



**MARCH 16–19, 2014**

Baltimore Marriott Waterfront Hotel  
Baltimore, Maryland

# ABSTRACTS & POSTER PRESENTATIONS

ENAR  
2014  
SPRING  
MEETING

*with IMS and Sections of ASA*

## 1. POSTERS: INVITED POSTER SESSION

### 1A. SuBLIME and OASIS for Multiple Sclerosis Lesion Segmentation in Structural MRI

**Elizabeth M. Sweeney\***, Johns Hopkins Bloomberg School of Public Health

**Russell T. Shinohara**, University of Pennsylvania

**Ciprian M. Crainiceanu**, Johns Hopkins Bloomberg School of Public Health

Magnetic resonance imaging (MRI) can be used to detect lesions in the brains of multiple sclerosis (MS) patients and is essential for evaluating disease-modifying therapies and monitoring disease progression. In practice, lesion load is often quantified by expert manual segmentation of MRI, which is time-consuming, costly, and associated with large inter- and intra- observer variability. We propose Subtraction-Based Logistic Inference for Modeling and Estimation (SuBLIME) and OASIS is Automated Statistical Inference for Segmentation (OASIS). SuBLIME is an automated method for segmenting incident lesion voxels between baseline and follow-up MRI studies. OASIS is an automated method for segmenting lesion voxels from a single MRI study. Both methods use carefully intensity-normalized T1-weighted, T2-weighted, fluid-attenuated inversion recovery (FLAIR) and proton density (PD) MRI volumes and are logistic regression models trained on manual lesion segmentations. We also present software implementations of SuBLIME and OASIS, where users can upload MRI studies to a website to produce lesion segmentations within minutes.

email: emsweene@jhsph.edu

### 1B. Elastic Statistical Shape Analysis of 3D Objects Using Square Root Normal Fields

**Sebastian Kurtek\***, The Ohio State University

We present a new Riemannian framework for comprehensive statistical shape analysis of 3D objects, represented by their boundaries. By comprehensive framework, we mean tools for registration, comparison, averaging, and modeling of observed surfaces. Registration is analogous to removing shape preserving transformations, which include translation, scale, rotation and re-parameterization. This framework is based on a special representation of surfaces termed square root normal fields and a closely related elastic metric. The main advantages of this method are: (1) the elastic metric provides a natural interpretation of shape deformations that are being quantified; (2) this metric is invariant to re-parameterizations of surfaces; (3) under the square root normal field transformation, the complicated elastic metric becomes the standard L2 metric, simplifying parts of the implementation. We present numerous examples of shape comparisons for various types of surfaces in different application areas including biometrics, medical imaging, and graphics. We also compute average shapes, covariances and perform principal component analysis. These quantities are used to define generative models and for random sampling.

email: kurtek.1@stat.osu.edu

### 1C. Epidemiological Models for Browser-Based Malware

**Natallia Katenka\***, University of Rhode Island

**Eric Kolaczyk**, Boston University

**Mark Crovella**, Boston University

**Tom Britton**, Stockholm University

The spread of the computer viruses over the Internet results in the losses of billions of dollars and an exposure of personal and highly classified information. Unlike previous research in the Internet security, our study (1) focuses specifically on epidemic models for browser-based malware threats, and (2) accounts for the structure of the web communication network. To infer and characterize client-server communication network that is specific for web application, we use the NetFlow traffic data collected at core networks from a large European Internet Service Provider (ISP). We adopt simple probability-based transmission models and explore dynamically vulnerability of the web-network to the existing and soon-to-be anticipated threats. In addition, by utilizing game theory, we investigate simple proactive strategies that would help to control and prevent an epidemic outbreak under different attack strategies. Our main findings suggest that the key properties of web-networks, namely heavy tail degree distribution, temporal, and strength/degree distribution prevent exponential growth of the infection among both clients and servers. In fact, selective communication behavior is a key to security. We also found that in order to protect clients, we need to defend high and medium degree servers; and in order to protect servers, we need to protect high and medium degree clients.

email: nkatanka@cs.uri.edu

### 1D. Meta-Analysis of Rare Variants Based on Single-Variant Statistics

**Yijuan Hu\***, Emory University

**Sonja I. Berndt**, National Cancer Institute, National Institutes of Health

**Stefan Gustafsson**, Uppsala University Hospital

**Andrea Ganna**, Uppsala University Hospital

**Joel Hirschhorn**, Boston Children's Hospital

**Kari E. North**, University of North Carolina, Chapel Hill

**Erik Ingelsson**, Uppsala University Hospital

**Danyu Lin**, University of North Carolina, Chapel Hill

There is a growing recognition that identifying 'causal' rare variants requires large-scale meta-analysis. The fact that association tests with rare variants are performed at the gene level rather than at the variant level poses unprecedented challenges in the meta-analysis. First, different studies may adopt different gene-level tests, so the results are not compatible. Second, gene-level tests require multivariate statistics, which are difficult to obtain. To overcome these challenges, we propose to perform gene-level tests for rare variants by combining the results of single-variant analysis (i.e., p-values of association tests and effect estimates) from participating studies. We show both theoretically and numerically that the proposed meta-analysis approach provides accurate control of the type I error and is as powerful as joint analysis of individual participant data. This approach accommodates any disease phenotype and any study design and produces all commonly used gene-level tests. An application to the GWAS summary results of the Genetic Investigation of Anthropometric Traits (GIANT) consortium reveals rare and low-frequency variants associated with human height. The relevant software is freely available.

email: yijuan.hu@emory.edu

## 1e. Spatial Quantile Regression for Neuroimaging Data

**Linglong Kong\***, University of Alberta  
**Hongtu Zhu**, University of North Carolina, Chapel Hill

Neuroimaging studies aim to analyze imaging data with complex spatial patterns in a large number of locations (called voxels) on a two-dimensional (2D) surface or in a 3D volume. We propose a multiscale adaptive composite quantile regression model (MACQRM) that has four attractive features: being robustness, being spatial, being hierarchical, and being adaptive. MACQRM utilizes imaging observations from the neighboring voxels of the current voxel and borrows strength from the nearby quantile regressions of the current regression to adaptively calculate parameter estimates and test statistics. Theoretically, we establish consistency and asymptotic normality of the adaptive estimates and the asymptotic distribution of the adaptive test statistics. Our simulation studies and real data analysis on ADHD confirm that MACQRM significantly outperforms MARM and conventional analyses of imaging data.

email: lkong@ualberta.ca

## 1f. Enhancements for Model-Based Clustering of Array-Based Dna Methylation Data

**Devin C. Koestler\***, University of Kansas Medical Center  
**Andres Houseman**, Oregon State University  
**Carmen J. Marsit**, Dartmouth College  
**Brock C. Christensen**, Dartmouth College

Motivated by idea of CpG Island Methylator Phenotypes (CIMPs), unsupervised clustering has emerged as one of the most popular techniques for the analysis of array-based DNA methylation data. While there are a profusion of different clustering methodologies, model-based clustering methods based on a finite mixture of underlying probability distributions have risen to the forefront, gaining acceptance as one of the premier methods for clustering array-based DNA methylation data. The recently proposed Recursively Partitioned Mixture Model (RPMM), a hierarchical model-based clustering methodology, is characterized by its computational efficiency and robustness for determining the number of clusters. Although RPMM has proved to be a successful strategy, an increasing understanding of the biology of DNA methylation and the increasing resolution and coverage of DNA methylation microarrays, necessitates refinements to the existing method. In this presentation, we describe recent advances in the RPMM methodology, with a specific emphasis on its application to DNA methylation microarray data. In addition, a summary of future directions and new avenues for this work will be discussed.

email: dkoestler@kumc.edu

## 1g. Disease Surveillance Using Dynamic Screening System

**Peihua Qiu\***, University of Florida

In the SHARe Framingham Heart Study of the National Heart, Lung and Blood Institute, one major task is to monitor several health variables (e.g., blood pressure and cholesterol level) so that their irregular longitudinal pattern can be detected as soon as possible and some medical treatments can be applied in a timely manner to avoid some deadly cardiovascular diseases (e.g., stroke). To handle this kind of applications effectively, we propose a new statistical methodology called dynamic screening system (DySS). The DySS method combines the major strengths of the multivariate longitudinal data analysis and the multivariate statistical process control, and it makes decisions about the longitudinal pattern of a subject by comparing it with other subjects cross-sectionally and by sequentially monitoring it as well. Numerical studies show that it works well in practice.

email: pqiu@ufl.edu

## 1h. Heat Kernel Wavelets on Manifolds and its Application to Brain Imaging

**Moo K. Chung\***, University of Wisconsin, Madison

Starting with a symmetric positive definite kernel as a basic building block, we show how to construct a new wavelet transform on an arbitrary manifold. This particular wavelet transform is shown to be equivalent to the weighted Fourier series (WFS) representation (Chung et al., 2007), which was originally introduced as a generalization of the traditional Fourier series expansion. The proposed framework offers probably the most simplistic and unified framework for constructing wavelets on a manifold and shown to be related to kernel regression and heat diffusion. Wavelet transform is a powerful tool decomposing a signal or function into a collection of components localized at both location and scale. The proposed wavelet transform applied in investigating the influence of age and gender on amygdala and hippocampus in human brain MRI. We detected a significant age effect on the posterior regions of hippocampi while there is no gender effect present in any of the structures.

email: mkchung@wisc.edu

## 1i. Data Visualizations Should be More Interactive

**Karl W. Broman\***, University of Wisconsin, Madison

The value of interactive graphics for making sense of high-dimensional data has long been appreciated but is still not in routine use. Interactive graphics facilitate data exploration, are great collaborative tools, allow compressed summary plots to be linked to the underlying details, and can be fabulous teaching tools. New web-browser-based tools, such as the JavaScript library D3, have greatly simplified the development of interactive visualizations. A number of R packages, including Shiny and rCharts, provide customizable interactive visualizations directly from R. I will provide a number of examples to illustrate the value of interactive graphics, with live demonstrations on handheld devices.

email: kbroman@biostat.wisc.edu

## 1J. Introducing the Evolving Evolutionary Spectrum, with Applications to a Learning Association Study

**Mark Fiecas\***, University of Warwick

We develop a novel approach to analyzing nonidentical nonstationary bivariate time series data. We consider two sources of nonstationarity: 1) within each replicate and 2) across the replications, so that the spectral properties of the bivariate time series data are evolving both over time within a replicate and also over the replicates in the experiment. We propose a novel model and a corresponding two-stage estimation method. In the first stage we account for nonstationarity within a replicate by using local periodogram matrices. In the second stage, we account for nonstationarity over the replications. We apply the method to a local field potential data set to study how the coherence between the nucleus accumbens and the hippocampus evolves over the course of a learning association experiment.

