF BAYESIAN MODEL DIAGNOSTIC TECHNIQUES

Guofen Yan *, Cleveland Clinic Foundation;
J. Sedransk, Case Western Reserve University and Energy Information Administration

While a number of Bayesian model diagnostic techniques have been proposed over the past twenty years, there has been little evaluation of their properties. Of interest is whether these techniques can identify a model failure when a single model is proposed and the nature of the departure from this model is not clearly known. We study the relatively simple, yet important, situation where the true data model is two-stage hierarchical while the fitted model doesn’t have any hierarchical structure. We develop a method to investigate properties of several different techniques, each of these techniques being based on posterior predictive assessment. Exact and asymptotic analytical expressions and the results from simulations are used to evaluate the techniques.

e-mail: gyan@bio.ri.ccf.org

COMBINING PHYLOGENETIC INFORMATION AND ENVIRONMENTAL FACTORS IN HIERARCHICAL MODELING TO EXPLAIN SPECIES DIVERSITY

Shanshan Wu*, University of Connecticut
 Paul O. Lewis, University of Connecticut
 Mark Holder, University of Connecticut
 John A. Silander, University of Connecticut
 Andrew Latimer, University of Connecticut
 Anthony G. Rebelo, National Botanic Institute, Kirstenbosch, South Africa
 Alan E. Gelfand, Duke University

A common result of speciation is that sister species tend to occupy separate but contiguous geographic regions but that this effect will diminish with time. However, sister species may have similar ecological characteristics and species attributes and models based solely on ecology may predict species to occur in areas occupied by its sister. With an interest in understanding and predicting species distribution, we propose the first integration of spatial environmental factors with phylogenetic tree data. The challenge has been to construct a model in which genetic history places some constraints (which decay with time) on the predicted occurrence of species relative to the sister species. In particular, following our earlier work (Gelfand et al, 2003), we add phylogenetic terms which constitute latent variables denoting positive/negative speciation effect, ancestor presence/absence, and magnitude as a function of relative amount of time since speciation. Ancestor presence/absence can be treated as known by inserting suitable estimates or can be treated as unknown. Relative amount of time since speciation can be independently determined through estimating phylogenies using DNA sequence data. We also consider spatial structure on prior distributions for speciation effects and ancestral effects.

e-mail: shanshan@stat.uconn.edu

90. Nonparametric Methods

NONPARAMETRIC CONFIDENCE SETS FOR DENSITIES AND CLUSTERS

Woncheol Jang*, Carnegie Mellon University;
 Christopher Genovese, Carnegie Mellon University;
 Larry Wasserman, Carnegie Mellon University

We present a method for constructing uniform asymptotic confidence sets for densities and clusters that is a density estimation version of Beran's (2000) REACT (Risk Estimation and Adaptation after Coordinate Transformation). Specifically, we derive the asymptotic distribution of the pivot process, $B_p(\lambda) \equiv \sqrt{n}(L_p(\lambda) - S_p(\lambda))/\tau(\lambda)$ where $L_p(\lambda)$ and $S_p(\lambda)$ are the loss function and the estimated risk function with smoothing parameter λ . Inverting the pivot provides confidence balls for the coefficients of orthogonal series estimators which one can extend to confidence sets for functionals of densities. The original regression version of the REACT method uses the classical empirical process theory to show stochastic convergence of the pivot process. In the density estimation case, the pivot process is a sum of dependent processes. We develop a modified version of the functional central limit theorem to deal with this dependence. While most of literature for dependent cases focused on mixing cases at the price of the entropy condition, we use specific features of the dependency structure to show asymptotic independence.

e-mail: wjang@stat.cmu.edu

ON NONPARAMETRIC REGRESSION FOR CURRENT STATUS DATA

Deborah Burr*, The Ohio State University
Shanti Gomatam, U.S. Food and Drug Administration

We study nonparametric estimation of the conditional distribution function when we have current status data on the outcome variable and a single continuous-valued covariate. An estimator of the conditional distribution function, called the local nonparametric maximum likelihood estimator (LNPMLE), is proposed. The asymptotic distribution of the LNPMLE of the conditional distribution function at a point is used to show that the asymptotically optimal bandwidth is of the order $N^{(-1/7)}$, and an expression for the optimal bandwidth is obtained. A plug-in estimate of the bandwidth is suggested, and its computation is illustrated on a simulated sample. We illustrate our methodology on a dataset in which the outcome of interest is time to infection with HIV.

e-mail: burr@stat.ohio-state.edu

RANK-SUM TESTS FOR CLUSTERED DATA

Somnath Datta *, University of Georgia
Glen A. Satten, Centers for Disease Control and Prevention

The Wilcoxon rank-sum test is widely used to test the equality of two populations because it makes fewer distributional assumptions than parametric procedures such as the t-test. However, the Wilcoxon rank-sum test can only be used if data are independent. When data are clustered, tests based on generalized estimating equations (GEEs) that generalize the t-test have been proposed. Here we develop a rank-sum test that can be used when data are clustered. As an application, we develop a non-parametric version of a family-based test of association between a genetic marker and a quantitative trait locus. We also give a rank-sum test for equivalence of three or more populations that generalizes the Kruskal-Wallis test to situations with clustered data.

e-mail: datta@stat.uga.edu

NONPARAMETRIC ESTIMATION OF AN INTENSITY FUNCTION AND ITS BREAKPOINT WITH APPLICATIONS TO SPIKE TRAIN DATA IN NEUROSCIENCE

Matt Gregas*, University of Minnesota

Motivated by a neuroscience experiment which observes spike trains from the primary motor cortex of Macaca Mulatta (rhesus monkey), we use local likelihood methods for estimating the intensity function of a non-homogeneous Poisson point process. Asymptotic distributions of our estimates are presented. Additionally, we develop tests for the presence of breakpoints in the intensity function at a given location. Critical values for the tests are based on the asymptotic distribution of the test statistics. Power and actual significance level of the test are described through a simulation study. The relationship between the power of the tests, the magnitude of the jump, the bandwidth, and the sample size are examined.

e-mail: matt@stat.umn.edu

LOCAL ISOTONIC REGRESSION

Derick R. Peterson *, University of Rochester

A new Local Isotonic (LOIS) regression estimator is proposed for the standard nonparametric regression problem. The idea is to approximate the regression function at each point by a local isotonic regression rather than, say, a local polynomial one. Since this local model is clearly more flexible than that of local linear regression, itself a locally monotone model, the LOIS regression estimator has smaller bias at most points. Moreover, in stark contrast to local polynomial regression, LOIS regression is very insensitive to the choice of bandwidth unless the regression function oscillates frequently. And if the data themselves are monotone then, for any bandwidth, LOIS reproduces the data, is thus conditionally unbiased at each observation, and is equivalent to the nonparametric maximum likelihood estimator for monotone regression functions. Since the price to be paid for the added flexibility of the local model is increased variability, and asymptotically this price is a steep one, the performance of LOIS for practical sample sizes is investigated via a simulation study.

e-mail: peterson@bst.rochester.edu

A PENALIZED NONPARAMETRIC MAXIMUM LIKELIHOOD APPROACH TO SPECIES RICHNESS ESTIMATION

Ji-Ping Z. Wang*, Northwestern University
Bruce G. Lindsay, Penn State University

We propose a class of penalized nonparametric maximum likelihood estimators (NPMLE) for the species richness problem. Our motivation for using a penalty term on the likelihood is that we have found that the estimators that lack it have an extreme instability problem. The estimators are constructed using a conditional likelihood that is simpler than the full likelihood. However, we show that the full likelihood NPMLE solution given by \cite{stat:NorrisPollock1998} can be found (with great accuracy) by using an appropriate penalty term on the conditional likelihood. A simple and fast algorithm for the penalized NPMLE is developed and shown to greatly speed up computation of the unconditional NPMLE. Based on our goals of attaining high stability while retaining sensitivity, we propose an adaptive quadratic penalty function. A systematic simulation study, using a wide range of scenarios, establishes the success of this method relative to its competitors.

e-mail: jzwang@northwestern.edu

NONPARAMETRIC VARIABLE SELECTION VIA COMPONENT SHRINKAGE

Yi Lin, University of Wisconsin, Madison;
Hao H. Zhang*, North Carolina State University

We propose a new method for model selection and model fitting in nonparametric regression models, in the framework of smoothing spline ANOVA. The “COSSO” is a method of regularization with the penalty functional being the sum of component norms, instead of the squared norm employed in the traditional smoothing spline method. The COSSO provides a unified framework for several recent proposals for model selection in linear models and smoothing spline ANOVA models. Theoretical properties, such as the existence and the rate of convergence of the COSSO estimator, are studied. In the special case of a tensor product design with periodic functions, a detailed analysis reveals that the COSSO applies a novel soft thresholding type operation to the function components and selects the correct model structure with probability tending to one. We give an equivalent formulation of the COSSO estimator which leads naturally to an iterative algorithm. A novel one-step update algorithm is proposed, which is computationally fast and efficient. Simulations and real examples show that the COSSO gives very competitive performances when compared with other methods.

e-mail: hzhang2@stat.ncsu.edu

A WEIGHTED ESTIMATED LIKELIHOOD METHOD FOR TWO-STAGE OUTCOME-DEPENDENT STUDY USING KERNEL SMOOTHER

Xiaofei Wang*, Duke University
Haibo Zhou, University of North Carolina at Chapel Hill

In two-stage epidemiological studies, disease outcome and some covariates are available for each member in the study population at the first stage, while the primary exposure variable is only observed on a selected subsample of the population at the second stage (e.g, Breslow and Cain, 1988). This work considers a two-stage outcome-dependent design in which the selection of the second-stage subsample depends on both the disease outcome and a continuous auxiliary covariate of the true exposure variable. Assuming randomly selected validation data is available for a mismeasured exposure variable, Pepe and Fleming (1991) and Carroll and Wand (1991) proposed an estimated likelihood method using the validation data to approximate empirically the unknown part of the likelihood. Due to the outcome-dependent nature of the second-stage subsample, however, the Pepe and Fleming estimator and the Carroll and Wand estimator yield inconsistent estimates in our setting. We propose a weighted kernel-based estimator to handle both the outcome- dependent nature of the subsampling component and the continuity of the auxiliary covariate. Simulation results show that the proposed estimator for a continuous auxiliary covariate performs better than the competing methods. We illustrate the proposed method with data example.

e-mail: xiaofei.wang@duke.edu

A LOGISTIC REGRESSION MODEL WITH MISCLASSIFICATION IN THE OUTCOME AND CATEGORICAL COVARIATE

Michelle A. Detry*, The University of Texas;
Wenyaw Chan, The University of Texas

Misclassification of binary data may frequently occur in public health research. For example, a study participant may be asked to report whether or not they have quit smoking, and he or she may respond that they do not smoke when (in fact) they do. In order to avoid the effects of such biases on study conclusions, misclassification should be incorporated in the model when analyzing the data. Previous research has been performed to develop methods to adjust for misclassification in a dichotomous outcome variable or misclassification in predictor variables when using logistic regression. The present research developed a model to account for misclassification in both the dichotomous dependent variable and an independent predictor variable when using logistic regression. The model performed well in simulation studies with parameter estimates similar to the theoretical values. The results show that including a continuous predictor variable (assumed to have no measurement error) hampers the ability to estimate misclassification in the dichotomous predictor variable. The misclassification model was also applied to a case study examining the effects of maternal smoking on low birth weight babies.

e-mail: mdetry@sph.uth.tmc.edu

IMPACT OF MEASUREMENT ACCURACY ON THE SAMPLE SIZE ESTIMATION USING PILOT PAIN ASSESSMENT: A SIMULATION STUDY

Yen-Hong Kuo*, Jersey Shore University Medical Center
Yen-Liang Kuo, Kinmen County Hospital

Visual analog scale is widely used for pain assessment in surgical clinical trials. The pain intensity is measured and reported as an integer in the unit of either centimeter or millimeter. Even the same intensity is more accurately assessed when it reaches millimeter, a value recorded in the unit of centimeter is commonly seen in medical literature. In order to estimate the sample size for conducting a clinical trial in which pain intensity is the main outcome, data from pilot studies or published results are used. How the levels of measurement accuracy on pain assessment in the pilot studies influence the sample size estimation is of interest. A simulation study was performed to assess the effects based on two groups of patients in a clinical trial. This presentation will compare the influences on the sample means and standard deviations from the levels of measurement accuracy and the sample sizes in a pilot study. The focus will be the consequent impacts on the sample size estimation.

e-mail: yhkuo@jhu.edu

MEASUREMENT ERROR IN TIME SERIES MODELS

John W. Staudenmayer*, University of Massachusetts
at Amherst; John P. Buonaccorsi, University of
Massachusetts at Amherst

Motivated by common experimental designs and models in ecology, we consider the problem of a time series observed with measurement error. Focusing on autoregressive models and additive measurement error, we derive the biases caused by ignoring measurement error, develop some simple new methods to correct for the effects of measurement error, and compare the new methods to existing methods both analytically and with simulation. The new methods prove to be much more effective than existing methods.

e-mail: jstauden@math.umass.edu

CLOSED FORM SEMIPARAMETRIC ESTIMATORS FOR MEASUREMENT ERROR MODELS

Yanyuan Ma*, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

We examine the locally efficient semiparametric estimator proposed by Tsiatis and Ma (2003) in the situation when a sufficient and complete statistic exists. We derive a closed form solution and show that when implemented to generalized linear models with normal measurement error, this estimator is equivalent to the efficient score estimator in Stefanski and Carroll (1987). We also demonstrate how other consistent semiparametric estimators naturally emerge.

e-mail: yma@unity.ncsu.edu

COVARIATE MEASUREMENT ERROR IN DUAL SYSTEMS MODELS

Gaitri Gunasekara * ; Michigan Technologies
University; Thomas D. Drummer, Michigan
Technological University

Double counts (capture-recapture or dual systems) estimators of abundance, can incorporate covariates to accommodate capture heterogeneity. Typically, the covariates, assumed to be measured without error, are used in logistic regression equations to model the probability of detection. The Horvitz-Thompson estimator can then be used to estimate total abundance. We consider here the case of two independent observers conducting surveys in Yellowstone National Park to obtain estimates of elk abundance, using group size as a covariate, where group size is measured with error. When both observers sight a group independently, there may be substantial variation in recorded group size. The measurement error impacts not only the estimates of logistic regression coefficients used to estimate probability of detection, but the group size used in the numerator of the Horvitz-Thomson estimator as well. This paper investigates the impact on the abundance estimate when measurement error is ignored, and suggests remedies for the occurrence of both additive and multiplicative measurement error. Key words: Abundance estimate, covariate measurement error, dual systems method, logistic regression.

e-mail: gpgunase@mtu.edu

A MODEL FOR A LATENT FACTOR PREDICTING A CATEGORICAL LATENT OUTCOME WITH APPLICATION TO EATING DISORDERS RISK AND BODY SATISFACTION

Jia Guo*, University of Minnesota
Melanie M. Wall, University of Minnesota

In the research of public health, psychology, and social sciences, many research questions have been proposed to investigate the relationship between a categorical outcome and a continuous predictor variable. The focus of this paper is to develop a model to build this relationship when both the categorical outcome and the predictor variable are latent (i.e. not observable directly). In this paper, we propose and compare four methods to estimate the parameters in the model. A simulation study will be performed to evaluate the different methods in terms of bias and efficiency. A data example involving adolescent health will be used for demonstration, where the latent classes of eating disorders risk are predicted by the latent factor body satisfaction.