email: M.Fiecas@warwick.ac.uk

## 1k. Improving Rare Variant Association Test with Prior Information

**Xin He\***, Carnegie Mellon University

**Li Liu**, Carnegie Mellon University

**Bernie Devlin**, University of Pittsburgh  
School of Medicine

**Kathryn Roeder**, Carnegie Mellon University

A number of rare variant association tests have been proposed to map the causal genes of complex traits from association data. In this work, we attempt to leverage prior information in the statistical test. A large amount of biological information is publicly available that may provide measure of functionality and other useful information of genetic variants, yet, most existing methods are not designed to take advantage of such resources. We built a flexible Bayesian framework to test gene-level association that can easily accommodate additional information of variants. Specifically our model takes three types of information: how likely a variant is detrimental based on their annotations and bioinformatic predictions; how often a variant occurs in large population samples (independent of the association data being analyzed); and how strongly the gene containing the variant is under selective constraint. We re-analyzed the recently published autism case-control data using our new method, with the prior information derived from PolyPhen2 and NHLBI Exome Sequencing Project. Our new method significantly improves the detection of autism risk genes than existing methods.

email: xinhe2@gmail.com

## 2. POSTERS: CLINICAL TRIALS AND STUDY DESIGN

### 2a. A New Statistical Test of Heterogeneity in Treatment Response

**Hongbo Lin\***, Indiana University, Indianapolis

**Changyu Shen**, Indiana University, Indianapolis

Randomized studies are designed to estimate the average treatment effect (ATE); however, individuals may derive quantitatively, or even qualitatively, different effects from the ATE. It is important to detect if heterogeneity exists in the treatment responses. We propose a method to test the hypothesis that the treatment has no effect on each of the smallest sub-populations, which are defined by discrete baseline covariates using randomized trial data. Our approach is nonparametric, which generates the null distribution of the test statistic by either the bootstrap or permutation principle. A key innovation of our method is that stochastic search is built into the test statistic to detect signals that may not be linearly related to the multiple covariates. This is important because in many real clinical problems, the treatment effect is not linearly correlated with relevant baseline characteristics. Simulations were performed to compare the proposed test with existing methods. We also applied the method to a real randomized study that compared the Implantable Cardioverter Defibrillator (ICD) with conventional medical therapy in reducing total mortality in a low ejection fraction population.

email: lin53@uemail.iu.edu

### 2b. Comparing Methods of Tuning Adaptively Randomized Trials

**John Cook**, University of Texas MD

Anderson Cancer Center

**Yining Du\***, University of Texas MD

Anderson Cancer Center

**Jack Lee**, University of Texas MD Anderson Cancer Center

The simplest Bayesian adaptive randomization (SAR) scheme is to randomize patients to a treatment with probability equal to the probability  $p$  that the treatment is better. We examine three variations on adaptive randomization which are used to compromise between this scheme and equal randomization (ER) by varying the tuning parameter. The first variation is to apply a power transformation to  $p$  to obtain randomization probabilities. The second is to clip  $p$  to live within specified lower and upper bounds. The last is to begin the trial with a burn-in period of equal randomization. In each method, statistical power increases as one gets closer to ER, while the proportions of patients treated on the superior arm and overall response rate increase as one gets closer to SAR. Absent a stopping rule, the power transformation method puts the most patients on the better arm given the power. With the stopping rule, the power transformation method appears to do roughly equally well with the clipping approach but requires fewer patients in the trial, with the burn-in method doing worse. We would recommend the power transformation method without an early stop and could also consider the clipping method when adopting a stopping rule.

email: yining.du@yahoo.com

## 2c. Multi-Regional Issues in Equivalence Assessment of Test and References

**Yi Tsong\***, U.S. Food and Drug Administration

In a multiple regional clinical trial designed to assess the equivalence of a test and reference treatments, one may need to design and analysis multi-regional trial with regional regulatory requirements on study population and reference treatment and even significance levels. Assuming for two regions, with the combinations of the requirements, we may have a clinical trial designed to have two arms (test and reference 1+reference 2 arm) and three-arms (test arm, reference 1 arm and reference 2 arm). Subjects of the two regions may be randomized into two arms of the two-arm trial with region 1 subjects receiving test or reference 1 treatment in the two arms. Subjects of region 2 are to receive test or reference 2 treatments. It leads to an equivalence test of single equivalence hypothesis. The question is should we use one or two significance levels to show equivalence. When we fail to show equivalence of test and the combined references in the combined population, do we proceed to test for the regions separately? If so, how do we adjust the type 1 error rate? On the other hand, for the three-arm trial, should we randomize subjects of both regions to the three arms or should we randomize subjects to region specific arms instead. In the first case, should we test equivalence using different significant levels but adjust for multiple comparisons? What about in the second case?

email: yitsong0115@gmail.com

## 2d. Statistical Methods for Analyzing Count Data—A Case Study on Adverse Event Data from Vaccine Trials

**Qin Jiang\***, Pfizer Inc.

In controlled clinical studies, count data are often used as outcome variables, such as number of adverse events reported or number of episodes of certain symptoms confirmed during the study defined period. Since these count data are not normally distributed t-tests, ANOVA, or ANCOVA can produce biased estimate. Therefore, Poisson regression has been widely used as alternative method for these count data. However, as data are frequently characterized by overdispersion and excess zeros additional methods to model the count data should be considered. This presentation applies clinical data to fit Poisson, Negative Binomial (NB), zero-inflated Poisson (ZIP), zero-inflated Negative Binomial (ZINB), Poisson hurdle model (PH) or zero-truncated Poisson model, Negative Binomial hurdle model (NBH). The differences between these distributions are illustrated and the results from these models are compared and discussed.

email: qin.jiang@pfizer.com

## 2e. An Alternate Study Design Approach for Multilevel Counts Subject to Overdispersion, with Illustrations Reflective of a Motivating Cluster-Randomized Community Trial

**Kenneth J. Wilkins\***, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health and Uniformed Services University of the Health Sciences

**Shweta Padmanaban**, Georgetown University

**Stephanie M. Rodriguez**, Uniformed Services University of the Health Sciences

Numerous biomedical studies assess the impact of interventions assigned on an aggregate rather than individual level. While commercial software readily accommodates clustered continuous and binary outcomes in designing such studies, instances using multilevel counts subject to overdispersion do not seem as directly addressed. We evaluate how alternate methods for determining sample size and statistical power compare in designing studies that target population-average (PA) intervention effects on overdispersed counts. Specifically, we exploit how specific regression coefficients retain a PA interpretation when assuming a (possibly overdispersed) Poisson model using a log link for the mean, and normally distributed random effects (e.g., Young, et al., 2007). This investigation is motivated by a trial whose stakeholders urged an investigation into whether its null findings were a consequence of its PA design/analysis approach. After trial initiation in 2010, Molenberghs and colleagues proposed an analysis approach to multilevel outcomes similarly parametrized, yet in combination with conjugate random effects to model overdispersed non-Gaussian outcomes. We illustrate how off-the-shelf software calculations compare to this combination model when aggregate-level data are available prior to design.

email: kwilkins@usuhs.edu

## 2f. Leveraging Baseline Variables to Improve Estimators of the Average Treatment Effect in Randomized Trials

**Elizabeth A. Colantuoni\***, Johns Hopkins Bloomberg School of Public Health

**Michael Rosenblum**, Johns Hopkins Bloomberg School of Public Health

There has been much controversy over the proper use of baseline variable information in analyzing the results of randomized trials. Although randomization ensures balance in the distribution of baseline variables on average if a trial were to be repeated many times, chance imbalance in baseline variables may occur in the actual, observed trial. Further, if baseline variables are correlated with the outcome then adjusting for these can provide gains in the precision of the estimated average treatment effect. In this work, we build on the estimators proposed by Rubin and van der Laan (2008), Rotnitzky et al (2012) and Gruber and van der Laan (2012) that account for baseline variables, and are asymptotically as or more precise than the standard unadjusted estimator. Our main contribution is a simplified procedure that reduces the computational complexity of the recent estimators while maintaining their key advantages.

We apply the estimators to two recent randomized trials and evaluate their properties via a simulation study. Given the large potential gains and minimal costs, we recommend a preplanned adjustment for a small set of a priori identified baseline variables in clinical trials.

email: elizabethcolantuoni@gmail.com

## 2g. Identifying Comparable Populations using Entropy Matching: The Comparison of Drug Effectiveness Between Clinical Trials and EMRs

**Haoda Fu**, Eli Lilly and Company  
**Jin Zhou\***, University of Arizona

Electronic medical records (EMRs) have been promoted as essential to improving healthcare quality and mining EMR data is a cost-effective way to estimate drug efficacy compared to use randomized clinical trials (RCT). However, evidence consistency of drug efficacy estimation is a concern. Specifically, the study of electronic records is referred to as one type of non-randomized studies (NRS) or observational studies. Bias can be introduced due to various confounding elements compared to RCT. In this paper, we modified an entropy-matching algorithm in political science to identify comparable populations from EMR to RCT. In simulation studies, comparing to traditional propensity score matching entropy-matching algorithm provides a more accurate treatment effect estimation. Finally, based on identified population we compared drug efficacy of diabetes.

email: jzhou@email.arizona.edu

## 2h. Re-Estimating Sample Size in a Randomized Clinical Trial Using Participant Compliance Data

**Peter D. Merrill\***, University of Alabama, Birmingham  
**Leslie A. McClure**, University of Alabama, Birmingham

Participants' compliance to assigned treatment is an issue that must be considered in any randomized clinical trial. Non-compliance in a study leads to a bias towards the null in superiority studies, leading to a reduction in power. One method of dealing with this issue is to include an estimate of participant compliance in the power calculation; however, new issues may arise if an inaccurate estimate of compliance is used. Internal pilot (IP) methods have been developed to deal with similar issues of misspecification of the variance in power calculations. After a set proportion of the initial sample size has been collected, the data collected are used to estimate the sample variance, which is then used to re-estimate the sample size. We examined the impact of re-estimating the sample size part-way through a trial based on estimates of the compliance from the IP sample. Using simulations, we assessed the impact of misspecifying the compliance during the planning phase of the trial and assessed the performance of the method. Across all combination of planning and estimated compliance, as well as for different IP sizes, the desired power was achieved, while small inflations of type I error rate were observed.

email: pdmerr@uab.edu

## 2i. Variable Group Sizes in Cluster Randomized Trials Reduces Power

**Stephen A. Lauer\***, University of Massachusetts, Amherst  
**Nicholas G. Reich**, University of Massachusetts, Amherst  
**Ken P. Kleinman**, Harvard Medical School and Harvard Pilgrim Health Care Institute