e-mail: jjguo@biostat.umn.edu

92. Longitudinal Data II

A JOINT LATENT AUTOREGRESSIVE MODEL FOR LONGITUDINAL ORDINAL DATA AND PATIENT DROPOUT WITH INFORMATIVE MISSINGNESS

George A. Capuano*, University of North Carolina at Chapel Hill;
Pranab K. Sen, North Carolina University at Chapel Hill

In many clinical trials where the endpoint is a time to event, longitudinal measurements are captured on variables that are also thought to influence the underlying data process. In order to better understand the interrelationship between the data processes, a common approach is to posit a joint model for the longitudinal measurements and the survival endpoint where the components are linked by a random or latent effect that represents an underlying propensity for failure or dropout. In many instances, patients under study in different treatment groups become more homogenous in time thereby making the proportional hazards assumption impractical. A joint longitudinal ordinal categorical data model and proportional odds model for dropout that incorporates a model for informative missingness is presented. Model components are linked by a common Gaussian latent autoregressive effect representing a propensity for patient dropout. An automated Monte Carlo EM algorithm extended from the generalized linear model framework is applied to yield exact maximum likelihood estimates while assessing the error from Monte Carlo sampling at each iteration. One step estimators are also constructed for comparing the efficiency against their fully iterated counterparts.

e-mail: georgecapuano@aol.com

CHOOSING THE OPTIMAL POCT DEVICE VIA A LONGITUDINAL STUDY

Jason T. Connor*, Carnegie Mellon University
Kenneth M. Shermock, The Johns Hopkins Hospital
Emily Stewart, Brigham Young University
Jodie Fink, Cleveland Clinic Foundation
Lee Bragg, Kaiser Permanente

Five FDA approved point-of-care (POC) fingerstick devices used to monitor warfarin patients' INR values are compared via a longitudinal study. Trial Design: 350 patients were randomized to 1 of 5 POC devices. Patients' INRs were monitored for an average of 87 days each. All warfarin dosing decisions were made by the treating clinician according to fingerstick readings from the POC device. Simultaneous readings were taken from venous arm draws sent to a reference laboratory which served as the gold standard. These reference laboratory measurements were used to determine the fraction of time each patient's INR values were within their therapeutic range (2.0 - 3.0 in 85% of study patients). Statistical Methods: A Bayesian beta hierarchical model with a smoothing component for the longitudinal blood draws was used to estimate the mean proportion of time each device kept patients within their therapeutic range. Results: This longitudinal study offers an improvement of simple measures of numeric agreement as it provides information on how numerical discrepancies between the POC and gold standard lead to compounding errors during long- term management of patients. Two POC devices, Coaguchek S (59.5% in range) and Coaguchek ProDM (55.5%) prove to be superior to their three competitors, Hemochron (50.5%), Protime (48.5%), and Rapidpoint 43.3%).

e-mail: jconnor@stat.cmu.edu

BAYESIAN ANALYSIS OF MIXED-EFFECTS TIME VARYING-COEFFICIENT MODELS FOR LONGITUDINAL DATA

Yi-Liang Tung * , National Taiwan University, Taiwan, R.O.C

Nonparametric regression modeling for longitudinal data has received great attention in recent years. Most of the statistical approaches are from frequentists point of view. For instance, Lin and Carroll (2001) proposed the semiparametric marginal models, and Wu et al. (1998) proposed the varying-coefficient models. The common features of these models are non-likelihood-based, marginal approach, and asymptotic base. It is still an unanswered question to compare the semiparametric models with other alternative parametric models. To overcome this problem, I proposed a general class of models, mixed-effects time varying-coefficient models, which extend the generalized linear mixed models by Breslow and Clayton (1993). The time- varying covariate-effects, heterogeneity across experiment units, and time series covariance structure can now be considered under this framework. I also develop a modified Gibbs sampler to make inference of the unknown quantities. This class of models can be extended to incorporate semiparametric Bayesian method under the Dirichlet process mixture settings. Formal model comparison procedure via Bayes factor is also provided. Finally, the MACS dataset (Kaslow et al. 1987) is used for illustration. Keywords: Bayes factor; Dirichlet process mixture; Heterogeneity.

e-mail: tongil@ms47.hinet.net

MULTIVARIATE TRANSITION MODELS FOR LONGITUDINAL BINARY DATA

Leilei Zeng*, University of Waterloo
Richard R.J. Cook, University of Waterloo

In many settings with longitudinal binary data interest lies in modeling covariate effects on transition probabilities (e.g. Muenz and Rubenstein, 1985). When interest lies in tracking how two processes change together, one may examine the degree to which changes in one process are correlated with changes in another process. In such settings the associations between the transition occurrences for the two processes is the scientific focus. Under a Markov assumption, use of marginal transition models permits separate modeling of covariate effects on the transition probabilities for univariate longitudinal binary data, but no insight into the associations can be obtained. Time-dependent covariates may be constructed for one process based on another, but the two processes are then treated asymmetrically. We propose a method of estimation and inference for joint transitional models for multivariate longitudinal binary data based on GEE2 (Zhao and Prentice, 1990) or alternating logistic regression (Carey et al., 1993). This approach enables one to model covariate effects on marginal transition probabilities as well as on the association parameters between the two processes. Such models are helpful for understanding how two or more processes change together while providing insight into covariate effects on features of the marginal Markov processes. We show that this approach results in consistent estimates of regression coefficients and association parameters. Moreover, we show through simulation that estimates from the joint transition models can lead to more efficient estimates of regression coefficients when the association between processes is strong. The method is illustrated by application to a longitudinal study in radiology. Extensions to deal with multivariate longitudinal categorical data are indicated.

e-mail: lzens@math.uwaterloo.ca

THE EFFICIENT ESTIMATION OF NONLINEAR EFFECTS IN BOTH CROSS-SECTIONAL AND LONGITUDINAL SEM

Andreas G. Klein*, University of Illinois at Urbana-Champaign;
Bengt O. Muthen, University of
California Los Angeles

This paper deals with an efficient methodology for the estimation of nonlinear SEM, which provides the analysis of nonlinear effects for a broad range of cross-sectional and longitudinal SEM. The proposed method provides an approximate ML estimator and outperforms the currently available methodology (covariance structure analysis, two-stage least squares) with respect to statistical power and flexibility. Parallel to the estimation method, a new general latent variable modeling framework is presented which incorporates the modeling of multiple nonlinear effects in both cross-sectional and longitudinal models. For the case of cross-sectional models, this framework includes models with multiple interaction effects among latent exogenous variables. For the case of longitudinal models, the applicability of the framework is illustrated by a heterogeneous growth curve model. Formally, heterogeneity of growth refers to the fact that some subgroups of individuals grow more consistently than others, and to model this heterogeneity statistically correct is essential for optimal prediction of individuals' development. The new methodology enables the researcher to analyze more efficiently what heterogeneity of growth could depend on and provides a tool for an improved prediction of individual growth, based on initial status or other covariate information.

e-mail: agklein@uiuc.edu

SAMPLE SIZE CALCULATION FOR STUDIES WITH CORRELATED ORDINAL OUTCOMES

Hae-Young Kim*, Centers for Disease Control and Prevention
John M. Williamson, Centers for Disease Control and Prevention
Cynthia M. Lyles, Centers for Disease Control and Prevention

Correlated ordinal response data often arise in public health studies. Sample size (power) calculations are a crucial step in designing such studies to ensure an adequate sample to detect a significant effect when it truly exists. Rochon (1998, *Statistics in Medicine* 17, 1643–1658) proposed sample size calculations for binary, count, and continuous responses by using GEE. Here we derive the minimum sample size for the correlated ordinal response data. We assume that study data will be analyzed with the approach of Lipsitz et al. (1994, *Statistics In Medicine* 13, 1149–1163) and use the non-central version of the Wald test statistic described in Rochon to calculate sample size for the correlated ordinal case. Analysis of an arthritis clinical trial is used for illustration.

e-mail: hykim730115@yahoo.com

93. Genetic Maps, Linkage Disequilibrium and Haplotype Analysis

A CLOSER LOOK AT HUMAN RECOMBINATION

John Barnard * , Cleveland Clinic Foundation

Recombination, the mechanism of DNA mixing from one generation to the next, is a major component of the forces that drive human evolution. A recent study of ours (Kong et al 2002) demonstrates a substantially higher correlation between recombination rates and sequence content (e.g. GC content, CpG motifs, and polyA/polyT stretches) than reported before. Most interestingly, we found that not only do mothers have variable recombination rates ('mother effect'), even gametes of the same mother have systematic difference in recombination rates ('gamete effect'). There is a positive correlation of recombination counts across chromosomes suggesting there is some underlying factor that affects recombination rates across the genome. Here we will report results that quantify the substantial contributions of the mother effect and the maternal gamete effect to the total variation of recombination counts. We will also take a close look at whether the underlying factor has an uniform effect on the whole genome or whether the effect is more pronounced on certain chromosomes or certain parts of the chromosomes. Finally, we will briefly review the novel methodological and computational tools employed in our analysis to handle the issues of incomplete data and crossover interference.

e-mail: jbarnard@bio.ri.ccf.org

ON THE USE OF PRIOR INFORMATION TO DEVELOP GENETIC MAPS

Chris A. Andrews*, Oberlin College

A dense, highly accurate genetic framework map is key to successful linkage analysis. Genetic components of numerous human diseases have been identified due to human maps. Similar maps have been constructed from scratch for other organisms to facilitate our understanding of them and us. The methodology presented in this talk can improve the process by which framework maps are constructed. Information available about current framework maps should be used to speed the construction of new maps. It should come as no surprise that the genomes of primates are very similar to each other, including the human genome. A Bayesian approach is employed to produce a baboon map and could be applied to the current effort to map the rhesus genome. The resulting baboon map has more markers than an existing map and is produced with fewer data.

e-mail: chris.andrews@oberlin.edu

A UNIFORM MEASURE OF LINKAGE DISEQUILIBRIUM FOR MARKERS WITH TWO OR MORE ALLELES

Chun Li*, Center for Human Genetics Research,
Vanderbilt University

Linkage disequilibrium (LD) indicates non-independence between two loci. It is an important concept in population genetics and has recently been used extensively in fine mapping genes. A few statistics are available to measure the level of LD between two polymorphisms, including D , D' , δ , etc. However, most statistics are defined only for bi-allelic markers and cannot be easily extended to markers with >2 alleles. To measure LD for markers with >2 alleles, a common approach is first to collapse all but one allele for each marker and calculate the allele-wise LD for the resulting 2×2 tables, then to calculate a weighted average of all the allele-wise LD statistics. The value of such a weighted average tends to be small when the number of alleles increases. We propose a new, uniform measure of LD between two polymorphisms, which is defined directly on markers, no matter how many alleles the markers have. The new measure is essentially a Kullback-Leibler distance from the true joint distribution of two markers to the joint distribution assuming independence, but further standardized so that it ranges from 0 to 1. Unlike D' , the new measure takes value 0 only when the two markers are in perfect LD.

e-mail: chun.li@vanderbilt.edu

ESTIMATING THE HAPLOTYPE SPECIFIC DISEASE RISK IN A CASE-CONTROL STUDY

Nandita Mitra*, Memorial Sloan-Kettering Cancer Center
E.S. Venkatraman, Memorial Sloan-Kettering Cancer Center
Colin B. Begg, Memorial Sloan-Kettering Cancer Center

Estimation of the association between haplotypes and disease from a case-control study is considered. Assuming a single 'disease haplotype' leads to increased risk, attention focusses on the relative risks associated with a single copy or two copies of the disease haplotype, relative to individuals with no copies. In this setting, case frequencies of the haplotype pairs are in Hardy-Weinberg Equilibrium (HWE) only if the influence of two copies of the disease haplotype on risk is multiplicative. Thus, imputation cannot rely on the assumption of HWE for cases. A method is presented for obtaining unbiased estimates of relative risks, making use of the EM algorithm and the assumption of HWE only for controls. The method accounts for the additional variation in the estimates due to the imputation of expected frequencies of haplotype pairs from ambiguous genotypes. A simulation study shows that the resulting confidence intervals have nominal coverage, and that the methods based on the assumption of HWE for both cases and controls can lead to bias.

e-mail: mitran@mskcc.org

PRUNING THE RECONSTRUCTED HAPLOTYPE FREQUENCIES IN ASSOCIATION ANALYSIS

Jung-Ying Tzeng*, North Carolina State University

Many test statistics of association in haplotype analysis use the information of haplotype frequencies. Two major issues encountered in such analysis are missing phase information and high dimensionality of haplotype distribution. As several promising algorithms are being proposed to provide better estimates of haplotype frequencies from genotypes, unreliable estimates still tend to emerge on haplotypes with very low frequencies. These sparse categories also result in a higher dimension of haplotype distribution, which leads to inefficiency in the haplotype analysis. To increase the power and efficiency of haplotype analysis, we continue to explore the idea of dimension reduction mentioned in Tzeng et al. (AJHG 2003), which proposes pruning the haplotype distribution by clustering the rare haplotypes. Here we present a more refined approach to estimate the pruned haplotype frequencies (PHF) from genotypes. In this approach, we set the parameter space of PHF as common haplotypes and unambiguous haplotypes to remove only the unreliable rare categories. We then relate the rare haplotypes to the preserved haplotypes through mutations, and obtain likelihood estimates of PHF from genotypes. Finally through simulation, we evaluate the performance of association tests based on the pruned version and the original version of the reconstructed haplotype frequencies.

e-mail: jytzeng@stat.ncsu.edu

INTEGRATION OF ASSOCIATION STATISTICS OVER GENOMIC REGIONS USING BAYESIAN ADAPTIVE REGRESSION SPLINES

Xiaohua Zhang*, Merck Research Laboratories
Kathryn Roeder, Carnegie Mellon University
Garrick Wallstrom, Carnegie Mellon University
Bernie Devlin, University of Pittsburgh School of Medicine

In the search for genetic determinants of complex disease, two approaches to association analysis are most often employed, testing single loci or testing a small group of loci jointly via haplotypes. The former has the advantage of simplicity but suffers severely when alleles at the tested loci are not in linkage disequilibrium (LD) with liability alleles; the latter should capture more of the signal encoded in LD, but is far from simple. The complexity of haplotype analysis could be especially troublesome for association scans over large genomic regions. For these reasons, we have been evaluating statistical methods that bridge the gap between the two approaches. We present one such method, which uses non-parametric regression techniques embodied by Bayesian Adaptive Regression Splines (BARS). For a set of markers falling within a common genomic region and a corresponding set of single locus association statistics, the BARS procedure integrates these results into a single test by examining the class of smooth curves consistent with the data. The BARS procedure provides a robust and potentially powerful alternative to classical tests of association, diminishes the multiple testing problem inherent in those tests, and can be applied to a wide range of data types, including genotype frequencies estimated from pooled samples in clinical trials, which is favored in the application of genomics to drug discovery and development.

e-mail: xiaohua_zhang@merck.com

94. Methods for Multiple Endpoints

CONDITIONAL ESTIMATION FOR GENERALIZED LINEAR MODELS WHEN COVARIATES ARE SUBJECT-SPECIFIC PARAMETERS IN A MIXED MODEL FOR LONGITUDINAL MEASUREMENTS

Ering Li *, North Carolina State University; Daowen
Zhang, North Carolina State University; Marie
Davidian, North Carolina State University

The relationship between a primary endpoint and features of longitudinal profiles of a continuous response is often of interest, and a relevant framework is that of a generalized linear model with covariates that are subject-specific random effects in a linear mixed model for the longitudinal measurements. Naive implementation by imputing subject-specific effects from individual regression fits yields biased inference, and several methods for reducing this bias have been proposed. These require a parametric (normality) assumption on the random effects, which may be unrealistic. Adapting a strategy of Stefanski and Carroll (1987 *Biometrika* 74:703-716), we propose estimators for the generalized linear model parameters that require no assumptions on the random effects and yield consistent inference regardless of the true distribution. The methods are illustrated via simulation and by application to a study of bone mineral density in women transitioning to menopause.

e-mail: eli@stat.ncsu.edu

JOINT MODELING OF ADHERENCE AND REPEATED MEASURES

Wei Lang*, Wake Forest University
Lisa A. Weissfeld; University of Pittsburgh
Ralph B. D'Agostino, Jr., Wake Forest University

In clinical trials evaluating the effect of treatment on outcomes measured repeatedly over time, participant's adherence to the treatment prescribed has a direct impact on the outcomes. Data analysis including participants who adhere to the treatment needs to be adjusted in order to draw valid conclusion concerning the treatment effect. Adherence can be analyzed as either a 0-1 outcome at each time point or as a time-to-event analysis with the event being time to nonadherence. We construct a copula-based parametric model (CPM) for the joint analysis of correlated discrete and continuous outcomes. A separate CPM is also proposed for the joint modeling of time-to-event and longitudinal outcomes. Data from a behavioral weight loss study and postmenopausal estrogen/progestin intervention clinical trial are analyzed using the proposed method.

e-mail: wlang@wfubmc.edu

K-MEANS CLUSTER ANALYSIS OF LONGITUDINAL TRAJECTORIES WITH SMOOTHING

David R. Gagnon * , Boston University School of Public Health;
Margaret Stedman, Boston University, School of Public Health

Longitudinal cohort studies provide rich sources of data that allow us to assess the effect of changing risk factor profiles over time on adverse outcomes. Techniques such as survival analysis with time dependent covariates focus on the effect of current risk factors present at the time of individual events, ignoring information present in the overall trajectories of risk factors over time. While this is a considerable improvement over the simpler approach of assessing risk factors at baseline and observing their effects on survival, there has been a lack of attention to the overall patterns of these risk factors over the course of follow-up and their effects on survival. Using cluster analysis, we group subjects with similar risk factor trajectories over time and then examine the effect of these groups on adverse events. We present a k-means approach to the cluster analysis of longitudinal risk factor trajectories which uses LOESS-smoothed trajectories and a mean squared distance metric and results in distinct classes of subjects with differing risk factor trajectories over time. The association of these groups to survival is then assessed. An example using the VA Normative Aging Study data is presented.

e-mail: gagnon@bu.edu

APPLICATION OF THE TIME-VARYING AUTOREGRESSIVE MODEL WITH COVARIATES

Tulay Koru-Sengul*, University of Pittsburgh
David S. Stoffer, University of Pittsburgh
Patricia R. Houck, University of Pittsburgh
Jean M. Miewald, University of Pittsburgh
Daniel J. Buysse, University of Pittsburgh
David J. Kupfer, University of Pittsburgh

The Time-Varying Autoregressive model with covariates (the time-varying ARX) can be thought of as a version of the transitional general linear models for analyzing longitudinal data. Variations of the time-varying ARX model have been seen in the time series and the econometrics literature but this model has not been used to analyze longitudinal data. Outcome variable over time is modeled as a linear regression model for each time point by taking into account the correlation between the repeated measurements over time. The major advantages are identifying the time points where response changes, handling of unequally spaced data, and being able to use staggered time-varying covariates. The time-varying ARX model is applied to the sleep measures in a healthy adult sample of The Pittsburgh Study of Normal Sleep (PSNS). The results show that flexibility of sleep habits is related with reduced time to sleep. Research supported in part by the grant from National Science Foundation (DMS0102511) and the Mental Health Intervention Research Center for Mood and Anxiety Disorder (MH30915).

e-mail: tkst7@pitt.edu

POINT ESTIMATE AFTER GROUP SEQUENTIAL TESTS

Keyue Ding*, National Cancer Institute of Canada,
Clinical Trials Group Queen's University

Repeated significance tests not only increase the overall type I error, but also introduce bias in estimating the unknown primary parameter. In this paper, the bias of mle is obtained by conditioning on the stopping time for monitoring a drift parameter of a Brownian motion, and a bias corrected estimate is proposed. Although it still overestimates the drift parameter, Monte-Carlo studies show that it is precise for the practical use. Comparisons with other existed estimates are presented, and the proposed estimate outperforms other estimates in term of having a smaller overall mean squared error.

e-mail: kding@ctg.queensu.ca

INVESTIGATION OF TWO KINDS OF DECISION RULES FOR BIVARIATE SEQUENTIAL TESTS BASED ON MARGINAL CRITERIA

Yanli Zhao*, University of Minnesota
Patricia M Grambsch, University of Minnesota
James D. Neaton, University of Minnesota

Many clinical trials have multiple outcomes and formal interim monitoring guidelines that consider them can be important. We consider the bivariate case. Previous research has focused on marginal criteria for bivariate endpoints that control the overall type I error. For example, for trials with an efficacy and safety endpoint, the trial is stopped when either endpoint crosses a boundary. In many trials, the two endpoints of interest can both be considered efficacy or a combination of efficacy and safety endpoints, e.g., survival and progression to AIDS. We consider this situation and an interim analysis guideline for a trial with two treatment groups that controls the overall type I error and stops early only if both endpoints indicate superiority for the same treatment. We develop sets of two boundaries, a higher and a lower, permitting one endpoint to be primary and the other supportive or secondary, with or without prespecifying which one is primary. The effect of the correlation between the two endpoints on the boundaries are considered and compared with the situation when the stopping guideline is based on crossing a boundary for either endpoint vs. both endpoints. The results show that for both Pocock and OBF error spending functions increasing the absolute value of the correlation has a much larger impact on the boundaries for the “and” rules than for the “or” rules and acts in the opposite direction. We discuss the reasons for this, in the context of examples of trials appropriate to each decision rule.

e-mail: yanliz@biostat.umn.edu

STATISTICAL ISSUES IN THE ANALYSIS OF EXPERIMENTAL PSYCHOPHARMACOLOGY STUDIES

Ralitza V. Gueorguieva*, Yale University; Yu-Te Wu,
Yale University; John Krystal, Yale University

Experimental psychopharmacology studies are widely used. The ethics of such studies in psychiatric clinical trials have been subject to a considerable debate, whereas important statistical design and analysis issues have been neglected. These studies commonly employ complex multivariate repeated measures designs. Their data also exhibit dose-related changes in the location and shape of the distribution and floor/ceiling effects. Traditionally repeated measures ANOVA's or linear mixed models have been used to analyze such data. However assumptions underlying these methods are usually not satisfied. Ordinal data GEE and recently proposed nonparametric procedures for repeated measures designs (Brunner, Domhof and Langer, 2002) are alternative approaches. In this presentation we assess the effect of violations of assumptions on the tests and on the estimates from mixed effects models. We also present simulation results from power and type I error comparisons between the mixed effects, GEE and nonparametric approaches. Data from a clinical trial designed to compare the effects of ketamine in healthy controls to the effects in alcohol dependent subjects is used for motivation and illustration.

e-mail: ralitza.gueorguieva@yale.edu

APPLICATIONS OF CONVOLUTION PARAMETER ESTIMATION

Calvin L. Williams*, National Science Foundation

The concept of convolutions in 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 Lundford, 1974 and Brown, 1977). The idea of estimating the parameters of these types of convolutions have been discussed with some intrepidation by Sclove and Van Ryzin (1969) and Gong and Samaniego(1981). More recently Sprott(1984) tackled the issue of estimating parameter(s) of convolutions. We review these proposed approaches and offer some additional application of these techniques basde on some additional observations.

e-mail: cwilliam@nsf.gov

BOUNDS ON THE BIASING PARAMETER AND THE MOMENTS OF THE STOCHASTIC SHRINKAGE
PARAMETERS OF THE LIU
TYPE ESTIMATOR

Fikri Akdeniz*, Cukurova University

One of the problems with the new biased estimator (Linear- Unified –Liu - estimator), is the appropriate value for the unknown biasing parameter d . In this paper we consider the optimum value for d and give upper bound for the expected value of the estimator of this biasing parameter. We also derive the general expressions for the moments of the stochastic shrinkage parameters of the Liu estimator and the generalized Liu estimator.

e-mail: akdeniz@mail.cu.edu.tr

ESTIMATION OF ATTRIBUTABLE FRACTION USING DATA FROM PARTIAL QUESTIONNAIRE

E. Andres Houseman*, Harvard School of Public Health
Donald K. Milton, Harvard School of Public Health

The attributable fraction (AF) is often used to explore the policy implications of an association between a disease and an exposure. To date, there have been no proposed estimators of AF in the context of partial questionnaire designs. We first propose a computationally efficient estimator of regression parameters using weighted estimating equations. In particular, assuming a log-linear model for the distribution of missing covariates, we employ the methods of Wacholder (1994) to motivate consistent estimating functions, and weight each subject's contribution by the inverse probability of responding to the questionnaire. Then we use the regression estimates to derive two types of AF estimates: a 'semiparametric' estimate that extends existing AF formulas by summing over weights; and a 'model-based' estimate that uses the assumed joint probability model for disease status and covariates. The latter approach admits an adjustment to the regression intercept, which may be necessary when the frequency of cases in the sample does not match the frequency of cases in the target population (e.g. case-control studies). Finally, we demonstrate our methods using data obtained from a study on adult occupational asthma, conducted within a Massachusetts HMO.

e-mail: ahousema@hsph.harvard.edu

ESTIMATE THE MEAN TREATMENT EFFECT IN THE PRESENCE OF MEASUREMENT ERRORS

Zhe Shang*, Wyeth Pharmaceuticals

Monitoring renal function is essential for transplantation patients undergoing immunosuppressive therapy with drugs which are potentially nephrotoxic. One common measure of renal function is the glomerular filtration rate (GFR), which usually served as an important end point in transplantation studies. Even though direct accurate measure GFR is possible, they are too expensive for routine use, instead, calculated GFR based on some empirical formulas are widely used. In this talk, two estimators, combining both the measured without error (directly measured in here) which obtained in an internal validation subsets of patients and measured with error (calculated in here) which is available for all the patients, are proposed to estimate the population mean in the presence of measurement error without assuming an error is unbiased or normally distributed. The large sample variance estimators of these two estimators are also derived. Then these two estimators are used to estimate treatment effects, i.e. the population mean difference. Four estimators, i.e., the two estimators proposed before, with two additional estimators, one only uses the measurements without error and another only uses the measurements with errors, are compared under different scenarios through simulation studies.

e-mail: shangz@wyeth.com

PROPERTIES OF WALD AND SCORE STATISTICS — THE TWO SAMPLE CASE FOR GENERALIZED LINEAR MODELS, AND IMPLICATIONS FOR MORE GENERAL MODELS

David I Warton*, University of New South Wales

Score and Wald statistics based on the canonical parameterization are routinely used for hypothesis testing using generalized estimating equations. However, both score and Wald statistics have some undesirable properties — score statistics often have relatively low power, the reasons for which are unclear, and Wald statistics based on the canonical parameterization can be undefined when the sample estimate of a parameter is on the boundary of the parameter space. I investigated why the score statistic occasionally has low power, and whether other parameterizations of the Wald statistic have better properties. These questions were investigated for a simple case — two-sample tests for a generalized linear model, with results having implications for more general models, in particular, for generalized estimating equations. Wald statistics under the variance stabilizing parameterization and the skewness reducing parameterization were found to be good practical approximations of the log-likelihood ratio statistic, although they are only first order approximations (as are other Wald statistics), and were found to have relatively high power for most alternatives. The score statistic can have poor power properties because the Fisher information matrix under H_0 can be a poor estimator of the variance matrix of parameters if H_0 is not true.

e-mail: dwardon@maths.unsw.edu.au

WHAT IS THE DEGREE OF DEVIATION FROM BALANCE IN COMPARING TWO INDEPENDENT GROUPS MEANS? A SIMULATION STUDY

Handan Çamdeviren*, Mersin Üniversitesi; Mehmet Mendep, Mersin Üniversitesi; Refik Burgut, Mersin Üniversitesi, Mehmet A. Sungur, Mersin Üniversitesi
Seval Kul, Mersin Üniversitesi

The purpose of this study is to evaluate the effect of deviation's degree from balance on Type I error and test power in comparison of two independent group means with samples drawn from normal distribution. Independent t test is used to compare to means from two independent samples which have different sample size and different variance ratio. Samples are drawn from normal distribution with IMSL library in FORTRAN 90 software. Type I error is accepted 5%. 100,000 trials are realized for each condition produced. For power calculation, small and large values of standardized mean difference are accepted 0.3 and 1.1 respectively. When assumptions of independent t test are valid, actual Type I error protect their level determined and they don't influence sample size and balance. Under different variance conditions, when much more observations are allocated to the group, big variance, generally actual Type I error is found smaller, when less observations are allocated to this group generally it is found bigger than its value expected. In addition, when standardized mean difference is large, power values obtained generally took medium or high. In this case, power values are found little bit smaller only on account of increasing in ratio between variances of groups.

e-mail: hcamdeviren@hotmail.com

ADJUSTING O'BRIEN'S TEST TO CONTROL TYPE I ERROR FOR THE GENERALIZED NONPARAMETRIC
BEHRENS-FISHER PROBLEM

Peng Huang*, Medical University of South Carolina
Barbara C. Tilley, Medical University of South Carolina
Stuart Lipsitz, Medical University of South Carolina
Robert F. Woolson, Medical University of South Carolina

O'Brien (1984) introduced a simple nonparametric test procedure for testing whether multiple outcomes from one treatment have consistently larger values than outcomes from the other treatment. We first explore the theoretical properties of O'Brien's test. We then extend it to the general nonparametric Behrens-Fisher hypothesis problem when no assumption is made regarding the shape of the distributions. We provide conditions when O'Brien's test controls its error probability asymptotically and when it fails. We also provide adjusted tests when the conditions do not hold. Throughout the paper, we do not assume that all outcomes are continuous. An application from a Parkinson's disease clinical trial is given.

e-mail: huangp@musc.edu



A COMPARISON OF ESTIMATORS OF POPULATION SLOPE UNDER INFORMATIVE DROP-OUT IN LONGITUDINAL STUDIES

Hai Lin and Elizabeth H. Slate, Medical University of South Carolina

We address estimation of the population slope of a longitudinal response in the presence of informative drop-out. Traditional methods, such as generalized mixed model analyses, do not account for informative drop-out, and hence can produce biased estimates of the population slope. We use a simulation study to investigate the performance of four estimators of population slope under drop-out of varying degrees of “informativeness.” The four estimators are the weighted and unweighted least squares methods, the empirical Bayes method (Mori, et al., *Stat. Med.*, 11:621-631, 1992) and the informative right censoring adjusted estimator (IRCAE) of Mori et al., *Biometrics*, 50:39-50, 1994). The estimators are evaluated by MSE and bias. Of particular interest is the new comparison of the empirical Bayes estimator to IRCAE, for which our results show empirical Bayes with less bias and less MSE than IRCAE for the differing degrees of informativeness. However, both empirical Bayes and IRCAE estimates are more attractive and competitive methods than both the weighted and unweighted methods under informative censoring. We apply all four methods to estimate the slope of longitudinal prostate specific antigen readings for men in a cancer prevention trial.