When randomizing each individual to a study arm is unethical or impossible, a cluster-randomized trial (CRT) design maintains some strengths of a randomized study design. Examples include interventions on physicians or hospitals. Instead of randomizing individuals, a CRT randomizes clusters of subjects, e.g., patients of a given physician to the same study condition. CRT designs are common in public health research. The CRT is a potent tool; however it is not without drawbacks. Power is affected by the correlation within each cluster, as well as the number of clusters and the number of subjects. Existing formulas to estimate power in CRTs assume equal subjects per cluster. Using the clusterPower package in R, we conducted a simulation study to assess the impact of cluster size variability on the statistical power of CRTs across a wide-range of parameters. Our goals were to illustrate how these features affect power and to show the utility of a simulation-based power calculation methodology. Our simulation study shows that increased variability in cluster sizes reduces statistical power. We derive concrete guidelines that can assist in the design of future CRTs.

email: stephenalauer@gmail.com

## 2j. CoGaussian Statistical Model for Right Skewed Data

**Govind Mudholkar**, University of Rochester  
**Ziji Yu\***, University of Rochester  
**Saria Awadalla**, University of Illinois, Chicago

Scientific data are often nonnegative, right skewed and unimodal. For such data, CoGaussian distribution, the R-symmetric Gaussian twin, with its mode as the centrality parameter, is a basic model. In this talk, the essentials, namely the concept of R-symmetry, the roles of the mode and harmonic variance as, respectively, the centrality and dispersion parameters, the pivotal role of the CoGaussian family in the class of R-symmetric distributions and the associated method of moments and maximum likelihood estimation results, are introduced. Also, applications of the CoGaussian distribution in areas such as Sequential Analysis and Survival Analysis are generally described.

email: ziji\_yu@urmc.rochester.edu

## 2k. Use of Historical Data in Clinical Trials

**Kert Viele\***, Berry Consultants

Clinical trials rarely, if ever, occur in a vacuum. Generally large amounts of clinical data are available prior to the start of a study, particularly on the current study control arm. There is obvious appeal in using (i.e. borrowing) this information. With historical data providing information on the control arm, more trial resources can be devoted to the novel treatment while retaining accurate estimates of

the current control arm parameters. This can result in more accurate point estimates, increased power, and reduced type I error in clinical trials, provided the historical information is sufficiently similar to the current control data. If this assumption of similarity is not satisfied, however, one can acquire increased mean square error of point estimates due to bias, and either reduced power or increased type I error depending on the direction of the bias. We review several methods for historical borrowing, illustrating how key parameters in each method affect borrowing behavior, and then we compare these methods on the basis of mean square error, power, and type I error. Our goal is to provide a clear review of the key issues involved in historical borrowing and provide a comparison of several methods useful for practitioners.

email: kert@berryconsultants.net

### 3. POSTERS: BAYESIAN METHODS

#### 3a. A Bayesian Approach to ROC Curve Estimation using Conditional Means Priors

**Jack S. Knorr\***, Baylor University  
**John W. Seaman**, Baylor University

The summary receiver operating characteristic (SROC) curve is frequently used to model the accuracy of a diagnostic test. The SROC curve was developed as a frequentist meta-analytic model, but has since been considered from a Bayesian perspective. It is often desirable in a Bayesian model to use relatively non-informative priors. However, transformation of variables within the model may unintentionally produce very informative priors. We review the Bayesian approach to constructing SROC curves. We demonstrate the dangers of using diffuse priors on the associated regression parameters. We then present a conditional means prior (CMP) distribution on the parameters of interest. This CMP model is compared to the diffuse prior model through their application to a data set: a set of studies estimating the accuracy of a regular duplex ultrasound for diagnosing peripheral artery stenosis.

email: jack.knorr@gmail.com

#### 3b. Benchmark Dose Model Averaging in Toxicology

**Otis R. Evans\***, University of North Carolina, Wilmington

Benchmark dose estimation is an important part of toxicology. Most current methods assume one risk model as a sufficient estimator of Benchmark Dosage. However, if a different model is selected, the benchmark dose can change drastically. Here in, we develop a bayesian model averaging that utilizes several different risk models. The methodology will be showcased through several simulations.

email: ore7005@uncw.edu

#### 3c. Dirichlet Process Mixture Extension Model to Accommodate Complex Sample Designs for Linear and Quantile Regression

**Xi Xia\***, University of Michigan  
**Michael Elliott**, University of Michigan

Complex survey sample designs typically have unequal inclusion probabilities for sampled elements, due to unequal probabilities of selection, to account for unit non-response, and/or to calibrate to known population totals. Typically weighted estimate with weights equal to inverse of the inclusion probabilities is applied to correct for bias introduced by correlations between inclusion probabilities and sampled data. However, if weights are uncorrelated with the quantity of interest, or extreme weight values occur due to highly variable sample weights, the weighted estimator could introduce extra variability, and result in an overall worse root mean square error (RMSE) comparing to unweighted methods. In this manuscript, we follow a Bayesian model-based approach that treats unobserved elements as 'missing data', and posits a mixture model to develop a predictive posterior distribution for inference on quantity of interest. More specifically, we adapt the Dirichlet Process Mixture Extension Model by Dunson (2007), which allows the mixture distribution to depend upon the covariates and leads to models with computationally stable complexity. We consider the repeated sampling properties of linear and quantile regression estimators in the presence of both model misspecification and informative sampling. We also estimate associations between dioxin levels and age using data from NHANES.

email: xiaksi@umich.edu

#### 3d. On Bayesian Model Selection for Robust Likelihood-Free Methods Based on Moment Conditions

**Cheng Li\***, Northwestern University  
**Wenxin Jiang**, Northwestern University

In biostatistics, an important practice is to use robust likelihood-free methods, such as the estimating equations, that only require assumptions on moments, without assuming the full probabilistic model. We propose a Bayesian-flavored model selection approach for such likelihood-free methods, based on (quasi-) posterior probabilities constructed from the Bayesian Generalized Method of Moments (BGMM). This novel concept allows us to incorporate two important advantages of the Bayesian approach: the expressiveness of posterior distributions and the convenient MCMC computational methods. Many different applications are possible, including modeling correlated longitudinal data, quantile regression, and graphical models based on partial correlation. We demonstrate numerically how this method works in these applications. In addition, under mild conditions, we show theoretically that the BGMM can achieve posterior consistency for selecting the unknown true model, and that it possesses the oracle property that the posterior

distribution for the parameters of interest is asymptotically normal and as informative as if the true model were known. These remain true even when the dimension of parameters diverges with the sample size and the true parameter is possibly sparse.

email: chengli2014@u.northwestern.edu

### 3e. Robustness of Multilevel Item Response Theory Model to Outliers Using Normal/Independent Distribution On Both Random Effects and Outcomes

**Geng Chen\***, University of Texas School of Public Health  
**Sheng Luo**, University of Texas School of Public Health

Parkinson's disease (PD) is one of the most common movement disorders chronic progressive neurovegetative diseases. Clinical trials that search for the neuroprotective treatments usually measure multiple longitudinal outcomes of various types. Among the multivariate outcomes, there are three sources of correlation, i.e. inter-source (different measures at the same visit), intra-source (same measure at different visits) and cross correlation (different measures at different visits). The multilevel item response theory (MLIRT) model accounts for all three sources of correlations and provide more accurate estimation on the overall disease progression rate. It is demonstrated that the misspecification of random effects in generalized linear mixed models does not have much impact on the prediction and inference about covariate effects. However, our simulation study showed that under MLIRT model settings, the outliers on random effects greatly impact the estimates of model parameters. Normal/Independent (NI) distribution as a distribution family with heavy tails could be an approach that handles the outliers' problems on both random errors and random effects. Our proposed methods were evaluated by simulations and applied to real PD data. The results showed that the MLIRT model with NI distribution assumptions on both random effects and outcomes provided more accurate inference for multivariate longitudinal data.

email: geng.chen@uth.tmc.edu

### 3f. Bayesian Sample Size Determination for Informative Hypotheses

**Kristen M. Tecson\***, Baylor University  
**John W. Seaman**, Baylor University

Researchers often analyze data assuming models with constrained parameters. In Bayesian hypothesis testing, incorporation of inequality constraints on model parameters has lead to "informative hypotheses" and associated priors. It is well known that in the frequentist context, incorporating inequality constraints into hypothesis tests increases power and efficiency by eliminating the need for ad hoc testing. In Bayesian hypothesis testing, similar improvements are seen

in operating characteristics. The effect of sample size in these problems has received little attention. In this poster session, we investigate informative hypotheses using Bayesian methods. We explore Bayesian sample size determination techniques for a variety of settings in the informative hypothesis context, including order violations.

email: kristen\_rose@baylor.edu

### 3g. Block Total Response Designs: A Bayesian Approach

**Michelle S. Marcovitz\***, Baylor University  
**Damaraju Raghavarao**, Temple University  
**John W. Seaman**, Baylor University

Statisticians working in behavioral research have long been interested in estimating the occurrences of sensitive behaviors in a population. Research about the prevalence of sensitive behaviors such as cheating on a test, taking illegal drugs, and acting dishonestly in the workplace is usually conducted through a survey. A well-planned survey can encourage participants to answer truthfully, making it possible to obtain stable estimates for the proportions of people within the population who participate in the sensitive behavior. The block total response method developed by Raghavarao and Federer asks respondents to answer a mixture of related sensitive and non-sensitive 'yes' or 'no' questions in a single questionnaire. Participants report the total number of questions they would answer 'yes' to rather than letting the interviewer know which specific questions they have responded 'yes' or 'no' to. In this poster session, we explore a Bayesian approach to estimating proportions of people answering 'yes' to sensitive questions when a block total response design has been used to create questionnaires. We employ a Markov chain Monte Carlo approach to obtain posterior probabilities for the sensitive questions and consider induced priors and Bayesian sample size determination.

email: tua03619@gmail.com

### 3h. Priors And Sample Size Determination for Hurdle Models

**Joyce Cheng\***, Baylor University  
**John W. Seaman**, Baylor University  
**David Kahle**, Baylor University

Hurdle models are often presented as an alternative to zero-inflated models for count data with excess zeros. Hurdle models consist of two parts: a binary model indicating a positive response (the 'hurdle') and a zero-truncated count model. One or both sides of the model can be dependent on covariates, which may or may not overlap. We consider new prior structures for Bayesian analysis of such models. We also consider Bayesian sample size determination for these models. We apply our methods to a hypothetical sleep disorder study.

email: joyce\_cheng@baylor.edu

### 3i. Bayesian Models for Facility-Level Adverse Medical Device Event Rates Among Hospitalized Children

**Laura A. Hatfield\***, Harvard Medical School

**Vanessa Azzone**, Harvard Medical School

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

When medical devices malfunction the health consequences can be serious. The Food and Drug Administration (FDA) has called for enhanced post-market surveillance of medical devices, that is, as they are used in the community. Children are an especially important population for surveillance because they are often excluded from pre-market approval studies, yet still receive approved devices. Safety surveillance depends on accurate expected failure rates and sensitivity to detect deviations from these rates. Voluntary surveillance is subject to reporting biases, and active surveillance (mandatory reporting on all safety events by selected facilities), suffers from lack of generalizability. Thus, we turn to administrative data, which address both challenges, though it lacks detail about patients and devices. To estimate and predict rates of medical device-related problems in a broad population of hospitalized children, we study all pediatric inpatient admissions in Massachusetts. Using diagnosis codes, we identify admissions with an adverse medical device event (AMDE). Significant differences emerge between the patient populations and AMDE rates of community and pediatric specialty hospitals. We compare the ability of different statistical models to generate accurate quarterly predictions using either patient-level or facility-level data. Our hierarchical Bayesian models improve predictions by addressing issues related to ecological fallacy.

email: hatfield@hcp.med.harvard.edu

### 3j. Group Comparison of Pulsatile Hormone Times Series

**TingTing Lu\***, University of Michigan

**Timothy D. Johnson**, University of Michigan

Due to its oscillatory and pulsatile nature, analyzing hormone time series data is challenging and many model-based methods have been proposed over the years. Typically, analyses are performed in two stages. First, the number and locations of the episodic events are determined. Second, a model is fit to the data conditional on the number of pulses. However, errors occurring in the first step are carried over to second. In 2007, Johnson proposed the first fully Bayesian deconvolution model that jointly estimates both the number and locations of secretion events and admits a non-constant basal concentration. Thus, both pulsatile and oscillatory components of hormone secretion are simultaneously modeled. Furthermore, the model allows for variation in pulse, shape and size. However, the model cannot handle groups of subject and cannot compare secretion patterns between groups. In this paper we extend Johnson's model in two ways. First, we admit group comparisons of the underlying pulse driving mechanism; second, we model the pulse driving mechanism via a Cox process where the intensity function is not assumed constant as is assumed in Johnson (2007). We take a fully Bayesian hierarchical approach to estimate model parameters, and then compare results with a smoothing spline functional analysis approach.

email: ttlu@umich.edu

### 3k. A Bayesian Approach To Detecting Changes in the Visual System

**Raymond G. Hoffmann\***, Medical College of Wisconsin

**Edgar A. Deyoe**, Medical College of Wisconsin

Data: The Visual Field Map (VFM) is obtained by activating the visual cortex in the brain with a dynamic target presented to the subject. Changes in the visual system can be simulated with images that have different size wedges (0, 18, 27, 36, 45 and 90 degrees) removed from a circular disk which is presented to the subject's eye. The output of the visual system is a set of activated voxels (from 295 to 619) in the visual cortex determined by functional MRI, which then is used to induce a figure on a virtual circular retina using an  $(r, \theta)$  representation for the location. The virtual retina will have more points in the center as does the real retina. Methods: A Bayesian non-parametric spatial model, a spatial Dirichlet Process model, is used to model the ratio of two different images induced by the two different angular wedges. Felfand, Kottas and MachEachern (JASA, 2005) introduced a Dirichlet Process as a prior mixing distribution on the family of densities  $DP(\nu G)$ . Under the null hypothesis of no difference, the ratio of the two densities will have a constant posterior density. Deviations from this will be used to indicate the probability of a perturbed visual system.

email: rhoffmann@mcw.edu

### 3l. A Comparison of Mcmc And Variational Bayes Algorithms for 3D Log-Gaussian Cox Processes

**Ming Teng\***, University of Michigan

**Farouk S. Nathoo**, University of Victoria

**Timothy D. Johnson**, University of Michigan

Log-Gaussian Cox Processes (LGCP) are flexible models for fitting spatial point pattern data. In order to estimate the intensity function, a Bayesian model with implementation based on Markov chain Monte Carlo (MCMC) simulation from the posterior, proposed by Møller et al. (1998), is commonly used. However, for LGCPs, MCMC is slow to converge to the posterior distribution and mixing is slow thereafter. Møller et al. proposed the use of the Metropolis adjusted Langevin algorithm (MALA), which helps considerably with mixing. More recently, Coeurjolly and Møller (2013) proposed a Variational estimator for LGCPs. We consider both MALA and variational Bayes methods based on mean field approximations for fitting LGCP models to 3D point pattern data with subject specific covariates and spatially varying coefficients. The application of VB to LGCP models is made challenging by the non-conjugate structure of the model. To develop tractable solutions, we incorporate Laplace approximations within the VB framework (Wang and Blei, 2013) which leads to Gaussian variational approximations. We make comparison between MALA and VB in terms of statistical and computation efficiency. Simulation studies are used to evaluate efficiency and we apply the algorithms to an imaging study of Multiple Sclerosis lesion locations with subject specific covariates.

email: tengming@umich.edu

### 3m. A Bayesian Hierarchical Model for Estimating HIV Testing Hazard

**Qian An\***, Emory University  
**Jian Kang**, Emory University

Human immunodeficiency virus (HIV) infection or Acquired immunodeficiency syndrome (AIDS) is a severe infectious disease actively spreading globally. The occurrence of new HIV diagnoses over time among a population infected with HIV, i.e. HIV testing hazard, is an important parameter for public health. In this paper, we propose a Bayesian hierarchical model with two levels of hierarchy to estimate the HIV testing hazard using the annual AIDS and HIV diagnoses data. At level one, we model the latent number of HIV infections for each year using a Poisson distribution with the intensity parameter representing the HIV incidence rate. At level two, the annual number of AIDS and HIV diagnosed cases and all undiagnosed cases stratified by the HIV infections at different years are modeled using a multinomial distribution with parameters including the HIV testing hazard. We propose a new class of prior for the HIV incidence rate taking into account the temporal dependence of these parameters to improve the estimation accuracy. We develop an efficient posterior computation algorithm based on the adaptive rejection sampling technique. We demonstrate our model using simulation studies.

email: qan2@emory.edu

### 3n. Bayesian Inference of the Asymmetric Laplace Distribution With Partial Information

**Shiyi Tu\***, Clemson University  
**Min Wang**, Clemson University  
**Xiaoqian Sun**, Clemson University

We make Bayesian inference on the asymmetric Laplace distribution which is a very important distribution in the Bayesian quantile regression. It is shown that the posterior distribution under the Jeffreys prior is improper, so we need to consider other noninformative priors. A good choice is reference prior with partial information (RPPI), we check all the scenarios of RPPI, and finally obtained two priors which could lead to the proper posterior distribution. Furthermore, based on RPPI, we also can extend our prior to a more general form.

email: stu@clemson.edu

### 3o. An Efficient Bayesian Sampling Approach for Continuous Bayesian Network Structure Learning

**Shengtong Han\***, University of Memphis  
**Hongmei Zhang**, University of Memphis

Bayesian network has an appealing property of encoding dependence structure explicitly. Hence there is an increasing effort devoted to Bayesian network structure learning. However, learning Bayesian network is known to be N-P hard. Structure learning becomes even more difficult for networks with a large number of nodes mainly because of the huge structure space. More recently, sampling techniques based on order space are developed, which is tested to mix more quickly and more advanced sampling approach-equal energy sampler with the aim of alleviating

local maximum problem, together with order based MCMC has been developed for discrete Bayesian networks. This paper extends the use of equal energy sampling approach over order space to continuous Bayesian networks. Better performance is obtained compared to other existing inference algorithms from simulations.

email: shengtonghan@gmail.com

### 3p. Two-Sample Empirical Likelihood Based Tests for Mean: From Frequentists to Bayesian Type Techniques With Applications To Case-Control Studies

**Ge Tao\***, State University of New York at Buffalo  
**Albert Vexler**, State University of New York at Buffalo

Many clinical experiments are designed to be in a form of case-control studies for detecting discriminating ability of biomarkers or comparing treatment effects. Two-sample statistical tests are common procedures applied in such investigations. Avoiding parametric assumptions regarding data distributions, we consider different forms of empirical likelihood ratio tests to be employed in case-control studies. In this paper, we also proposed and evaluate novel two-sample empirical likelihood ratio based tests that involve Bayes factor type mechanisms in a nonparametric manner. The asymptotic properties of the proposed techniques are presented. We employ an extensive Monte Carlo study to evaluate the theoretical results as well as to compare the power of various two-sample tests for mean. The applicability of the proposed methods is demonstrated via a study associated with myocardial infarction disease.

email: getao@buffalo.edu

### 3q. Bayes Regularized Graphical Model Estimation in High Dimensions

**Suprateek Kundu\***, Texas A&M University  
**Bani Mallick**, Texas A&M University  
**Amin Momin**, University of Texas  
MD Anderson Cancer Research Center  
**Veera Baladandayuthapani**, University of Texas  
MD Anderson Cancer Research Center

There has been an intense development of Bayes graphical model literature over the past decade - however, most of the existing methods are restricted to moderate dimensions. We propose a novel approach scalable to high dimensional settings, by decoupling model fitting and covariance selection. First, a full model based on a complete graph is under a novel class of mixtures of inverse Wishart priors, which induce shrinkage on the precision matrix under an equivalence with Cholesky-based regularization, while enabling conjugate updates. Subsequently, we propose a post-fitting model selection step that uses penalized joint credible regions to perform neighborhood selection sequentially for each node. The posterior computation proceeds using straightforward fully Gibbs sampling. The proposed approach is shown to be asymptotically consistent in estimating the graph structure for fixed  $p$  when the truth

is a Gaussian graphical model. Simulations show that the proposed approach compares favorably with competitors in terms of model selection and computational efficiency. We apply our methods to high dimensional cancer genomics applications.

email: sk@stat.tamu.edu

## 4. POSTERS: STATISTICAL GENETICS AND GENOMICS

### 4a. LDA Topic Model of an Unknown Number of Topics Via MCMC

**Zhe Chen\***, University of Florida

**Hani Doss**, University of Florida

Latent Dirichlet Allocation (LDA) is a Bayesian topic model well known as a powerful technique for the document classification purpose by associating a document with a mixture of multiple topics. Unlike the traditional LDA with the number of topics pre-specified, we consider a situation where the number of topics is an unknown parameter which causes the dimensional changes of the component parameters in the model. The goal is to design an effective and efficient data-driven mechanism to decide an optimal range of values for the number of topics and to estimate the relevant component parameters. A novel "automatic" reversible jump MCMC algorithm, with no tunings of the proposal parameters and mapping rules as in a usual reversible MCMC, is designed to achieve the goal. By fully utilizing the mathematical properties of the relevant posterior distributions, our MCMC algorithm achieves a high mixing rate and provides an effective and easy-implemented way of simulating a chain of samples used for making the decision of the number of topics and estimation of components parameters. We illustrate our method on both synthetic and real data and report the good results that show the promising prospect of this method.