e-mail: linh@musc.edu

A BAYESIAN NONPARAMETRIC REPEATED FRACTIONAL DATA MODEL

Ying Yang, M.D. Anderson Cancer Center
Peter Mueller, M.D. Anderson Cancer Center

The linear mixed effects model with normal errors is a popular model for the analysis of repeated measures. The logit transformation is often used if data takes values within 0 and 1 which are excluded. When the measurements of interest are fraction data, with range of possible values from 0 to 1 with positive probability masses on 0 and 1, the typical logit transformation is inappropriate. We introduce a model augmentation with latent variables that address this by including probability point masses at 0 and 1 in the model. The linear mixed effect model is imposed on the latent variables. In the linear mixed model, the random effects are typically assumed to be independently and identically distributed from some known parametric family. This assumption may be inappropriate if the distribution of random effects is multimodal and/or with unpredictable types of skewness. Misspecification of the random effect distribution may lead the parameter estimates of interest to be poor. We therefore propose a Bayesian nonparametric prior, Polya tree, for the random effects to capture the possible modality and skewness. Computation in the proposed nonparametric model is not as straightforward as in the empirical Bayes scheme. We discuss the posterior inference in three different cases. The proposed method is illustrated in two examples, including a simulation study and a cancer study in dogs.

e-mail: yingyang@mdanderson.org

MODELING TWO OR MORE CATEGORICAL VARIABLES THAT ALLOW FOR MULTIPLE CATEGORY CHOICES

Christopher R. Bilder, University of Nebraska-Lincoln;
Thomas M. Loughin, Kansas State University

Multiple-response (or pick any/c) categorical variables summarize responses to survey questions which ask ‘pick any’ or ‘choose all that apply’ from a set of item responses. The purpose of this research is to introduce extensions to loglinear modeling in order to model the associations between these types of variables simultaneously across all their items. Since individual item responses to a multiple-response categorical variable are likely to be correlated, the usual chi-square approximations to loglinear model goodness-of-fit statistics are not appropriate. A new bootstrap procedure is proposed to approximate the distribution of these statistics. Approximation methods modified from Rao and Scott (1984, *Annals of Statistics*) are also presented for the problem. Simulations show the bootstrap procedure performs very similar to the non-model based procedures proposed by Bilder and Loughin (2004, *Biometrics*) to test for simultaneous pairwise marginal independence. The main advantage to the modeling approach proposed here is that it provides a set of association structures to use when simultaneous pairwise marginal independence does not hold.

e-mail: cbilder3@unl.edu

MODEL SELECTION FOR CLUSTERED RECURRENT EVENT DATA USING NESTED GAMMA FRAILTY MODELS

Xin Zhi, University of Minnesota School of Public Health;
Lynn E. Eberly, University of Minnesota School of Public Health

Correlated time-to-event data are commonly encountered in biomedical research, often as recurrent event data or clustered data. To accommodate the correlation among survival times within the proportional hazards model framework, frailty models have been developed. For recurrent event data collected in a multi-center clinical trial, patients within a clinic share a common frailty, while multiple events for each patient share another common frailty; the second frailty is nested in the first one. A shared gamma frailty model (one level of frailty) and a multiplicative double gamma frailty model (two levels of frailty) can both be fitted to this type of data using EM-based algorithm. This leads to a need for appropriate tests for frailty model selection. We carried out a simulation study to examine the performance of a likelihood ratio test in choosing between the single and double frailty model. Simulation results show that when the single frailty is the correct model, the LRT statistic is approximately distributed as 50:50 mixture of $\text{chisq}(0) + \text{chisq}(1)$. The power of the test to choose the correct model varies according to number of clinics, patients per clinic, and events per patient, as well as the strengths of the frailty effects. An example is provided to illustrate the application of the LRT in practice.

e-mail: xinzhi@biostat.umn.edu

RELIABILITY AND VALIDITY OF DAILY COLLECTION OF SYMPTOM DATA IN MIDLIFE WOMEN

Imke Janssen, Rush University;
Peter M. Meyer, Rush University;
Kelly Karavolos, Rush University;
Lynda H. Powell, Rush University

Missing data not only present challenges to the analysis, they may lead to false conclusions and therefore jeopardize a study's validity. We present several graphical methods appropriate in large longitudinal studies to assess questions about the quality of data collection. We use data from 667 women who filled out a daily diary, a collection of symptom data in the Study of Women's Health Across the Nation (SWAN). The diary consists of 18 questions about mood, vasomotor symptoms, sleep, forgetfulness, sexual desire, and pain during the past 24 hours, for each day of the menstrual cycle. Does the completion rate change over the time of collection? Are there differences between ethnic or educational subgroups? Is the incompleteness or reliability worse for longer cycles? Our graphs identify subgroups with lower completion rates. They also reveal some variables that are more problematic than others. The results will help investigators in redesigning the diary and in identifying subgroups for whom additional resources may increase completion rates. **ACKNOWLEDGMENT** The Study of Women's Health Across the Nation (SWAN) was funded by the National Institute on Aging, the National Institute of Nursing Research, and the NIH Office of Research on Women's Health.

e-mail: Imke_Janssen@rush.edu

NON-PARAMETRIC ANALYSIS OF SEQUENCE HETEROGENEITY ASSOCIATED WITH INJECTION DRUG USAGE

Mei-Fen Yeh, Johns Hopkins University Bloomberg School of Public Health;
Jeanne Kowalski, Johns Hopkins University School of Medicine;
Guang Wen Zhang, Johns Hopkins University Bloomberg School of Public Health;
Qiajia Shao, Johns Hopkins University Bloomberg School of Public Health;
Mike Schneider, Johns Hopkins University Bloomberg School of Public Health;
Alan Templeton, Washington University;
Richard Markham, Johns Hopkins University Bloomberg School of Public Health.

In this article, we propose a non-parametric method for the analysis of genetic diversity by constructing a clonal heterogeneity estimate within each patient and visit. By first mapping the genetic sequence into a numeric representation, we obtain a single estimate through defining a genetic distance measure to summarize the heterogeneity among pairs of clones. This approach is especially useful for dealing with clonal sequencing method with a quasi-species nature, such as the HIV virus. Furthermore, this single measure is readily incorporated into traditional model based analyses to examine relationships among several variables for inference, such as Generalized Estimating Equations, which was implemented in this study. We performed this methodology to examine the effect of a history of injection drug usage on diversity in the HIV-1 protease region, on a subset of longitudinal data collected from the Women's Interagency HIV Study (WIHS). A total of 124 subject visits by 58 subjects were analyzed. The result suggested that a history of injection drug usage appeared to be associated with a higher level of genetic diversity among subjects with CD4 T cell levels > 650 (estimate=0.98, standard error=0.50, 90% confidence interval = (0.10, 1.73), p-value=0.0639).

e-mail: myeh1@jhsph.edu

SCHEMATIC: A PROGRAM FOR CLASSIFYING AND DIAGRAMMING MEDIATORS, MODERATORS, AND INDEPENDENT, PROXY AND OVERLAPPING RISK FACTORS

John Scott, University of Pittsburgh School of Medicine;
Helena C. Kraemer, Stanford University;
Marilyn J. Essex, University of Wisconsin – Madison;
Ellen Frank, University of Pittsburgh School of Medicine;
David J. Kupfer, University of Pittsburgh School of Medicine

Terms such as ‘mediator’ and ‘moderator’ are commonly used in the risk factor literature, but poorly defined. A conceptual model recently proposed by Kraemer et al. [American Journal of Psychiatry 2001; 158: 848 – 856] classifies bivariate relationships among risk factors into five categories based on temporal order, correlation, and dominance of one risk factor over another or codominance between risk factors. This classification can be operationalized for continuous outcomes with a multiple regression model and for binary outcomes with a logistic regression model. The authors present a program that implements such an operationalization for large datasets and produces schematic diagrams depicting the relationships among risk factors. The program, Schematic, is written in S, for the S-Plus or R computing environments. An application of the program to a large psychiatric dataset is presented. Supported by the Mental Health Intervention Research Center for Mood and Anxiety Disorders (NIMH Grant MH30915) and by the John D. and Catherine T. MacArthur Foundation Research Network on Psychopathology and Development.

e-mail: jas96@pitt.edu

SAMPLE SIZE DETERMINATION IN STUDIES WHERE HEALTH STATE UTILITY ASSESSMENTS ARE COMPARED ACROSS GROUPS AND TIME

Barbara H. Hanusa, University of Pittsburgh;
Christopher R.H. Hanusa, University of Washington;
Chung-Chou H. Hanusa, University of Pittsburgh;
Kevin L. Kraemer, University of Pittsburgh.

An initial aim of cost effectiveness and utility analyses of health care programs is to establish the validity of the numerical utility values assigned to different health states. There are 3 methods of assigning values but there is no gold standard to validate these values. The usual standard is to compare the ranking of these numerical values to the ranking of the health states when all health states are viewed simultaneously. Differences in the rankings derived from these two sources are described by measures of consistency. Studies that aim to compare consistency measures across different methods of assigning values or observations must rely on the distributions of these measures to estimate required sample sizes. Using an example from a recent proposal designed to place valid numerical ratings to 6 alcohol related health states, we will present the results from sample size calculations for 3 consistency measures: Spearman’s rho, the inconsistency index (Geisler, et al, 1999) and the numbers of inversions needed to make the ranking and the values from the ratings coincide. The distributions of these measures differ and change the required sample size, thus demonstrating the importance of considering the underlying distributions of these indices for sample size and analysis plans.

e-mail: hanusabh@msx.upmc.edu

SPATIAL CLUSTER DETECTION USING BAYES FACTORS FROM OVERPARAMETERIZED MODELS

Ronald E. Gangnon, University of Wisconsin - Madison;
Murray K. Clayton, University of Wisconsin - Madison

We consider a partition model for estimation of regional disease rates and for detection of spatial clusters. Formal inference regarding the number of partitions (or clusters) can be obtained using a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. As an alternative, we consider models with a fixed, but overly large, number of partitions. Localized Bayes factors can be used to calibrate the results, thus providing informal inferences about the number and locations of clusters. We illustrate these two approaches using the well-known New York leukemia data.

e-mail: ronald@biostat.wisc.edu

EXTENDING THE TWO-PART MODEL TO ACCOMMODATE MULTIPLE RESPONSE OPPORTUNITIES

Leann Myers, Tulane University School of Public Health & Tropical Medicine;
Yeonjoo Yi, Tulane University School of Public Health & Tropical Medicine

Zero heavy data include both nonresponders clustered at zero and responders whose values are often characterized by positive skew. In two-part models, a regression model is used to predict who will respond in part one. In part two, level of response is predicted based on a regression model using data from responders. The two parts are combined into a single model. When there is a single score per subject, logistic regression is used in part one and standard multiple regression is used in part two. Two-part models can be applied in epidemiological contexts where the study sample includes both responders (smokers reporting how many cigarettes smoked daily) and nonresponders (nonsmokers). We extend the two-part model to incorporate multiple measures per subject. We use generalized estimating equations (GEE) methods with a logit link to predict who will respond at different time points. Using only the non-zero responses, GEE methods using a normal link are used to predict the mean level of response at different time points. The two parts are then combined to obtain a single prediction model. Examples are provided and limitations are discussed.

e-mail: myersl@tulane.edu

EFFECTS OF MISCLASSIFICATION ERROR IN HIERARCHICAL POISSON MODELS

Leslie H. Morgan, Tulane University;
Leann Myers, Tulane University;
Frances J. Mather, Tulane University

One application of hierarchical Poisson models is to analyze the spatial distribution of cancer cases. For cases that have been geocoded to residential address and aggregated to larger geographic areas, the counts are not completely accurate. An empirical study of the hierarchical Poisson model's performance to varying levels of misclassification error was conducted. The population structure of the Louisiana counties was used to generate cases that follow a Poisson process with three fixed effects, age, year of diagnosis, and a county-level deprivation score. A random intercept model with a normally distributed county-level effect was simulated with three levels of variance. The first level was no variance attributable to the group level. The second and third levels of variance resulted in expected risk ratios between counties of 2.0 and 4.0, respectively. The levels of misclassification error were set at 0, 5, and 10 percent. The models were simulated with and without a cluster effect. For the 18 resulting models, the bias of the estimates was examined as well as the distribution of the scale parameter.

lmorgan1@tulane.edu

A COMPARISON OF GOODNESS-OF-FIT TESTS FOR BINOMIAL GENERALIZED ESTIMATING EQUATIONS MODELS

Huiyi Lin, Tulane University
Leann Myers, Tulane University

Binary outcomes are very common in medical and epidemiological studies. When scores are independent, logistic regression models are used. When outcomes are dependent, such as those in longitudinal and cluster studies, generalized estimating equation methods (GEE) are often used to analyze the correlated binary data. For any modeling procedure, an essential step is to determine the goodness-of-fit (GoF) for the final model. For logistic regression, standard GoF tests have been developed and are available in most statistical software. GoF statistics for GEE, however, have been developed more recently, and have not been incorporated into the software. Several GoF statistics for GEE have been proposed (Barnhart & Williamson 1998; Horton et al 1999; Pan 2002). The objective of this study is to compare these seven GoF statistics for GEE models using simulation data under different conditions. Sample sizes were varied, as were the possible covariates (discrete or continuous). Different models included time-dependent and time-independent covariates, quadratic components and interactions. We report here the results when there were two scores per subject, with either low or moderate correlation between scores. No single GoF statistic performed best under all conditions. The optimally performing GoF for each model type is described.

e-mail: hlin4@tulane.edu

QTL ANALYSIS IN GENOTYPICALLY SELECTED EXPERIMENTAL POPULATIONS

Zongli Xu, University of North Carolina at Chapel Hill; Fei
Zou, University of North Carolina at Chapel Hill;
Todd J. Vision, University of North Carolina at Chapel Hill

Selective mapping is a strategy originally conceived for the construction of a high-density linkage map in an experimental cross with a minimum of genotyping. The idea is to first build a sparse linkage map based on a large number of individuals, from which a smaller number of individuals are selected. The selected sample is optimized to have a higher map resolution than would a random sample. The selected sample is subsequently genotyped for a large number of markers. In principle, a selected sample could also be used for the mapping of quantitative trait loci (QTL). However, the statistical properties of QTL mapping with such a sample have not been studied. We empirically and analytically investigate the bias, resolution and power of QTL mapping in a selected sample. We find that use of a selected sample substantially improves QTL mapping resolution relative to a random sample without introducing bias in the estimate of QTL position and effect size. However, marker density needs to be relatively high in order to achieve the equivalent detection power of random samples. Currently available QTL mapping methods can be easily adapted to analyze selected samples. **Keywords:** selective mapping, QTL, experimental crosses **Acknowledgements.** This work is supported by NSF grant DBI-0110069.

e-mail: zongli@email.unc.edu

A COST EFFECTIVENESS ANALYSIS OF MEAN TIME IN MODERATE – TO-SEVERE ALZHEIMER’S DISEASE

Xiaoning (Joanie) Zhu, Medical University of South Carolina;
Kit N. Simpson, University of South Carolina