email: zhe.chen@ufl.edu

### 4b. Controlling The Local False Discovery Rate in the Adaptive LASSO

**Joshua N. Sampson\***, National Cancer Institute, National Institutes of Health

**Nilanjan Chatterjee**, National Cancer Institute, National Institutes of Health

**Raymond Carroll**, Texas A&M University

**Samuel Muller**, University of Sydney

The Lasso shrinkage procedure achieved its popularity, in part, by its tendency to shrink estimated coefficients to zero, and its ability to serve as a variable selection procedure. Using data-adaptive weights, the adaptive Lasso modified the original procedure to increase the penalty terms for those variables estimated to be less important by ordinary least squares. Although this modified procedure attained the oracle properties, the resulting models tend to include a large number of "false positives" in practice. Here, we adapt the concept of local false discovery rates (lFDRs) so that it applies to the sequence,  $\{p_n\}$ , of smoothing parameters for the adaptive Lasso. We define the lFDR for a given  $p_n$  to be

the probability that the variable added to the model by decreasing  $p_n$  to  $p_{n-1}$  is not associated with the outcome, where  $p_n$  is a small value. We derive the relationship between the lFDR and  $p_n$ , show lFDR = 1 for traditional smoothing parameters, and show how to select  $p_n$  so as to achieve a desired lFDR. We compare the smoothing parameters chosen to achieve a specified lFDR and those chosen to achieve the oracle properties, as well as their resulting estimates for model coefficients, with both simulation and an example from a genetic study of prostate specific antigen.

email: joshua.sampson@nih.gov

### 4c. Integrated Analysis of MicroRNA and Messenger RNA Expression Profiles of Essential Thrombocytosis

**Erya Huang\***, Stony Brook University

**Wei Zhu**, Stony Brook University

**Dmitri V. Gnatenko**, Stony Brook University

**Wadie F. Bahou**, Stony Brook University

Human blood platelets play essential roles in hemostasis and thrombosis. Although platelets are anucleate blood cells, they retain messenger RNAs (mRNAs) and microRNAs (miRNAs) derived from their precursor megakaryocyte. Overproduction of platelets can cause essential thrombocytosis (ET) disease, which oftentimes leads to blood clots and heart attack, or even death. miRNA and mRNA do not work separately, but in a complicated network of interactions. The analysis of the interactions would help understand the molecular basis of ET. In this project we have utilized general linear model and partial correlation network analysis to compare the gene expression profiles between a 13-member ET patient cohort and a 30-member normal cohort to identify putative joint miRNA and mRNA pathways that underlie the disease of ET.

email: e.huang@live.com

### 4d. The Power Comparison of the Haplotype-Based Collapsing Tests and the Variant-Based Collapsing Tests for Detecting Rare Variants in Pedigrees

**Wei Guo\***, National Institute of Mental Health, National Institutes of Health

**Yin Yao Shugart**, National Institute of Mental Health, National Institutes of Health

Background: Both common and rare genetic variants contribute to the etiology of complex diseases. Recent genome-wide association studies (GWAS) have successfully investigated how common variants contribute to the genetic factors associated with common human diseases. However, understanding the impact of rare variants, which are abundant in the human population (one in every 17 bases), remains challenging. A number of statistical tests have been developed to analyze collapsed rare variants identified by association tests. Here, we propose a weighted haplotype-based approach built on an existing statistical framework of the pedigree disequilibrium test (PDT); the

PDT uses sequencing data to assess the effects of rare variants in general pedigrees. Results: Extensive simulations in the sequencing setting were carried out to evaluate and compare the novel weighted haplotype-based PDT (hPDT) method with a rare variant PDT (rvPDT) that drew on a more conventional collapsing strategy. As assessed through a variety of scenarios, hPDT had enhanced statistical power compared with rvPDT when the variants were extremely rare  
email: wei.guo3@nih.gov

#### 4e. Functional Normalization (FUNNORM): A Better Alternative to Quantile Normalization for Methylation Data

**Jean-Philippe Fortin\***, Johns Hopkins University

**Aurélie Labbe**, McGill University

**Mathieu Lemire**, University of Toronto

**Brent W. Zanke**, Ottawa Hospital Research Institute

**Thomas J. Hudson**, University of Toronto

**Elana J. Fertig**, Sidney Kimmel Cancer Center at Johns Hopkins University

**Celia M.T. Greenwood**, McGill University

**Kasper D. Hansen**, Johns Hopkins University

Motivation: DNA methylation levels can be estimated with microarrays, but the signals require normalization to adjust for technical artifacts. Current methods for Illumina methylation arrays use different versions of quantile normalization, but none is completely appropriate for methylation data since global methylation profiles can vary substantially between subjects. Results: We present a new method, Functional Normalization (FunNorm), for preprocessing DNA methylation data from the Illumina Infinium HumanMethylation450 BeadChip, using a function-on-scalar correction model for background noise, dye bias, design bias and spatial effects. It is a novel approach containing adjustments that vary with the percentiles of the channel intensities and taking advantage of the different technical control probes. The key feature of the method is that contrary to methods based on quantile normalization, it does not rely on an assumption of distributional similarity across subjects. Our method is carefully evaluated against several other approaches for normalization, using large public cancer studies from the The Cancer Genome Atlas (TCGA) and using cross-validation with internal datasets. Our method is shown to be highly competitive for large-scale studies of a condition associated with large changes in DNA methylation, such as cancer.

email: fortin946@gmail.com

#### 4f. MetaOC: META-Analysis With One-Sided Correction to Detect Differentially Expressed Genes with Concordant Direction

**Xingbin Wang\***, University of Pittsburgh

**M. Ilyas Kamboh**, University of Pittsburgh

**George C. Tseng**, University of Pittsburgh

With the rapid advances and prevalence of high throughput genomic technologies, integrating information of multiple relevant genomic studies has brought new challenges.

Microarray meta-analysis has become a frequently used tool in biomedical research. Among the four categories of meta-analysis methods, combining p-values from multiple studies for DE gene detection has a long history in statistical science. However, when combining two-sided p-values for binary outcomes, the existing methods do not have the advantage of filtering discordant biomarkers such that DE genes with discordant DE direction can often be identified. In this paper, we systematically extended Owen's one-sided correction method to six existing meta-analysis methods of combining p-values. The results of simulations and applications in real data showed that the methods with one-sided correction are helpful to guarantee identification of DE genes with concordant DE direction without losing statistical power.

email: xingbinw@gmail.com

#### 4g. Normalization of DNA Methylation Microarrays Using Technical Covariates

**Paul T. Manser\***, Virginia Commonwealth University

**Mark Reimers**, Virginia Commonwealth University

Studies of DNA methylation hold enormous promise for uncovering regulatory epigenetic differences. Recently the Infinium HumanMethylation450 BeadChip was released as a major extension to Illumina's previous chip, the Infinium HumanMethylation27. The new array provides much greater coverage of CpG sites but its complex design can make normalization difficult and tedious. Previous normalization methods have used peak alignment and various quantile-normalization-based approaches that may not be suitable for samples from complex tissues or cancers, in which global changes in methylation levels may be expected. We propose to incorporate more available information by modeling variability in microarray signals as a function of a set of known technical covariates using a flexible local regression surface as an indirect model of real technical variability. We also incorporate biological information to achieve a normalization that is robust to global changes in methylation by selecting an empirical set of biologically invariant control probes from housekeeping and imprinted genes for fitting the surface. We show that our method performs favorably when compared with current normalization methods, both in terms of overall performance and in robustness to global changes in DNA methylation due to aberrant methylation in cancer, or shifting cell admixtures in complex tissue.

email: manserpt@vcu.edu

#### 4h. Sequence Kernel Association Test for Quantitative Traits in Twin Samples

**Kai Xia\***, University of North Carolina, Chapel Hill

**Wonil Chung**, University of North Carolina, Chapel Hill

**Zhaoyu Yin**, University of North Carolina, Chapel Hill

**Rebecca C. Santelli**, University of North Carolina, Chapel Hill

**Fei Zou**, University of North Carolina, Chapel Hill

Sequencing technologies have made it possible to identify rare variants that associate with complex traits. However, genome-wide association studies between rare variants and complex traits usually suffer from inefficiency, high false positive rate and low detection power. Many statistical methods have been proposed and the sequence kernel

association test (SKAT) is a powerful and fast method among others and its extensions such as famSKAT can handle big family data. However type I error is still not well controlled when applying SKAT or famSKAT to twin studies. Here, we extend SKAT to twinSKAT using two-step approach: first we estimate the parameters under null by implementing ACE mixed model using restricted maximum likelihood method; then score test using the parameters estimated is used for the association test. Through simulation study, we evaluated the type I error and power and found that twinSKAT has good control of type I error and good statistical power. The null linear mixed model only needs to be estimated once, so the computing time for twinSKAT is almost equivalent to SKAT. Since both additive genetic and shared environmental effects have been controlled, our model is able to achieve much better control type I error for twin samples.

email: argossy@gmail.com

#### **4i. An Alternative Approach to Model RNA-seq Data with GLMM**

**Han Sun\***, Cleveland Clinic

**Jiayang Sun**, Case Western Reserve University

The next generation sequencing technology has gradually gained dominance over the micro-array in the genomic profiling experiments. RNA-seq (RNA sequencing) is widely applied as a high-dimensional, high throughput gene expression technology (Ozsolak, 2011). RNA-seq has shown less noise level compared to the micro-array technology. One of the central questions in RNA-seq analysis is how to identify Differentially Expressed (DE) genes. Poisson distribution has been a popular choice for modeling the reads count data. Negative Binomial (NB) models are adopted to address the over-dispersion problem when biological replicates are available. Wiel et al (Van De Wiel, 2013) proposed a Bayesian method to estimate multiple shrinkage parameters in a generalized linear model context. Here we propose an alternative approach using Generalized Linear Mixed Model to accommodate the correlation structure in the biological replicated samples, thus aiming to provide a less computational expensive approach to for high throughput data analysis.

email: han.sunny@gmail.com

#### **4j. Classifying Family Relationships Using Dense SNP Data and Putative Pedigree Information**

**Zhen Zeng\***, University of Pittsburgh

**Eleanor Feingold**, University of Pittsburgh

When GWAS or sequencing studies are performed on family-based datasets, the genetic data can be used to check putative pedigrees. Even in datasets of putatively unrelated people, both close and distant relationships can often be detected using dense SNP data. A number of methods for finding relationships using genetic data exist, but most are intended for linkage-analysis type data: small numbers of uncorrelated genetic markers genotyped on small numbers of pedigrees. Another limitation of existing methods is that many use only a subset of the available information that can be used to classify pedigrees. In this paper we propose a set of approaches for classifying relationship types in GWAS datasets or large-scale sequencing datasets. We first

propose a method for finding regions of identity-by-descent in closely-related individuals using dense SNP data. We then demonstrate how that information can be used in principle to distinguish relationships. Finally, we propose classification pipelines for checking and identifying relationships aimed at datasets containing a large number of small pedigrees.