Alzheimer’s Disease (AD), which affects about 4.5 million Americans, is a chronic illness with a progressive loss of cognitive and intellectual abilities, and is associated with substantial medical and non-medical costs. Patients with moderate AD require help with some activities of daily living, and severe cases need costly 24-hour supervision. The cost of AD in the US is estimated at \$61 billion a year, including caregiver time, productivity loss and medical expenses associated with AD. 75% of the total healthcare costs associated with AD occur during the stage of severe dementia. A previous study has shown that memantine treatment of patients with moderate to severe AD is cost saving comparing with placebo. However, no studies has been performed that examines the potential economic differences between memantin and the drugs used in current practice. We use a decision analysis approach to examine this question. Our analysis is structured as a decision tree model. The model parameters are based on the results of two published clinical trials, and the use of institutional and outpatient resources by severity of illness published by the Alzheimer’s association in 2002. A key model assumption linking a drug’s effectiveness and cost-effectiveness related to the risk of a change in patients’ residential setting. **Result:** Our model predicts that the use of memantine to treat moderate-to-severe AD is slightly less cost-effective than the most commonly used current drug: donepezil. Sensitivity analysis revealed that this model is only marginally sensitive to variations in the assumptions related to the efficacy measurement variables used, or to assumptions related to the patients’ residential status at baseline. **Conclusion:** Based on clinical trial reports to date and current data on treatment patterns and cost of care, we may expect negligible economic benefits from the use of memantine compared to other approved drugs for moderate-to-sever AD.

e-mail: zhu@musc.edu

A FULL PEDIGREE BASED METHOD FOR THE STATISTICAL ASSESSMENT OF GENETIC ANTICIPATION

Tracy J. Costello, M.D. Anderson Cancer Center and The University of Texas - Houston;
Christopher I. Amos, M.D. Anderson Cancer Center

Genetic anticipation is defined as decreasing age of onset or increasing severity as a disease is passed through generations. Statistical methods applied to evaluate anticipation include t-tests (compare age of onset in children & adults) and an affected parent/affected child pair method correcting for selection through affected children (Huang, Vieland 1997). These methods have been shown to be insufficient for assessing anticipation due to familial correlation and low power. First, we use a likelihood-based method (regressive logistic hazard) to model anticipation in families by a generational covariate. The second method allows alleles to mutate as they are transmitted from the parents to the children, modeling triplet repeat diseases where disease alleles can become more deleterious as they are transmitted. Using extended families dramatically improves power to detect anticipation. We present simulation studies comparing data sets with no anticipation, differing levels of anticipation, genetic heterogeneity and missing data to evaluate the power and type I error rates of this algorithm and results from analysis of Huntington's Disease and Li-Fraumeni Syndrome data sets.

e-mail: tthiel@request.mdacc.tmc.edu

LATENT TRAIT RATER AGREEMENT MODELS WITH APPLICATIONS IN ORAL HEALTH

Elizabeth G. Hill, Medical University of South Carolina;
Elizabeth H. Slate, Medical University of South Carolina

We demonstrate the use of latent trait models for the analysis of rater agreement for ordinal category ratings, using the approach of Uebersax and Grove (Biometrics, 1993). These models are intuitively appealing and offer easily interpretable results. For example, parameter estimates can be used to quantify and assess: rater classification thresholds, rater variability, latent correlation between the true and apparent trait, sensitivity, specificity, positive predictive value, and negative predictive value. We extend these models to include both subject- and rater-level covariates. Applications are shown for both simulated data and a caries diagnosis study.

hille@musc.edu

UNDERSTANDING SYSTEMATIC VARIATION IN GEL ELECTROPHORESIS

Jeffrey C. Miecznikowski, Carnegie Mellon University,
Kim Sellers, Carnegie Mellon University;
William F. Eddy, Carnegie Mellon University;
Jon Minden, Carnegie Mellon University

Within the study of proteomics, a major goal is to determine the proteins that are present within a tissue sample. Through fluorescence difference gel electrophoresis, scientists have a method for detecting quantitative changes in the amount and type of proteins between two tissue samples. A major advantage to this system is that both samples are run on the same gel, eliminating the gel to gel variation. This technology allows for two samples to be compared and the differences in proteins between the samples to be recorded. Because of the infancy of this technology, it is dependent on the understanding and control of several systematic sources of variation, such as the bias in the camera and the light within the imaging apparatus. Ultimately, in order to automate a process that accurately determines the amount and identification of proteins in a sample, we must understand and correct for the systematic sources of variation that are present within this technology.

e-mail: jcm3@stat.cmu.edu

SOME NEW MULTIVARIATE CUSUM QUALITY MONITORING PROCEDURES

Richard A. Johnson, University of Wisconsin-Madison;
Ruoja Li, University of Wisconsin-Madison

Our research concerns multivariate quality monitoring schemes for detecting a shift in the mean of a process. For a small but persistent shift in the mean of the process, it is known that the cumulative sum (CUSUM) schemes perform better than the classic Shewhart charts. We propose some new multivariate CUSUM schemes by generalizing the conventional univariate CUSUM scheme to multivariate settings. One of our CUSUM statistics has components which are the difference between the univariate statistic for detecting an increase and that for detecting a decrease. Another scheme projects the multivariate observations on the direction having the largest univariate CUSUM statistic. We compare our new schemes with a number of existing CUSUM schemes, including the scheme proposed by Crosier. In our simulations, among existing procedures, Crosier's approach has the best overall ARL (average run length) performance. Our new multivariate schemes have ARL performances that are generally comparable with Crosier's scheme and are better in some circumstance. Moreover, they provide a natural estimate of the direction of the shift. The large sample properties of both our CUSUM statistics and Crosier's statistic are discussed.

e-mail: ruojia@stat.wisc.edu

FACTORS AFFECTING THE PERFORMANCE OF A SYNDROMIC SURVEILLANCE OUTBREAK DETECTION ALGORITHM

Ling Wang, Boston University School of Public Health;
Paola Sebastiani, Boston University School of Public Health;
Kenneth D. Mandl, Harvard Medical School

How to detect and how fast to detect different types of outbreaks is a crucial problem in syndromic surveillance. The goal is to find out what is going on as soon as possible, as well as to keep the false detection rate as low as possible. We use an autoregressive plus seasonal components model to monitor online the daily counts of chief complaints for respiratory syndromes at the emergency department of an urban pediatric hospital. The online monitoring system is tested to assess the false detection rate. We simulate outbreaks of different shapes and sizes to assess the true detection rate, and we consider the effect of exogenous factors such as air quality or time of the year on the performance of the system by modeling the interplay of the exogenous factors on the true detection rate with a Bayesian network. We found that the online monitoring system has an overall 84.8% detection accuracy across all shapes of outbreaks, while the outbreak size influences the earliness to detection. The false and true positive rates are also associated with the exogenous factors.

e-mail: wangling@bu.edu

BAYESIAN METHODS FOR INDIVIDUAL BIOEQUIVALENCE STUDIES

Janelle S. Erickson, Eli Lilly and Company;
John W. Seaman, Jr., Baylor University

Bioequivalence studies focus on determining if drug formulations are therapeutically equivalent. This is important, for example, in the development of generic drugs. Once a brand-name drug goes off patent, other pharmaceutical companies can develop and market a generic drug provided they can demonstrate that the two drug formulations are bioequivalent. More recently, there has been concern about the interchangeability of brand name drugs and generic drugs for use in a given patient. We present a new Bayesian approach for assessing individual bioequivalence which can accommodate several models of bioequivalence.

e-mail: erickson_janelle_s@lilly.com

RECEIVER OPERATING CHARACTERISTIC CURVE ANALYSIS IN FOLLOW-UP STUDIES WITH CENSORING

James B. Kampert, The Cooper Institute

Receiver operating characteristic (ROC) curve analysis is used in medical studies to evaluate the sensitivity (true positive rate) and specificity (true negative rate) of a diagnostic marker of a disease. ROC analysis is sometimes used in epidemiologic follow-up studies to evaluate the predictive accuracy of risk markers for endpoint events observed over time. Here the empirical true positive rate is an unbiased estimate of the population true positive rate. In cohort studies, however, ascertainment of endpoints is often incomplete due to censoring. Then the empirical false positive rate overestimates the population false positive rate, biasing the ROC curve. We present nonparametric and semi-parametric methods for ROC curve analysis in censored follow-up studies with continuous and ordinal predictors. The nonparametric method invokes Bayes' theorem and the Kaplan-Meier estimator to yield unbiased false positive rates. The semi-parametric method combines a proportional-hazard risk function with the Nelson-Aalen cumulative hazard estimator to produce a relatively smooth unbiased ROC curve. We illustrate and discuss limitations of each method. Monte Carlo simulations compare the accuracy of each method. We discuss sampling errors of the ROC curve and its integral, and inference for correlated ROC curves.

e-mail: jkampert@cooperinst.org

MISSING DATA IN PALLIATIVE CARE

J. Lynn Palmer, M.D. Anderson Cancer Center

Missing data in palliative care research are often not missing at random. Reasons for missing data could be due to increased illness of the patient, a temporary side effect due to medication, or other reasons associated with the current therapy. This discussion will summarize some recent results found in this area, graphically showing some of the associations found, and discuss possible effects this missing data could be having on the reported outcomes of the studies.

e-mail: jlp@odin.mdacc.tmc.edu

COMPARISON FOR STOCHASTIC COMPARTMENT MODELS OF A DIGESTIVE SYSTEM IN FORAGE-FEED ANIMALS

Yoonsung Jung, Kansas State University

Stochastic models of compartmental systems are useful in applications in physiology, pharmacology, biochemistry and related biomedical science. Generalized stochastic compartmental models have been proposed for describing animal nutrition processes, and the new models include gamma retention times and heterogeneous particles (Matis, Wehrly, and Ellis, 1989). The proposed models are relatively easy to fit to data using standard estimation procedures and they discriminate between proposed nutritional hypotheses. This project investigates the applicability of generalized stochastic models to experimental data on four treatments given to each of two cows. The overall best fitting model to these data is the G(8)/E model, which consists of two sequential compartments, the first having age-dependent (γ) retention times and the second having age-independent (exponential) retention times.

e-mail: ysjung72@empal.com

SHRINKAGE CORRECTION FOR PREDICTION – FUTURE RANDOM X CASES

Xue Xin, Tulane University School of Public Health and Tropical Medicine;
Leann Myers, Tulane University School of Public Health and Tropical Medicine

One of the major uses of regression models in medical research is to predict. When the x_i 's at which future predictions are made are not specified in advance but will be observed over certain population, we are essentially predicting the response at a future random covariate vector x . The Stein-type predictors give a uniformly lower mean squared error for prediction than least squares estimators under certain assumptions. Different forms of the Stein-type and nonparametric shrinkage predictors were computed and compared using re-sampling and sample re-using techniques. Data were generated from models with normally and nonnormally distributed error terms. The parameters p (number of predictors), n (sample size) and β were varied. The results showed that Stein correction is only justified when p is equal or greater than 3. When normality and constant variance assumption were violated, shrinkage prediction can still be expected to give a worthwhile improvement over least squares estimator. When the error part in the model is large, the magnitude of β is small or the sample size is small relative to the number of covariates, shrinkage correction is recommended and risk saving is quite substantial.

e-mail: xxin@tulane.edu

EM ALGORITHM BASED GENETIC TESTS OF ASSOCIATION AND/OR LINKAGE WITH MISSING PARENTAL DATA

Chao-Yu Guo, Boston University School of Public Health;
Kathryn Lunetta, Genome Therapeutics, Corp.;
Anita DeStefano, Boston University School of Public Health;
Josée Dupuis, Boston University School of Public Health;
L. Adrienne Cupples, Boston University School of Public Health

Sun et al. (1999) proposed the 1-TDT to detect linkage disequilibrium between a candidate locus and a disease locus using genotypes of the affected individuals and one available parent. In this paper, we apply the EM algorithm to deal with the situation where one parent has missing genotype information, a method that can provide unbiased information about transmitted and non-transmitted alleles. We then combine these one-parent families with families where genotypes are available for both parents to formulate the EM-TDT and EM-HRR statistics. Additional subjects with no parental data can also be incorporated in the EM-HRR statistic. Simulation results reveal that the EM-TDT and 1-TDT perform similarly in detecting association. Due to the haplotype relative risk's data structure, the transmitted markers are always present regardless of missing one or two parents. As a result of having a larger sample size, the EM-HRR is more powerful in detecting linkage disequilibrium than either the EM-TDT or 1-TDT in a population under HWE. If admixture is not extreme, the EM-HRR is still more powerful. When a large degree of admixture exists, a circumstance that is unlikely to occur in samples of the same ethnicity, the EM-HRR performs better when the association is strong, though not as well when the association is weak.

e-mail: guo_chao_yu@hotmail.com

SYSTEMATIC ASSESSMENT OF AFFYMETRIX MICROARRAY DATA QUALITY

Chaohui Wang, Cleveland Clinic Foundation;
Xuejun Peng, Cleveland Clinic Foundation;
Eric Blalock, University of Kentucky

Microarray data present big challenges for “classical” statistical analysis methods because assumptions of normality, homoscedasticity, independence, etc. are usually not valid. To gauge how much these assumptions are violated, we use Affymetrix microarray data collected from a well designed experiment with reasonable sample sizes and study the above properties empirically. Data are processed using several analysis algorithms at both probe level and probe set level as well as different transformation methods and the results are compared. Our conclusion is that interesting genes may be found via multiple ways but no single approach fits all genes. We report our findings in this poster presentation.

e-mail: cwang@bio.ri.ccf.org

THE EXTENDED CLINICAL TRIAL

Robert A. Parker, Center for Biostatistics in AIDS Research, Harvard School of Public Health

The randomized controlled trial is the standard of evidence for assessing the efficacy of a therapeutic intervention for disease. A major question, however, is whether the results are generalizable. This problem includes the population to whom the results apply. Moreover, the selective nature of the population enrolled in a randomized controlled trial cannot be eliminated since each subject (or an appropriate surrogate) must consent to participate in the study and to receive a randomly determined treatment. If many potential subjects refuse to participate, the generalizability of the results becomes very questionable. This problem is particularly severe when the therapy being studied is available without participating in the study. To resolve this problem, I propose the 'extended clinical trial' which involves both the conventional randomized controlled trial plus an observational non-randomized study of four additional groups: those who are ineligible because they require immediate treatment or have contraindications to immediate treatment; and those who choose not to participate either because they want immediate treatment or do not want immediate treatment. I describe methods to combine results from these two separate studies and assess the consistency between the two studies.

e-mail: rparker@sdac.harvard.edu

A SENSITIVITY ANALYSIS OF LONGITUDINAL DROPOUT DATA USING OSWALD

Amy E. Begley, University of Pittsburgh
Gong Tang, University of Pittsburgh
Patricia R. Houck, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh
Benoit H. Mulsant, University of Pittsburgh
Charles F. Reynolds, University of Pittsburgh

Most methods for analyzing longitudinal data assume that the dropouts only depend on observed data; that is, they assume the dropout process is ignorable. The widely used SAS Proc Mixed assumes this mechanism; however, standard error estimates for fixed effect parameters are usually under estimated. OSWALD is a flexible and powerful object-oriented software, written in S-PLUS, for the analysis of longitudinal dropout data (<http://www.maths.lancs.ac.uk/software/Oswald>). In addition to random dropout, OSWALD models longitudinal data with informative dropout (which depends on the missing data). We applied OSWALD to a longitudinal dataset from a clinical trial comparing antidepressant effects in an elderly depressed sample. Several mechanisms such as completely random dropout (MCAR), random dropout (MAR) and informative dropout are applied and the results are compared. As OSWALD does not provide standard error estimates in the informative dropout situation, a bootstrap method was implemented. Research supported in part by the Mental Health Intervention Research Center for the Study of Late Life Mood Disorders (MH52247) and the Mental Health Intervention Research Center for Mood and Anxiety Disorder (MH30915)

e-mail: begleyae@upmc.edu

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

INDEX OF PARTICIPANTS

Abecasis, Goncalo R	30	Black, Michael	55
Abrams, Allyson	69	Blalock, Eric	Poster
Aeschliman, Dana	27	Bloomfield, Clara D	50
Agresti, Alan	64	Boehnke, Michael	30
Akdeniz, Fikri	95	Bollen, Kenneth A	1
Akritas, Michael G	20	Bonangelino, Pablo E	11
Albert, Paul	44, 66	Boos, Dennis D	16, 52
Alderson, Donald	21, 39	Booth, James	6, 64
Alemayehu, Demissie	36	Borkowf, Craig B	29
Allison, David B	62, 78	Bortnick, Steven	41
Amaratunga, Dhammika	32	Bouman, Peter	27
Amemiya, Yasuo	64	Bowman, F. Dubois	28
Amos, Christopher I	Poster	Bragg, Lee	92
An, Hyongjin	39	Braun, Thomas	3, 87
An, Seung-Won	31	Breidt, Jay	25
Anderson, Steven A	24	Broman, Karl W	30
Anderson, Stewart J	32, 42	Brookmeyer, Ron	71
Andrei, Adin-Cristian	3, 54	Broyles, Stephanie T	71
Andrews, Chris A	93	Brun, Marcel	21
Apanasovich, Tatiyana V	23	Bull, Shelley	88
Arena, Vincent C	32, 36, 41	Buonaccorsi, John P	64, 91
Atwatter, Michael	63	Burgut, H.Refik	52, 95
Auh, Sungyoung	76	Burr, Deborah	90
Baggerly, Keith	68	Buyse, Daniel J	94
Bailer, John A	22	Cai, Jianwen	34, 73
Bailey-Kellogg, Chris	53	Calder, Catherine	41
Balasubramani, G.K.	31	Calvin, James A	11
Banerjee, Sudipto	Short Course, 11, 69	Çamdeviren, Handan	95
Banks, David	2, 36	Cao, Dachuang	62
Bao, Xiaoming	15, 50	Capuano, George A	92
Bar-Joseph, Ziv	58	Carlin, Bradley P	Short Course, 83
Barnard, John	93	Carrasco, Josep L	86
Barnhart, Huiman X	86	Carroll, Raymond J	18, 23, 39, 65, 67
Barry, Sarah J E	71	Casella, George	6, 32, 62
Bartolucci, Alfred	62	Chakraborty, Sounak	89
Basu, Saonli	8	Chan, Ivan	65
Beck, Gerald J	37	Chan, Jiahua	30
Begg, Colin B	93	Chan, Wenyaw	76, 91
Begley, Amy E	Poster	Chappell, Rick	29, 79
Begum, Munni	83	Chatterjee, Nilanjan	18
Beiden, Sergey V	52	Chaudhuri, Arin	83
Bellamy, Scarlett	76	Chauhan, Chand K	33
Bellera, Carine	31	Chen, Cong	40
Berbaum, Kevin S	52	Chen, Dung-Tsa	72
Berhane, Zekarias	73	Chen, Hua Yun	82
Berry, Donald A	10, 26	Chen, Jiahua	30
Bertolet, Marianne H	31	Chen, Jie	20
Betensky, Rebecca A	64	Chen, Kun	8
Biernacka, Joanna M	88	Chen, Lu	18
Biggerstaff, Brad J	24	Chen, Min	60
Bilder, Christopher R	Poster	Chen, Ming-Hui	27, 39, 57, 74
Billheimer, Dean	53	Chen, Pai-Lien	85

INDEX OF PARTICIPANTS

Chen, Qingxia	39	Devlin, Bernie	93
Chen, Shu-Chuan	43	Diao, Guoqing	62
Chen, Tai-Tsang	29	Diaz, Rafael E	74
Chen, Wei-Min	30	Diehr, Paula	75
Chen, Ying Qing	82	Diggle, Peter	Tutorial
Chen, Zhen	51, 85	Dignam, James	27
Cheng, Cheng	20	Ding, Keyue	94
Cheung, Ken	3	Ding, Meichun	10
Cheung, Natalie W	61	Dinse, Gregg E	51
Chinchilli, Vernon M	63, 86	Dixon, Dennis O	16
Chiou, Jeng-Min	23	Dixon, Philip M	53
Chiu, Yen-Feng	18	Do, Kim-Anh	57
Christensen, William F	25	Dodd, Kevin W	49
Christman, Mary	25	Doerge, Rebecca W	62
Chute, Christopher G	37	Dominik, Rosalie C	80
Clayton, Murray K	Poster	Dougherty, Edward R	21
Cleves, Mario A	9	Downing, Darryl	79
Coffey, Christopher S	73	Draghicescu, Dana	15
Colaianne, James J	32, 52	Drummer, Thomas D	91
Collings, Bruce J	16	Duan, Fenghai	43
Connor, Jason T	92	Dukic, Vanja	15, 27, 50
Cook, Richard J	17, 92	Dunning, Andrew J	76
Cook, Thomas D	40, 54	Dunson, David B	1, 51
Cordell, Heather J	88	Dupuis, Josée	Poster
Costello, Tracy J	Poster	Durkalski, Valerie L	76
Coull, Brent A	64	Eberly, Lynn E	Poster
Craig, Bruce A	53, 62, 85	Eddy, William F	53, Poster
Craigmile, Peter F	76	Edenberg, Howard	43
Craiu, Radu	27	Edwards, Jode W	78
Cui, Yue	33	Edwards, Lloyd J	84
Cui, Yuehua	62	Eickhoff, Jens C	64
Cupples, L. Adrienne	Poster	Ellenberg, Susan	Short Course
D'Agostino, Jr., Ralph B	94	Elliott, Michael	13, 49
D'Angelo, Gina M	74	Elm, Jordan	44
Dain, Brad	68	Ely, E. Wesley	63
Dalgaard, Peter	Tutorial	Ely, John W	74
Dang, Qianyu	42	Emerson, Scott S	54
Datta, Somnath	20, 29, 87, 90	Eng, Kevin H	40
Datta, Susmita	20	Epstein, Michael P	30
Davidian, Marie	Roundtable, 18, 47, 94	Erickson, Janelle	94, Poster
Davis, Clarence E	Roundtable	Esserman, Denise A	73
Davis, William W	49	Essex, Marilyn J	Poster
Dawson, Jeffrey D	74	Etzel, Carol J	39
de la Chapelle, Albert	50	Faigman, David L	2
De la Torre, Fernando	55	Fan, Jianqing	34, 59
DeGruttola, Victor	85	Fan, Ruzong	19, 62
DeMets, David L	Short Course, 37	Fang, Hongbin	42, 64, 66
DeStefano, Anita	Poster	Fedorov, Valerii V	61
Dendukuri, Nandini	84	Feingold, Eleanor	8, 35, 93
Deng, Shibing	32	Feiveson, Alan H	10
Deng, Wei	7	Feldman, Harold I	21, 65
Detry, Michelle A	91	Feng, Rui	30

INDEX OF PARTICIPANTS

Feng, Zeny Z	30	Graham, Dionne A	11
Ferketich, Amy K	50	Grambsch, Patricia M	94
Feuer, Eric J	49, 63	Gray, Alexander G	72
Fieuws, Steffen	64	Greenhouse, Joel	Roundtable, 5, 31
Fine, Jason P	9	Gregas, Matt	90
Fine, Michael J	80	Gregoire, Timothy G	Roundtable
Fink, Jodie	92	Grill, Warren M	33
Fleming, Thomas	Short Course	Gross, Kevin	65
Foulkes, Andrea S	85	Gross, Ralph	55
Frank, Ellen	Poster	Gu, Charles	19
Frederick, John	15	Gu, Chong	67
Freeman, Jean L	76	Guan, Yongtao	11
Friel, Nial	81	Gueorguieva, Ralitza V	94
Frisén, Marianne E	69	Guha, Subharup	89
Fu, Wenjiang M	6	Guimaraes, Paulo D	44
Fu, Yuejiao	30	Gunasekara, Gaitri	91
Fuentes, Montserrat	Roundtable, 15, 83	Guo, Chao-Yu	Poster
Furukawa, Kyoji	25	Guo, Jia	91
Gabrielson, Edward	46	Guo, Kevin	44
Gadbury, Gary	78	Guo, Wensheng	42, 75
Gagnon, David R	94	Guo, Ying	76
Gail, Mitchell	18, 35, 45	Gurka, Matthew J	84
Galai, Noya	60	Gustafson, Kenneth J	33
Gallop, Robert J	85	Gustafson, Paul	27
Gangnon, Ronald E	Poster	Gwise, Thomas E	22
Gardiner, Joseph C	9	Hachey, Mark	63
Garrett, Elizabeth	46	Hagan, Joseph L	76
Gastwirth, Joseph	35	Halloran, M. Elizabeth	Roundtable, 24
Gaughan, Christina	65	Hanley, James A	31, 84
Gelfand, Alan E	Short Course, 59, 89	Hansen, Ben B	31
Genovese, Christopher	90	Hanusa, Barbara H	Poster
Genton, Marc G	15	Hanusa, Christopher R H	Poster
George, Edward I	85	Hanusa, Chung-Chou H	Poster
Gerber, Georg	58	Hardin, Michael J	50
Ghaffar, Sadia	9	Harezlak, Jaroslaw	42
Ghosal, Subhashis	51, 75	Harlow, Sioban	9, 18
Ghosh, Debashis	46, 81	Harrington, David P	26, 60
Ghosh, Kaushik	63, 82	Hart, Jeffrey D	48
Ghosh, Malay	39, 89	Hartman, Jessica	77
Ghosh, Pulak	10	Hastie, David	33
Ghosh, Sujit	15, 51, 75	He, Shui	41
Gifford, David	58	He, Yulei	7
Gilbert, Peter B	40	Heagerty, Patrick J	23
Gleser, Leon J	22	Hedeker, Don	21, 42, 64
Gneiting, Tilmann	11, 15	Heffernan, Rick	77
Goetghebeur, Els J T	17	Heilig, Charles M	80
Gomatam, Shanti	90	Heitjan, Daniel F	29, 66, 71
Gonzalez, Juan R	73	Heo, Tae-Young	41
Goodman, Steven	5	Herring, Amy H	1
Gorman, Dennis	83	Hill, Andrew N	11
Gould, A. Lawrence	54	Hill, Elizabeth G	Poster
Gould, Larry	9	Hillis, Stephen L	52

INDEX OF PARTICIPANTS

Hoffmann, Raymond G	28	James, Gareth M	58
Hogan, Joseph W	21, 39	Jang, Woncheol	90
Hohmann, Ann	14	Janssen, Imke	18, Poster
Holder, Mark	89	Jemal, Ahmedin	63
Holland, David	15, 83	Jeong, Jong-Hyeon	40
Holloman, Christopher	41	Jiang, Guoyong	61
Holmans, Peter A	88	Jiang, Ruhong	68
Hong, Bo	50	Jiang, Wenyu	87
Hoobs, Charlotte A	9	Jin, Chunfang	62
Horel, Scott	83	Jin, Xiaoping	83
Horvath, Steve	81	Joffe, Marshall M	21, 65
Horvitz-Lennon, Marcela	14	Johnson, Brent A	21, 58, 81
Houck, Patricia R	7, 94, Poster	Johnson, Dallas	71
Houseman, E. Andres	95	Johnson, Laura L	75
Howard, Sandler M	51	Johnson, Richard A	Poster
Hsing, Tailen	21	Johnson, Timothy D	44
Hsu, Chiu-Hsieh	60	Jones, Michael P	71
Hsu, Jason C	72	Joo, Yongsung	6
Hsu, Li	18	Judson, Richard	68
Hsu, Nan-Jung	25	Jung, Inkyung	84
Hu, Chengcheng	59	Jung, Jeesun	19
Hu, Feifang	9, 22	Jung, Sin-Ho	32, 81
Huang, Chiang-Ching	81	Jung, Yoonsung	Poster
Huang, Chiung-Yu	17	Kaiser, Mark S	25
Huang, Hsin-Cheng	25	Kalbfleisch, Jack	30, 60, 87
Huang, Jian	6	Kampert, James B	Poster
Huang, Jianhua	48, 56	Kanade, Takeo	55
Huang, Kun	31	Kang, Dae Ryong	29
Huang, Lan	39	Karavolos, Kelly	Poster
Huang, Peng	74, 95	Karrison, Theodore G	9
Huang, Xiaohong	81	Kass, Rob	4
Huang, Xuelin	17, 40, 73	Katki, Hormuzd A	8
Huang, Yijian	17	Kaufman, Cari	4
Huang, Yunda	81	Kelleher, Kelly	31
Hughes, Timothy R	46	Kendziorski, Christina	62, 72, 85
Hughes-Oliver, Jacqueline M	41, 72	Kennedy, William R	83
Hulin, Wu	70	Kepner, James L	54
Hung, H.M. James	79	Khattree, Ravindra	10
Hunt, Daniel	65	Khuri, Andre I	33
Huo, Dezheng	9	Kiebertz, Karl	44
Hwang, J. T. Gene	20	Kim, Hae-Young	92
Hyrien, Ollivier	65	Kim, Hyung Woo	10
Ibrahim, Joseph G	27, 39, 51, 57, 89	Kim, Jinheum	29
Iglewicz, Boris	33	Kim, Jong-Min	31
Ilk, Ozlem	53	Kim, Kyoungmi	6, 62
Imrey, Peter B	2	Kim, KyungMann	7
Indurkha, Alka	63	Kimmel, Marek	8
Ishwaran, Hemant	52	King, Tonya S	86
Ivanova, Anastasia	74, 80	King, Wayne M	76
Jaakkola, Tommi	58	Kistner, Emily O	8
Jackson, Kirby L	17	Kittelson, John M	54
Jackson, Monica C	11	Klein, Andreas G	92