email: zhenhouse@msn.com

#### **4k. Identifying Multiple-Role Genes Dynamic in Distinct Environments**

**Yaqun Wang\***, The Pennsylvania State University

**Ningtao Wang**, The Pennsylvania State University

**Han Hao**, The Pennsylvania State University

**Rongling Wu**, The Pennsylvania State University

A gene can play multiple roles for adjusting the active proteins when organisms cope with change in the environment. Those genes cannot be discovered through classical clustering methods such as finite mixture model which assumes that a gene only belongs to one cluster. In this paper we describe an approach for identifying multiple-role genes by modeling the interactions between clusters for dynamic gene expression data measured at a set of discrete time points in differential environments. It is an extension of Gaussian mixture model by including link terms between clusters. We considered the covariance structures not only for the correlation-ship over time points but also for the correlation-ship over changing environments. We developed a maximum-likelihood method implemented with an EM algorithm for estimating model parameters. A number of quantitative testable hypotheses have been outlined about the properties of dynamic gene expression in distinct environments. The results were compared with the results from a standard finite normal mixture model.

email: yxw179@psu.edu

#### **4l. ChIP-seq META-CALLER: An Assembly Method to Combine Multiple ChIP-seq Peak Callers to Identify And Reprioritize The Peaks**

**Rui Chen\***, University of Pittsburgh

**Qunhua Li**, The Pennsylvania State University

**George C. Tseng**, University of Pittsburgh

ChIP-seq, which combining chromatin immunoprecipitation with next generation DNA sequencing, is primarily used to provide quantitative, genome-wide mapping of target protein and DNA interaction events. Although there are existing programs for previous ChIP-Chip analysis, it is still a computational challenge of bioinformatics to identify signal peaks of protein binding site from large sequencing read count based datasets. Popular peak calling methods, such as MACS, SPP, CisGenome, and SISR, are widely used but each of them has different emphasis on sensitivity, specificity and different size and shape of peaks. In this talk, we proposed a meta-analysis framework to combine multiple top-performing methods to identify and reprioritize the peaks. We showed that the result has improved sensitivity and specificity and is more trackable by biologists for further validation and hypothesis generation.

email: chenrui0728@gmail.com

#### 4m. Fast Annotation-Agnostic Differential Expression Analysis

**Leonardo Collado-Torres\***, Johns Hopkins University  
Bloomberg School of Public Health and  
Maltz Research Laboratories

**Andrew E. Jaffe**, Maltz Research Laboratories

**Jeffrey T. Leek**, Johns Hopkins University  
Bloomberg School of Public Health

Since the development of high-throughput technologies for probing the genome researchers have been interested in finding differences across groups that could potentially explain the observable phenotypic differences. In other words, methods for large-scale hypothesis generation that filters out as many artifacts as possible. The traditional tools have focused on the known transcriptome and are highly dependent on existing annotation. Frazee et al (Biostatistics 2013, in review) developed a statistical framework to find candidate Differentially Expressed Regions (DERs) without relying on annotation that produced sensible results. We have implemented a faster version of this approach in order to handle larger data sets: up to a few hundred samples. The total processing time is comparable to other tools for differential expression analysis such as DESeq (Anders et al, Genome Biology 2010). We will explain our software and show examples of the results produced from applying it to various publicly available data sets. The software is available at <https://github.com/lcollador/derfinder>.

email: [lcollado@jhsp.edu](mailto:lcollado@jhsp.edu)

#### 4n. Sample Size and Power Determination for Association Tests in Case-Parent Trio Studies

**Holger Schwender\***, Heinrich Heine  
University Duesseldorf

**Christoph Neumann**, TU Dortmund University

**Margaret A. Taub**, Johns Hopkins University

**Samuel G. Younkin**, Johns Hopkins University

**Terri H. Beaty**, Johns Hopkins University

**Ingo Ruczinski**, Johns Hopkins University

Transmission/disequilibrium tests (TDTs) are the most popular statistical tests for detecting single nucleotide polymorphisms (SNPs) associated with disease in case-parent trio studies considering genotype data from children affected by a disease and from their parents. Since several types of these TDTs have been devised, e.g., approaches based on alleles or on genotypes, it is of interest to evaluate which of these TDTs have the highest power in the detection of SNPs associated with disease. Since the test statistic of the genotypic TDT which is equivalent to a Wald test in a conditional logistic regression model had to be computed numerically, comparisons of other TDTs with the genotypic TDT have so far been based on simulation studies. Recently, we, however, have derived a closed-form solution for the genotypic TDT so that this analytic solution can be used to

derive equations for power and sample size calculation for the genotypic TDT. In this presentation, we show how these equations can be derived and compare the power of the genotypic TDT with the one of the corresponding score test assuming the same underlying genetic mode of inheritance as well as the allelic TDT based on a multiplicative mode of inheritance.

email: [holger.schwender@udo.edu](mailto:holger.schwender@udo.edu)

#### 4o. A Hierarchical Bayesian Approach to Detect Differential Methylation in Both Mean and Variance for Next Generation Sequencing

**Shuang Li\***, Georgia Regents University

**Varghese George**, Georgia Regents University

**Duchwan Ryu**, Georgia Regents University

**Xiaoling Wang**, Georgia Regents University

**Shaoyong Su**, Georgia Regents University

**Huidong Shi**, Georgia Regents University

**Robert H. Podolsky**, Wayne State University

**Hongyan Xu**, Georgia Regents University

DNA methylation at CpG loci is a very important epigenetic process involved in many complex diseases including cancer. In recent years, next-generation sequencing (NGS) has been widely used to generate genome-wide DNA methylation data. Although substantial evidence indicates that difference in mean methylation rate between normal and disease is meaningful, it has recently been proposed that it may be important to consider DNA methylation variability underlying common complex diseases. We propose a robust hierarchical Bayesian approach using a Latent Gaussian model to detect CpG sites that are differentially methylated with respect to both the mean and the variance. To identify differentially methylation associated with the disease, we consider jointly testing for the mean and the variance under the proposed hierarchical Bayesian framework. To improve computational efficiency, we use Integrated Nested Laplace Approximation (INLA), which combines Laplace approximations and numerical integration in a very efficient manner for deriving marginal posterior distributions. Our approach also overcomes the potential for non-convergence inherent in the alternate MCMC approach. We performed simulations to compare our proposed method to a score test. The simulation results indicate that our proposed approach detects more truly differentially methylated sites and fewer false positives, and it is computationally faster.

email: [shli@gru.edu](mailto:shli@gru.edu)

#### 4p. Bayesian Mixture Models for Complex Copy Number Polymorphisms Inferred from Genotyping Arrays

**Stephen Cristiano\***, Johns Hopkins University

**Robert B. Scharpf**, Johns Hopkins University

**Lynn Mireless**, Johns Hopkins University

Gain and loss of DNA copy number in the germline is a major source of genetic variation and has been implicated in common diseases such as schizophrenia and autism spectrum disorders. Here, we extend previously proposed

mixture models for copy number polymorphisms with a fully Bayesian implementation that more flexibly models the mean, variance, and skew of the copy number-induced mixture components. While the marker-level estimates are well known to be sensitive to batch effects, we demonstrate that copy number estimates derived from the Bayesian mixture model are relatively robust to these artifacts.

email: scristia@jhsph.edu

#### 4q. Multiple Phenotype Analysis for Genome-Wide Association Studies

**Shelley Liu\***, Harvard School of Public Health  
**Sheng Feng**, Biogen-Idec

Genome-wide association studies have allowed for the discovery of thousands of genetic variants associated with diseases and quantitative traits, but there are limitations to the standard approach of testing for associations between a single SNP and a single phenotype. By accounting for the correlation structure of phenotypes using joint modeling, it is reported that there is increased power in testing for associations, and thus joint modeling has the potential to increase discovery in GWAS. Three methods from the literature on multiple phenotype analysis (MultiPhen, Canonical Correlation Analysis, and Scaled Multiple-Phenotype Association Test) will be compared through simulations, in regards to power and Type I error rate considerations. We seek to compare the performance of these methods under different scenarios, such as varying sample sizes, effect sizes, allele frequencies; different correlation structures between phenotypes; and incorporating both categorical and continuous phenotypes. The three methods for multiple phenotype analysis are then applied to an Alzheimer's Disease Genetic dataset, and their results will be compared.

email: shelleyliu@fas.harvard.edu

#### 4r. EBSeq-HMM: An Empirical Bayes Hidden Markov Model for Ordered RNA-seq Experiments

**Ning Leng\***, University of Wisconsin, Madison

**Brian E. Mcintosh**, Morgridge Institute for Research  
**Yuan Li**, University of Wisconsin, Madison  
**Bao K. Nguyen**, Morgridge Institute for Research  
**Bret Duffin**, Morgridge Institute for Research  
**Shulan Tian**, Morgridge Institute for Research  
**James A. Thomson**, Morgridge Institute for Research  
**Ron Stewart**, Morgridge Institute for Research  
**Christina Kendzierski**, University of Wisconsin, Madison

RNA-Seq is now widely used to study gene expression across two or more biological conditions; and a number of statistical methods are available to identify differences between or among groups. Although useful, these methods assume that groups are unordered and consequently sacrifice power and precision when measurements are collected over time or space. Here we propose an empirical Bayes hidden Markov model called EBSeq-HMM. EBSeq-HMM extends EBSeq which was developed to identify differentially expressed genes and/or isoforms across two or

more conditions. In EBSeq-HMM, a hidden Markov model is implemented to accommodate dependence in gene and/or isoform expression over time (or space). EBSeq-HMM provides posterior probabilities for each gene (or isoform) and each possible expression path. As demonstrated in simulation and case studies, the output may prove useful in identifying DE genes, classifying their changes over time or space, and clustering genes with similar profiles.

email: nleng@wisc.edu

## 5. POSTERS: PREDICTION, PROGNOSTICS, DIAGNOSTIC TESTING

#### 5a. Joint Confidence Region Estimation for Area Under ROC Curve and Youden Index

**Jingjing Yin\***, University at Buffalo  
**Lili Tian**, University at Buffalo

In the field of diagnostic studies, the area under the Receiver Operating Characteristic (ROC) curve (AUC) serves as an overall measure of a biomarker/diagnostic test's accuracy. Youden index, defined as the maximum overall correct classification rate minus one at the optimal cut-off point, is another popular index. For continuous biomarkers of binary disease status, although researchers mainly evaluate the diagnostic accuracy using AUC, for the purpose of making diagnosis, Youden index provides an important and direct measure of the diagnostic accuracy at the optimal threshold and hence should be taken into consideration in addition to AUC. Furthermore, AUC and Youden index are generally correlated. We initiate the idea of evaluating diagnostic accuracy based on AUC and Youden index simultaneously. As the first step towards this direction, we only focus on the confidence region estimation of AUC and Youden index for a single marker. Both parametric and non-parametric approaches for estimating joint confidence region of AUC and Youden index are considered. Extensive simulation study is carried out to evaluate the performance of the proposed methods, and for illustration, a real data set is analyzed by the proposed methods.

email: jjyin.scu@gmail.com

#### 5b. Building Risk Models with Calibrated Margins

**Paige Maas\***, National Cancer Institute, National Institutes of Health  
**Raymond Carroll**, Texas A&M University  
**Nilanjan Chatterjee**, National Cancer Institute, National Institutes of Health

Risk models are used to weigh the risks and benefits of preventative interventions in clinical and public health settings. For many diseases, established risk models have been developed from large representative cohorts and validated in independent studies. As new risk factors are identified, there is a need to update existing risk models to fully use the most up-to-date information in predicting

disease risk. It is often not reasonable, or possible, to conduct a new cohort study to collect the few additional risk factors needed to refit a given risk model. In fact, a more efficient method would add new risk factors while incorporating information from existing models as much as possible. We investigate two approaches for using existing models to calibrate the new model. First, we explore using a regression calibration approach in this context, utilizing a method from sample-survey literature which is traditionally used for increasing the efficiency of parameter estimation from a given survey by leveraging information from external data sources. Second, we develop a likelihood-based inference procedure that incorporates information from the existing risk model through the Kulback-Likelihood distance measure. We present analytic and numerical results that evaluate the performance of these approaches in various relevant scenarios.

email: pmaas@jhsph.edu

### 5c. Meta-TSP: A Meta-Analysis Framework of Top Scoring Pair Algorithm to Combine Multiple Transcriptomic Studies in Inter-Study Prediction Analysis

**SungHwan Kim\***, University of Pittsburgh  
**George C. Tseng**, University of Pittsburgh

The top scoring pair algorithm (TSP) and its variants have been shown to be a simple, yet robust and accurate rank-based method in the cross-validation of a single transcriptomic study. When it is applied to inter-study prediction, the accuracy is, however, low. As high-throughput genomic experiments become affordable and prevalent, multiple data sets of a relevant biological or clinical hypothesis are often available. Conceptually, the information integration of multiple studies can potentially improve the inter-study prediction accuracy. In this paper, we develop meta-analysis frameworks for TSP to generate variations of Meta-TSP classifiers when multiple transcriptomic studies are combined. We extend from three TSP variations (the original single TSP, top K-TSP and a variance optimized TSP) using two meta-analysis strategies. Through simulations and three real data applications, we show that all Meta-TSP variations obtain significantly better prediction accuracy than original TSP variations. Meta-kTSP performs better than Meta-TSP as expected. Simple average of TSP scores generally performs better than Fisher and Stouffer's p-value combination methods. In conclusion, the Meta-TSP methods help facilitate inter-study prediction analysis in both perspective and retrospective studies and will enhance translational research of the high-throughput diagnostic assays.

email: suk73@pitt.edu

### 5d. A Modified Tree-Based Method for Personalized Medicine Decisions

**Wan-Min Tsai\***, Yale University  
**Heping Zhang**, Yale University  
**Stephanie O'Malley**, Yale University  
**Ralitza Gueorguieva**, Yale University

The tree-based methodology has been extensively applied to identify predictors of health outcome in medical studies. However, in clinical trials, this approach in general does not pay particular attention to treatment assignment. In recent years, attention is shifting from average treatment effects to identifying moderators of treatment response. Thus it is desirable to develop a simple and efficient statistical tool that would provide information about clinical decision making regarding a personalized medical treatment. In this study, we extend and present modifications to the classical recursive partitioning technique proposed by Zhang et al. (2010) to efficiently identify subgroups of subjects who respond more favorably to one treatment than another based on their baseline characteristics. We introduce an automatic pruning step in the tree building process and program the tree building and tree pruning steps in R. We evaluate the performance of the proposed method through a simulation study and illustrate the proposed approach using a real dataset from a study of alcoholism.

email: wan-min.tsai@yale.edu

### 5e. A Simple Method for Evaluating within-Sample Prognostic Balance Achieved by Published Comorbidity Summary Measures

**Brian L. Egleston\***, Fox Chase Cancer Center, Temple University Health System  
**Robert G. Uzzo**, Fox Chase Cancer Center, Temple University Health System  
**J. Robert Beck**, Fox Chase Cancer Center, Temple University Health System  
**Yu-Ning Wong**, Fox Chase Cancer Center, Temple University Health System

Researchers are often interested in using published comorbidity summary scores such as the Charlson Comorbidity Index, Elixhauser Score, or ACE-27 measures. These summary scores can be useful in assessing the combined impact of multiple diseases on an individual's health outcomes. Many of these summary measures were created using regression methods, and their use is valid under the appropriate conditions. One such condition is that the relationships of the comorbidities with the outcome of interest in a researcher's own population are comparable to the relationships in a published algorithm's population. We demonstrate how the examination of survival curves for individual diseases within strata of a comorbidity summary measure can give guidance as to whether the use of a published measure is appropriate in a given sample. We provide an example using early stage kidney cancer cases identified in the Surveillance Epidemiology and End Results (SEER) database linked to Medicare claims data.

email: Brian.Egleston@fccc.edu

## 5f. Effect Size Measures for Functional Modifiers of Treatment Response

**Adam Ciarleglio\***, New York University  
School of Medicine

Heterogeneity in treatment response is an important consideration when assigning patients to treatment. The effectiveness of a treatment or even whether a treatment is harmful or beneficial can depend on a host of patient characteristics. With increasingly better technology and computing ability, the list of potential treatment effect modifiers has grown to include variables that can be classified as functional data (e.g., EEG data, etc.). Because of their structure and infinite dimensional nature, functional data pose many challenges not inherent in non-functional data. We propose several measures that quantify the strength of functional variables as treatment effect modifiers. They can be used as measures of effect size and can help to identify functional effect modifiers. The new measures are developed for the case of a continuous treatment response under the assumptions of a functional linear model. We study the properties of these measures through simulations and apply them to clinical trial data in order to compare treatments for depression. Furthermore, we propose and evaluate a procedure that uses our new measures to combine both functional and non-functional modifiers of treatment effects into a composite effect modifier that might better account for heterogeneity in treatment response than any individual one.

email: ajc2171@columbia.edu

## 5g. Power Calculations for Prognostic Biomarker Validation Studies with Time to Event Data

**Marshall D. Brown\***, Fred Hutchinson Cancer  
Research Center

**Yingye Zheng**, Fred Hutchinson Cancer Research Center  
**Tianxi Cai**, Harvard School of Public Health

A validation study to evaluate the clinical utility of a prognostic biomarker is critical for translating novel biomarkers into routine clinical practice. One important question in designing such a study is: what sample size should one use to ensure that definite conclusions can be drawn? Few publicly available tools exist to guide researchers in this important decision especially when the outcome is measured in censored failure time. We have developed an interactive web based application to allow investigators to implement sample size calculations and power simulations for both semiparametric and nonparametric estimates of common accuracy measures. This poster session will explain our methods, and display the application in use. The application was built using Rstudio's R package Shiny, and is hosted on Rstudio's Shiny server.

email: mdbrown@fhcrc.org

## 5h. Generalized Incremental Forward Stagewise Ordinal Models: Application Predicting Stage of Alzheimer's Disease

**Kellie J. Archer\***, Virginia Commonwealth University  
**Jiayi Hou**, Virginia Commonwealth University

Penalized methods have been successfully applied to high-throughput genomic datasets in fitting linear, logistic, and Cox proportional hazards models. However, extensions for fitting penalized models for predicting ordinal responses have not yet been fully characterized, even though clinical and histological outcomes are frequently recorded using an ordinal scale. Herein we describe an algorithm that leverages Hastie et al's (2007) generalized monotone incremental forward stagewise (GMIFS) method to fit cumulative logit, adjacent category, and continuation ratio models. These ordinal GMIFS models are capable of predicting an ordinal phenotype when high-dimensional genomic data comprise the predictor space. We illustrate application of the ordinal GMIFS models for the purpose of predicting stage of Alzheimer's disease (normal, incipient, moderate, severe) using post-mortem hippocampal samples that were profiled using Affymetrix gene expression microarrays. The genes included in the models may be indicative of pathogenic mechanisms involved in Alzheimer's disease progression.

email: kjarcher@vcu.edu

## 6. POSTERS: SURVIVAL ANALYSIS

### 6a. Non-Parametric Confidence Bands for Survival Function Using Martingale Method

**Eun-Joo Lee\***, Millikin University

A simple computer-assisted method of constructing non-parametric simultaneous confidence bands for survival function with right censored data is introduced. This method requires no distributional assumptions. The procedures are based on the integrated martingale process whose distribution is approximated by a Gaussian process. The supremum distribution of the Gaussian process generated by simulation leads to a construction of the confidence bands. To improve the inference procedures for the finite sample sizes, the log-minus-log transformation is employed.

email: elee@millikin.edu

### 6b. On The Estimators and Tests for the Semiparametric Hazards Regression Model

**Seung-Hwan Lee\***, Illinois Wesleyan University

In the accelerated hazards regression model with censored data, estimation of the covariance matrices of the regression parameters is difficult, since it involves the unknown baseline hazard function and its derivative. This paper provides simple but reliable procedures that yield asymptotically normal estimators whose covariance matrices can be easily estimated. A class of weight functions is introduced to result in the estimators whose asymptotic covariance matrices do not involve the derivative of the unknown hazard function. Based on the estimators obtained from different weight functions, some goodness-of-fit tests are constructed to

check the adequacy of the accelerated hazards regression model. Numerical simulations show that the estimators and tests perform well. The procedures are illustrated in the real world example of leukemia cancer. For the leukemia cancer data, the issue of interest is a comparison of two groups of patients that had two different kinds of bone marrow transplants. It is found that the differences of the two groups are well described by a time-scale change in hazard functions, i.e., the accelerated hazards model.

email: slee2@iwu.edu

#### 6c. Regression Analysis of Bivariate Current Status Data with the Proportional Hazards Model and Bernstein Polynomials

**Tao Hu**, Capital Normal University

**Qingning Zhou\***, University of Missouri, Columbia

**Jianguo Sun**, University of Missouri, Columbia

We consider regression analysis of bivariate current status or case I interval-censored failure time data under the marginal proportional hazards model. For the problem, several estimation procedures have been proposed, but each of these existing methods either applies only to limited situations or has no theoretical justification available. By using Bernstein polynomials and an unspecified copula model, we develop a sieve maximum likelihood estimation approach that applies to much more general situations. In particular, it allows one to estimate the underlying copula model and can be easily implemented. We show that the proposed estimates are consistent and can achieve the optimal convergence rate. The asymptotic normality and efficiency of the estimates of regression parameters are also established. A simulation study is conducted for the assessment of the performance of the developed approach and suggests that it works well for practical situations. An illustrative example is also provided.