INDEX OF PARTICIPANTS

Kleinman, Ken	50, 69	Levy, Douglas E	14
Klingenberg, Bernhard	64	Lewis, Paul O	89
Knauss, Jill	21	Li, Bing	56
Kocak, Fatin	52	Li, Chin-Shang	34, 82
Koch, Gary G	54	Li, Chun	93
Kohane, Isaac S	50	Li, Erning	94
Kong, Augustine	38	Li, Hongying	32
Kong, Lan	54	Li, Hongzhe	8
Koru-Sengul, Tulay	7, 94	Li, Huiling	71
Kosorok, Michael R	40, 54	Li, Jianbo	6
Koti, Kallappa M	22	Li, Lang	43
Kou, Samuel	59	Li, Lei	38
Kowalski, Jeanne	Poster	Li, Liang	65
Koyama, Tatsuki	22	Li, Ming	10
Kraemer, Helena C	Poster	Li, Mingyao	30
Kraemer, Kevin L	Poster	Li, Na	19
Krystal, John	94	Li, Quan	52
Kubiak, Rene	44	Li, Runze	70
Kul, Seval	95	Li, Ruoqia	Poster
Kulldorff, Martin	69, 77	Li, Wenjun	84
Kumar, Madan	68	Li, Yan	29, 72
Kuo, Yen-Hong	91	Li, Zhaohai	35
Kuo, Yen-Liang	91	Liang, Hua	70
Kupfer, David J	94, Poster	Liang, Kung-Yee	18, 30, 33
LaFleur, Bonnie	16	Liao, Jason	65
Lai, Tze L	56	Liao, Jiangang	20
Laird, Nan M	42	Lin, Danyu	17, 47, 58, 62
Lam, Raymond L	72	Lin, Hai	Poster
Lancaster, Tony	21	Lin, Huiyi	Poster
Landis, J. Richard	56, 65	Lin, Julia Y	21
Landsittel, Doug	32	Lin, Lawrence	86
Lang, Wei	94	Lin, Min	19
Latimer, Andrew	89	Lin, Rongheng	89
LaVange, Lisa	Roundtable	Lin, Xihong	Roundtable, 9, 17, 21, 30, 67, 81
Law, Ngayee J	51	Lin, Yan	74
Lawrence, Natalia S	28	Lin, Yi	29, 90
Lawson, Andrew B	69	Lin, Yong	20
Lee, J. Jack	10	Lindsay, Bruce G	43, 44, 90
Lee, Jae Won	53	Lipsitz, Stuart R	60, 74, 76, 95
Lee, Ji-Hyun	42	Lisabeth, Lynda D	9
Lee, JungBok	53	Little, Roderick J	Short Course, 21, 31, 39
Lee, Keunbaik	6	Liu, Dawei	81
Lee, Sang-Joon	3	Liu, Diane D	10
Lee, Seung Yeoun	81	Liu, Grace X	33
Lemeshow, Stanley	50	Liu, Jun	59
Leng, Chenlei	29	Liu, Lei	73
Leon, Andrew C	21	Liu, Li C	64
Leonardson, Amy	68	Liu, Li-yu	21
Leong, Traci L	11	Liu, Xuan	11
Lesaffre, Emmanuel	64	Liu, Xueli	81
Levin, Adrian E	63	Liu, Yali	85
Levy, Adrian R	27	Lo, Shaw-Hwa	38

INDEX OF PARTICIPANTS

Localio, Russell	5	Meyer, Peter M	18, Poster
Loecke, David	85	Miecznikowski, Jeffrey	53, Poster
Loeys, Tom	17	Miewald, Jean M	94
Lokhnygina, Yuliya	22	Miller, Forrest R	43
Long, Qi	21	Miller, J. Philip	52
Longini, Ira M	24	Millikan, Randall E	3
Looney, Stephen W	76	Milton, Donald K	95
Loughin, Thomas M	Poster	Min, Young Kyoung	82
Louis, Thomas A	Roundtable, 11, 84, 89	Minden, Jonathan S	53, Poster
Lu, Chengxing	28	Minin, Vladimir	81
Lu, Haolan	84	Mitra, Nandita	63, 93
Lu, Mei	39	Moffitt, Michael A	33
Lu, Shou-En	18, 66	Mokdad, Ali H	49
Lu, Wenbin	60	Molenberghs, Geert	Roundtable, 7, 12
Luan, Jingyu	60	Monks, Stephanie A	68
Luborsky, Judith L	18	Moore, Charity	69
Lumley, Thomas	36	Moore, Dirk F	33
Lunceford, Jared K	18	Moore, Renee H	50
Lunetta, Kathryn	Poster	Morara, Michele	41
Lyles, Cynthia M	92	Morel, Jorge G	17
Lyles, Robert H	50, 86	Morgan, Leslie H	Poster
Ma, Shuangge	54	Morris, Jeffrey S	68
Ma, Yanyuan	91	Morris, John C	52
MacEachern, Steven N	50, 89	Morris, Quaid	46
Mahnken, Jonathan D	76	Mosley, Bridget S	9
Majumdar, Dibyen	12	Mostashari, Farzad	77
Mallick, Bani K	39, 89	Mueller, Peter	3, 10, 57, 67, Poster
Mallick, Madhuj	51	Mukherjee, Bhramar	39
Manatunga, Amita K	50, 76	Mukhopadhyay, Siuli	33
Mandl, Kenneth D	50, Poster	Muller, Keith E	73
Mandrekar, Jayawant N	37	Mulsant, Benoit	7, Poster
Mandrekar, Sumithra J	37	Munoz Maldonado, Yolanda	75
Mann, John	28	Murray, Susan	3, 54, 60
Mariano, Louis T	2	Muthen, Bengt O	1, 92
Marin, Juan-Miguel	53	Myers, Kary L	28
Markham, Richard	Poster	Myers, Leann	18, Poster
Marron, J. S.	58	Naiman, Daniel Q	30
Mather, Frances J	Poster	Nam, Chung Mo	29
Mayrand, Marie-Helene	84	Nan, Bin	9
Mazumdar, Sati	7, 28, 41, 42, Poster	Nandabalan, Krishnan	68
Mazurick, Carissa A	56	Natarajan, Jaya	52
McCann, Melinda H	52	Naureckas, Edward	15
McClure, John D	81	Neaton, James D	94
McConnell, Daniel S	18	Neill, James W	43
McHenry, M. Brent	60	Nelson, Kerrie P	41
McKeel, Jr, Daniel W	52	Newton, Michael A	33
Mehta, Cyrus R	22, 80	Newton, Michael A	53
Meiring, Wendy	15	Nichols, Thomas E	4
Mendep, Mehmet	95	Nicolae, Daniel	38
Meng, Xiao-Li	27, 38, 51	Nie, Lei	12, 62
Meng, Zhaoling	19, 30	Nizam, Azhar	24
Matthew, Stephen J	43	Normand, Sharon Lise	14

INDEX OF PARTICIPANTS

O'Brien, Ralph	45, 61	Qu, Annie	70
O'Brien, Sean M	50	Raghavan, Nandini	32
O'Malley, A. James	14	Raghunathan, Trivellore	Short Course, 7, 13, 49
O'Neill, Robert	Roundtable	Rai, Shesh N	65
Oakes, David	47	Ramanujam, Ramanan	50
Obuchowski, Nancy A	52	Ramoni, Marco F	50
Ogden, Lorraine G	18	Rao, D. C	19
Ogden, Todd	28	Rathouz, Paul	15, 41, 50
Ogle, Stephen	25	Ravishanker, Nalini	51
Olson, Jane M	32, 88	Ray, Surajit	44
Owzar, Kouros	32	Rebelo, Anthony G	89
Paddock, Susan	89	Reed, Carol	68
Pagano, Marcello	11	Reed, George	84
Page, Grier P	20, 78	Reiter, Jerome P	13
Palesch, Yuko	44, 76	Remlinger, Katja S	72
Palmer, J. Lynn	Poster	Reynolds, Charles F	7, Poster
Palta, Mari	7, 63	Ridgeway, Greg	89
Pan, Wei	81	Ritchie, Marylyn D	78
Park, Mira	53	Roayaei, Jean	53
Park, Sohee	41	Robins, James M	23
Park, Soomin	7	Robinson, Bruce	21
Park, Taesung	81	Rockette, Howard E	39, 52
Parker, Robert A	Poster	Roeder, Kathryn	93
Parmigiani, Giovanni	8, 46	Rogerson, Peter A	69
Parsey, Ramin V	28	Root, Tammy L	70
Parsons, Van L	49	Rosenbaum, Paul	Tutorial
Patel, Rajan S	28	Rosenberg, Noah A	38
Paterline, Jennifer E	80	Rosenheck, Robert	14
Patterson, Joan	42	Rosner, Gary L	10, 67
Patterson, Scott D	79	Ross, Candice M	31
Pavlidis, Paul	46	Rotnitzky, Andrea	23
Pena, Edsel A	73	Roy, Jason	21, 39, 71
Peng, Huang	44	Rubin, Donald	13
Peng, Xuejun	6, 44, Poster	Rukhin, Andrew L	2
Peruggia, Mario	72	Ruppert, David	Short Course
Petersen, Lyle R	24	Russek-Cohen, Estelle	25
Peterson, Derick R	90	Rust, Phil	76
Platt, Richard	69	Ryan, Louise M	Presidential Invited Address, 41
Platt, Robert	84	Sabatti, Chiara	78
Pollock, Bruce	7	Saha, Chandan	71
Pong, Annpey	82	Sampson, Allan R	22, 76
Pounds, Stan B	20	Sandikci, Macit U	52
Powell, Lynda H	Poster	Sarkar, Sanat K	20
Price, Julie	28	Sarkka, Aila	83
Pritchard, Jonathan K	38	Sashegyi, Andreas	10
Propert, Kathleen J	61	Satten, Glen A	87, 90
Province, Mike	19	Sawrie, David	73
Putt, Mary E	61	Schabenberger, Oliver	Tutorial, 12, 84
Qaqish, Bahjat F	42, 74, 80, 84	Schadt, Eric	68
Qin, Li	75	Schaffer, Joe	70
Qiu, Jing	20	Scharfstein, Daniel O	60
Qiu, Peihua	48	Schartz, Kevin M	52

INDEX OF PARTICIPANTS

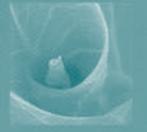
Schell, Michael	80	Snyder, Scott	33
Schenker, Nathaniel	49	Soh, Chang-Heok	60
Schlather, Martin	11	Song, Hae Hiang	52
Schneider, Mike	Poster	Song, Hae-Ryoung	15
Schonlau, Matt	2	Song, Jingli	86
Schuirmann, Donald J	61	Song, Peter X-K	30
Scott, John	Poster	Song, Xiao	17
Seaman, Jr., John W	Poster	Speroff, Theodore	63
Sebastiani, Paola	50, Poster	Spiessens, Bart	64
Sedransk, J.	89	Spinka, Christine	18
Seldin, Michael F	8	Spitz, Margeret	39
Seligson, David B	81	Stanek, Ed	84
Sellers, Kimberly F	53, Poster	Staniswalis, Joan G	82, 90
Seltman, Howard	31	Staudenmayer, John W	91
Selvanayagam, Zachariah	20	Stedman, Margaret	94
Sen, Pranab K	53, 83, 84, 92	Stefanescu, Catalina	27
Sen, Pranab Kumar	32	Stein, Elliot A	28
Serban, Nicoleta	43, 58	Stephens, Matthew	19
Sereika, Susan M	24	Sterrett, Andrew	72
Sha, Najjun	57	Stewart, Emily	92
Shackelford, Linda C	10	Stober, Paul	79
Shang, Zhe	95	Stoffer, David S	94
Shao, Jun	7	Stoto, Michael A	2
Shao, Qiajia	Poster	Strauss, Warren	41
Shardell, Michelle D	60	Strawderman, Robert L	34
Sharples, Katrina	54	Stromberg, Arnold	44
Shen, Lei	7, 50, 72	Sun, Jianguo (Tony)	34
Shera, David M	89	Sun, Jiayang	20, 33
Sherman, Michael	11	Sun, Jingran	48
Shermock, Kenneth M	92	Sun, Lei	88
Shih, Joanna H	18, 66	Sun, Liuquan	34
Shih, Joe	20	Sun, Tony	40
Shintani, Ayumi K	63	Sun, Yanqing	40
Shults, Justine N	56, 83	Sundaram, Rajeshwari	29
Shyr, Yu	10	Sungur, Mehmet A	95
Siegmund, David O	59	Sutradhar, Rinku	17
Silander, John A	89	Sutton, Shae	69
Silverman, Bernard W	58	Swartz, Michael D	8
Simon, Itamar	58	Szatkiewicz, Jin P	8, 35
Simpson, Kit N	Poster	Szolovits, Peter	50
Singer, Julio M	84	Tadesse, Mahlet G	57
Singh, Harshinder	32	Tan, Ming	42, 64, 66
Sinha, Debajyoti	27, 74	Tan, Wai-Yuan	65, 80
Sinha, Samiran	39	Tang, Feng	57
Sirbu, Corina M	9	Tang, Gong	7, 71, 92, Poster
Sirbu, George	22	Tang, Yongqiang	75
Skinner, Jonathan B	61	Tarpey, Thaddeus	28
Slate, Elizabeth H	73, Poster	Tashima, Karen	39
Small, Dylan S	56	Tatikola, Kanaka D	52
Smith, Christina	71	Taylor, Douglas J	61, 80
Smith, Wendell C	44	Taylor, Jeremy	81
Snappin, Steven M	40	Taylor, Jeremy M G	44, 51, 60

INDEX OF PARTICIPANTS

Tchernis, Rusty	14	Wang, Hongwei	40
Tebbs, Joshua M	31, 52	Wang, Ji-Ping Z	90
Templeton, Alan	Poster	Wang, Jiping	52
Ten Have, Thomas	21, 85	Wang, Lily	53
Teres, Jedediah	21	Wang, Lin	7
Thalamuthu, Anbupalam	50	Wang, Ling	Poster
Thall, Peter F	3, 22	Wang, Mei-Cheng	17
Thompson, Mary E	30	Wang, Naisyin	23, 67
Thompson, Mary-Lou	41	Wang, Sue-Jane	61, 68, 79
Thun, Michael J	63	Wang, Wei	77
Tian, Guoliang	42, 64, 66	Wang, Weizhen	74
Tian, Lili	63	Wang, William	65
Tian, Lu	87	Wang, Xiaofei	91
Tien, Hsiao-Chuan	85	Wang, Xiaofeng	33
Tilley, Barbara C	44, 95	Wang, Zengri	84
Tiwari, Ram C	63, 82	Wang, Zuoheng	19
Todem, David	7	Ward, Elizabeth	63
Town, Machell	49	Ward, Ryk	38
Traci, Leong L	83	Ware, James H	5
Troxel, Andrea B	39, 73	Warton, David I	95
Truong, Young	32	Wasserman, Larry	43, 58, 90
Trzaskoma, Ben L	10	Weaver, Mark A	61, 80
Tsiatis, Anastasios A	22, 40, 47, 91, 95	Wei, LJ	87
Tung, Yi-Liang	92	Wei, Qingy	39
Turnbull, Bruce	27	Weijer, Charles	80
Tzeng, Jung-Ying	81, 93	Weinberg, Clarice R	8
van der Laan, Mark	9	Weinrich, Martin C	54
Vannucci, Marina	57, 67	Weissfeld, Lisa	5, 73, 74, 94
Vansteelandt, Stijn	23	Weldon, Steve	44
Vartivarian, Sonya	31	Wells, Martin T	75
Venkatraman, E. S	93	Welsh, Alan	67
Ventura, Valerie	4, 89	Wendelschafer-Crabb, Gwen	83
Verbeke, Geert	12, 64	Westman, Judith A	50
Vickers, Andrew J	37	Williams, Calvin L	82, 95
Vision, Todd J	Poster	Williams, George W	26
Viswanathan, Surya	53	Williams, Malgorzata L	65
Vitek, Jan	53	Williamson, John M	74, 87, 92
Vitek, Olga	53	Wisniewski, Stephen R	31
Vlahov, David	60	Wit, Ernst C	81
Vonesh, Edward F	12	Wittes, Janet	45, 54
Wagner, Robert F	52	Wolfe, Robert A	73
Wahba, Grace	29	Wood, Constance	44
Wahed, Abdus S	40, 91	Woolson, Robert F	95
Wall, Melanie M	11, 33, 42, 64, 91	Wright, Fred A	72
Waller, Lance A	11, 83	Wright, Steve	22
Walls, Theodore A	70	Wu, Chi-tung	25
Wallstrom, Garrick	93	Wu, Chunlei	68
Wand, Matthew	Short Course	Wu, Colin O	56
Wang, Bin	31	Wu, Jianrong	61
Wang, Chaohui	Poster	Wu, Rongling	19, 62
Wang, Deli	6	Wu, Samuel W	32
Wang, Haiyan	20	Wu, Shanshan	89

INDEX OF PARTICIPANTS

Wu, Songmei	28	Zhang, Heping	8, 30, 43, 44
Wu, Wei Biao	48	Zhang, Jiameng	71
Wu, Xiaoling	30	Zhang, Li	68
Wu, Xifeng	39	Zhang, Lijun	65
Wu, Yu-te	94	Zhang, Ping	80
Wu, Yuehui	61	Zhang, Wen	46
Xie, Dawei	31, 49	Zhang, Xiaohua	93
Xie, Sharon X	63	Zhang, Zhiwei	39
Xie, Sunney	59	Zhang, Zhongfa	20
Xie, Yang	6	Zhao, Hongyu	35
Xin, Xue	Poster	Zhao, Jing	85
Xiong, Chengjie	19, 52	Zhao, Weizhi	9
Xiong, Xiaoping	42, 80	Zhao, Xin	75
Xu, Chuanbo	68	Zhao, Yanli	94
Xu, Lei	28	Zheng, Tian	38
Xu, Yaomin	6	Zhi, Xin	Poster
Xu, Yu	31	Zhong , Jinglin	12
Xu, Zongli	Poster	Zhou, Haibo	34, 91
Yakovlev, Andrei Y	66	Zhou, Hongling	33
Yan, Guofen	89	Zhou, Hui	39
Yan, Jun	9	Zhou, Lan	56
Yang, Yang	24	Zhou, Mai	60
Yang, Ying	Poster	Zhou, Rong	17
Yeo, Adeline	80	Zhou, Yong	34
Yi, Nengjun	62, 78	Zhu, Haoyuan	68
Yi, Yeonjoo	Poster	Zhu, Hongtu	44
Yin, Guosheng	51, 68	Zhu, Jun	25, 64
Ying, Gui-shuang	29	Zhu, Li	83
Ying, Zhiliang	60	Zhu, Xiaoning (Joanie)	Poster
Young, Linda	16, 45	Zhu, Yiliang	77
Young, S. Stanley	72	Ziegler, Kathryn A	1
Yu, Jihnhee	54	Zou, Fei	19, 62, Poster
Yu, Kai	19, 35	Zubovic, Yvonne M	33
Yu, Menggang	51	Zubrow, Alexis	15
Yu, Zhuo	9		
Yuan, Jinyu	35		
Yuan, Ming	85		
Yuan, Zheng	3		
Yuhong, Yang	48		
Zaborsky , Larry	14		
Zaslavsky, Alen	13		
Zaykin, Dmitri V	19		
Zeger, Scott L	1		
Zelen, Marvin	26		
Zelerman, Daniel	29		
Zeng, Donglin	17		
Zeng, Leilei	92		
Zhang, Chun	8		
Zhang, Cuihui	6		
Zhang, Daowen	94		
Zhang, Guang Wen	Poster		
Zhang, Hao Helen	29, 43, 90		



Duxbury...

Innovation and Dedication in Statistics

TEXTBOOKS FOR BIOSTATISTICS



Introductory Applied Biostatistics
Preliminary Edition
RALPH D'AGOSTINO,

SOFTWARE AND HANDBOOKS

JMP IN®
New JMP IN® Version 5, Release 5.1!
0-534-99747-3
JMP™: A BUSINESS UNIT OF SAS



PLEASE USE THE FILM

JESSICA UTTS and
ROBERT HECKARD
0-534-39305-5

Updated for Office XP
MICHAEL R. MIDDLETON
0-534-40293-3

Finally, you can test the way you teach!



For Utts and Heckard's *Mind on Statistics, Second Edition*, and selected other titles, iLrn with DuxStat offers:

- Automated grading for homework, quizzing, and testing
- Over 90% of the kinds of exercises and test problems you use in your class
- The DuxStat Data Analysis Tool, an easy-to-use Java statistical analysis applet that is fully integrated into questions

Visit <http://www.duxstat.com> for a demonstration of these powerful tools!

Excel Essentials:
Using Microsoft Excel for Data Analysis and Decision Making
(Video Tutorials)

A.J. HAMISH TAYLOR
0-534-39309-8

DUXBURY

THOMSON
★
BROOKS/COLE

Visit the Duxbury table at ENAR for more information about our latest titles and teaching tools!

5BCSTENR

To request a review copy, please call (800) 423-0563 or visit our Web site: <http://www.duxbury.com>.

S
T
A
T
I
S
T
I
C
S

THOMSON
★
BROOKS/COLE

Clinical Trials

Journal of the Society for Clinical Trials

EDITOR: Steven Goodman, MD, PhD, Johns Hopkins School of Medicine, USA

DEPUTY EDITOR:

Constantine Frangakis, Johns Hopkins University, School of Public Health, USA

We are pleased to announce that starting in January 2004, Clinical Trials: Journal of the Society for Clinical Trials (instead of Controlled Clinical Trials) will be the official journal of the Society for Clinical Trials, and will be published by Arnold Journals. Clinical Trials aims to be the primary focus for the dissemination and development of knowledge about the design, conduct, analysis, synthesis, history, ethics, regulation and clinical or policy impact of all types of clinical trials and related medical research methodologies.

Submissions in all these areas are encouraged.

To submit papers to the official SCT journal please email your paper, as a pdf or word file if possible, to clinicaltrials@jhmi.edu or contact: The Editor: Steven Goodman at: sgoodman@jhmi.edu

Before submitting your manuscript please consult the full information for authors guide at www.sctjournal.com

Members of the Society for Clinical Trials will receive 'Clinical Trials', the official

SCT journal, as part of their membership from 2004. Member subscriptions will include electronic access as well as print copy



www.sctjournal.com



Try our journals for FREE!
- contact us today at
ejournal@hodder.co.uk

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

NOTES

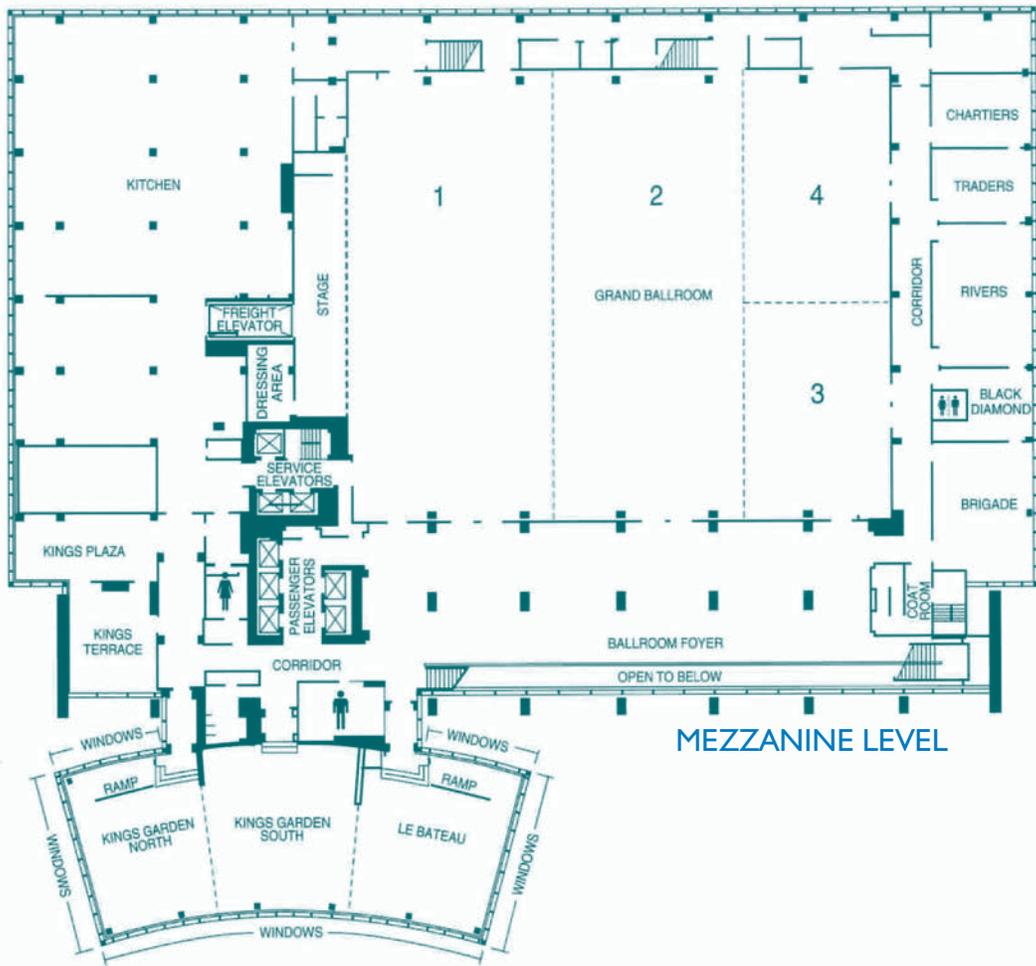
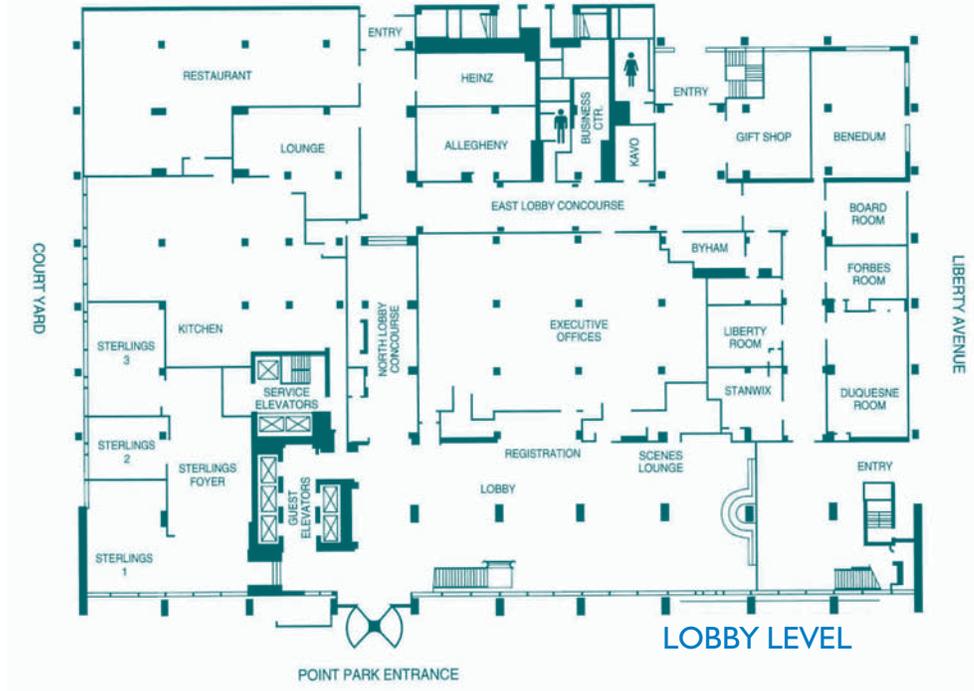
ENAR WEBSITE
WWW.ENAR.ORG

NOTES

ENAR WEBSITE
WWW.ENAR.ORG

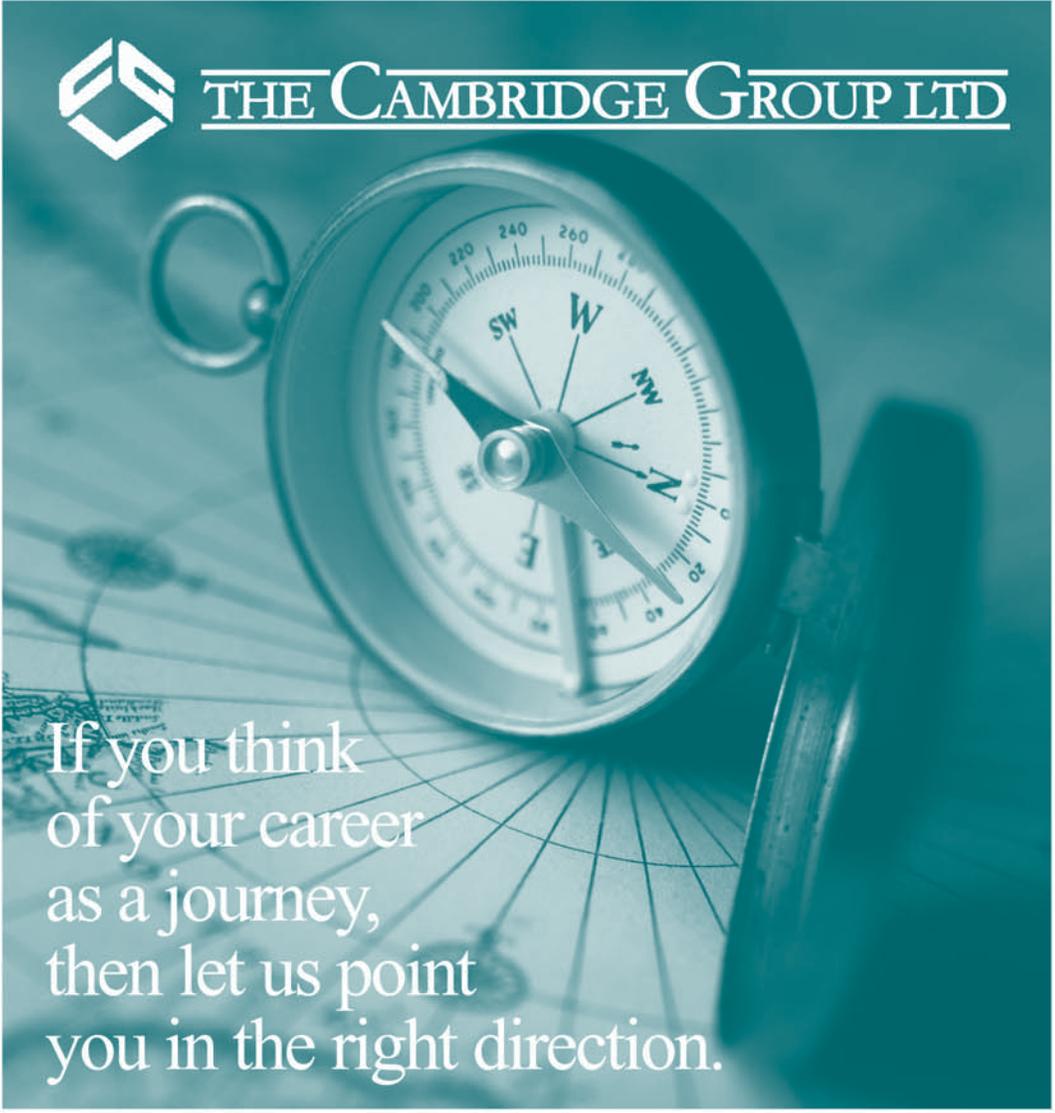
HILTON PITTSBURGH FLOOR PLAN

Executive Center at the Pointe





THE CAMBRIDGE GROUP LTD



If you think
of your career
as a journey,
then let us point
you in the right direction.

▶ **Biostatistics**

biostat@cambridgegroup.com

▶ **Clinical Computing
& Data Management**

QA@cambridgegroup.com

▶ **Consultant
& Contract Staffing**

contract@cambridgegroup.com

For over 25 years,
we've been pointing
to success.



THE CAMBRIDGE GROUP LTD

Your Success is Our Business

(800) 525-3396

www.cambridgegroup.com