email: qz4z3@mail.missouri.edu

#### 6d. Joint Structure Selection and Estimation in the Time-Varying Coefficient Cox Model

**Wei Xiao\***, North Carolina State University

**Wenbin Lu**, North Carolina State University

**Hao Helen Zhang**, University of Arizona

Time-varying coefficient Cox model has been widely studied and popularly used in survival data analysis due to its flexibility for modeling covariate effects. It is of great practical interest to accurately identify the structure of covariate effects in a time-varying coefficient Cox model, i.e. covariates with null effect, constant effect and truly time-varying effect, and estimate the corresponding regression coefficients. Combining the ideas of local polynomial smoothing and group non-negative garrote, we develop a new penalization approach to achieve such goals. Our method is able to identify the underlying true model structure with probability tending to one and simultaneously estimate the time-varying coefficients consistently. The asymptotic properties of the resulting estimators are established. We demonstrate the performance of our method using simulations and an application to the primary biliary cirrhosis data.

email: wxiao0421@gmail.com

#### 6e. Weighted Log-Rank Tests for 'Flipped-Data' Survival Analysis of Data With Non-Detects

**Eric R. Siegel\***, University of Arkansas for Medical Sciences

**Songthip T. Ounpraseuth**, University of Arkansas for Medical Sciences

**Ralph L. Kodell**, University of Arkansas for Medical Sciences

Non-detects are data whose values are left-censored at a limit of detection (LOD). Data with non-detects arise in fields as diverse as metabolomics, environmental monitoring, and AIDS research. To analyze data with non-detects, sophisticated methods such as maximum-likelihood estimation and multiple imputation have been deployed, but these methods require fitting a model whose error term follows the Normal or other parametric distribution. A simple, non-parametric alternative was proposed by Helsel (2005), in which data with non-detects are 'flipped' or converted into right-censored forms by subtracting them from a suitably large number, then analyzed via Kaplan-Meier curves and the log-rank test. In a simulation study, we investigated the performance of Helsel's method on normally distributed data subjected to left-censoring at an LOD. We found that the Gehan (or Wilcoxon) test, a weighted version of the log-rank test, had significantly more power to detect group differences than the standard log-rank test. Here, we explore whether the Gehan test continues to be superior to the log-rank test when the left-censored data are generated using alternatives to the normal distribution.

email: siegeleric@uams.edu

#### 6f. A Frailty Approach for Survival Analysis with Error-Prone Covariate

**Sehee Kim\***, University of Michigan

**Yi Li**, University of Michigan

**Donna Spiegelman**, Harvard School of Public Health

This paper discovers an inherent relationship between the survival model with covariate measurement error and the frailty model. We further consider a semi-parametric frailty approach for the correction of bias due to covariate measurement error in survival data analysis. Using a frailty-based estimating equation, we draw inference for the estimated regression coefficients in the proportional hazard model with error-prone covariates. Our established framework accommodates general distributional structures for the error-prone covariates, not restricted to a linear additive measurement error model or Gaussian measurement error. When the conditional distribution of the frailty given the surrogate is unknown, it is estimated semi-parametrically through a copula as a function of the nonparametric marginal distributions of the true exposure and the surrogate. The proposed approach via the copula enables us to fit flexible measurement error models without the curse of dimensionality as in nonparametric approaches, and to be applicable with an external validation study. Finite sample properties are investigated through extensive simulation studies. The methods are applied to a study of physical activity in relation to breast cancer mortality in the Nurses' Health Study.

email: seheek@umich.edu

## 6g. LC-Morph: A Morphological Image Signature for Predicting Lung Cancer Survival

**Yuchen Yang\***, University of Kentucky

**Fuyong Xing**, University of Kentucky

**Hai Su**, University of Kentucky

**Chi Wang**, University of Kentucky

**Li Chen**, University of Kentucky

**Lin Yang**, University of Kentucky

**Arnold Stromberg**, University of Kentucky

An appropriate statistical model for survival analysis on lung cancer can provide precise prognosis for treatment planning. Usually the traditional prognostic decisions are made purely based on pathologists' subjective evaluations. Our whole process includes cell detection, segmentation, and building statistical model for survival analysis. 122 patients information extracted from the TCGA data set has been used in this study. A robust seed detection-based cell segmentation algorithm is proposed to accurately segment each individual cell in the image. Based on the cell segmentation results, a set of cellular features are extracted using some efficient image feature descriptors. To build a prognostic image signature for patient overall survival, the study data set is randomly split into a training data set (82 patients) and a testing data set (40 patients). Based on the training data, univariate Cox models are used to identify informative image features. A lasso-penalized Cox model is used to derive an image feature-based prognostic model and calculate the corresponding risk score (LC-Morph score). The score is externally validated on the testing data set. We also stratify patients into high- and low-risk groups based on the LC-Morph score and find significantly longer survival time in the low-risk group than the high-risk group (log-rank  $P=0.013$ ).

email: yuchen.y@uky.edu

## 7. POSTERS: IMAGING, HIGH DIMENSIONAL DATA, BIOMARKERS, AND MICROARRAY

### 7a. On the Distribution Of Photon Counts with Censoring in Two-Photon Microscopy

**Burcin Simsek\***, University of Pittsburgh

**Satish Iyengar**, University of Pittsburgh

**David Kleinfeld**, University of California, San Diego

Photon counting methods can give better images than analog methods in two-photon laser scanning microscopy. However, photon detectors have a dead period that leads to undercounts. We describe a model for photon generation and derive the distribution of the observed counts, which we then use to estimate the number of photons emitted.

email: bus5@pitt.edu

### 7b. Bayesian Gaussian Process Regression for High-Dimensional Data

**Qing He\***, Emory University

**Jian Kang**, Emory University

**Qi Long**, Emory University

There has been a growing interest in fitting non-parametric regression models via Gaussian processes (GPs) from a Bayesian perspective. Although the posterior inference for the GP regression model is mathematically tractable, the computational costs can be very expensive, typically at the scale of  $O(n^3)$  which limits its use for high-dimensional data analyses. A rich variety of methods have been proposed to address the problem using approximation techniques. Some recent options include Gaussian predictive process model for large spatial data (Banerjee et. al. 2008) and a random projection method on a lower dimensional subspace (Banerjee et. al. 2011). Alternatively, in this paper, we investigate an efficient posterior computation method that is developed from a fast GP simulation procedure (Wood & Chan 1994). This method provides exact posterior inference and is particularly useful when the observed data points are equally spaced. We apply the Metropolis adjusted Langevin algorithm (Roberts & Stramer, 2003) and Riemann manifold Langevin/Hamiltonian Monte Carlo algorithm (Girolami & Calderhead 2011) to this method. We conduct numerical studies to compare the performance of these different methods and demonstrate the superiority of our approaches.

email: qing.he@emory.edu

### 7c. Effects of Alcohol Use on Brain Networks: A Dynamic Causal Model Study with EEG Data

**Benjamin T. Brown\***, University of Minnesota

**Lynn Eberly**, University of Minnesota

**Steve Malone**, University of Minnesota

**Kathleen Thomas**, University of Minnesota

Brain connectivity, the task-related activation of the brain within networks among regions, can be estimated from neuroimaging data using Dynamic Causal Modeling (DCM). One or more brain network models can be estimated for each person based on EEG or fMRI data. Model comparison is done within-person using a Bayesian approach. The excellent temporal resolution of EEG makes it well suited for analysis of connectivity but its poor spatial resolution makes it difficult to find significant results for connection strengths. We describe the structure of DCMs and illustrate their use on EEG data from young adults who participated in a mismatch-negativity task. In this passive attention task, subjects view a closed-caption video while hearing a sequence of pure acoustic tones. Tones are of two types differing in frequency. Infrequent tones occur 12% of the time and elicit a "preattentive" orienting response. Scientific interest is in connectivity related to brain activation during infrequent tones relative to resting (no tones). Utilizing our empirical distributions of connectivity parameters, a group comparison is conducted using a permutation test to assess if networks are different between two groups: one consisting of individuals with a lifetime alcohol abuse disorder, and a control group with no substance abusers.

email: brow3774@umn.edu

#### 7d. C.Logic: A Classification Algorithm for Discovering Interactions that Lead to Disease Susceptibility

**Sybil L. Nelson\***, Medical University of South Carolina  
**Bethany Wolf**, Medical University of South Carolina  
**Viswanathan Ramakrishnan**, Medical University of South Carolina

Finding reliable statistical methods to predict disease outcome from genetic and environmental factors and their interactions is an important problem in medicine. Many scientists argue that the interaction of factors make a substantial contribution to disease susceptibility but are too often ignored in genetic studies. Two methods capable of classifying a person's disease risk in terms of interactions are Classification and Regression Trees (CART) and Logic Regression. However, CART tends to preferentially include continuous variables in a model and Logic Regression is designed exclusively for dichotomous variables. The goals of this project is to first, develop a protocol that simulates gene-gene and gene-environmental interactions second, develop a method that focuses on identifying interactions between variables that predict disease and third, use the simulated data to test the properties of the new method for correctly identifying interactions associated with the outcome. We have so far completed the simulation protocol and developed an algorithm, called C.Logic, for inclusion of continuous variables in a Logic Regression framework. The preliminary simulation results indicate that C.Logic is as good as or better than CART for recovering continuous and binary predictors and predictor interactions known to be associated with the outcome.

email: slp23@musc.edu

#### 7e. A Direct Approach to False Discovery Rate Regression

**Simina M. Boca\***, National Cancer Institute, National Institutes of Health  
**Jeffrey T. Leek**, Johns Hopkins Bloomberg School of Public Health

Adjusting for multiple testing by using false discovery rates is now standard for many scientific problems. The false discovery rate may be estimated by first calculating a p-value corresponding to each null hypothesis considered. Here we develop a method which allows for the incorporation of additional meta-data to use in the decision about whether to reject the null hypothesis for a specific test. We do this by proposing a framework for false discovery rate regression. We show via theoretical results and simulation that we can generally obtain conservative estimates of the covariate-specific false discovery rate.

email: simina.boca@nih.gov

#### 7f. A Study of the Correlation Structure of Microarray Gene Expression Data Based on Mechanistic Modelling of Cell Population Kinetics

**Linlin Chen\***, Rochester Institute of Technology  
**Lev Klebnov**, Charles University  
**Anthony Almudevar**, University of Rochester  
**Christoph Proschel**, University of Rochester

Sample correlations between gene pairs within expression profiles are potentially informative regarding gene regulatory pathway structure. However, as is the case with other statistical summaries, observed correlation may be induced or suppressed by factors which are unrelated to gene functionality. In this paper, we consider the effect of heterogeneity on observed correlations, both at the tissue and subject level. Using gene expression profiles from highly enriched samples of three distinct embryonic glial precursors of the rodent neural tube, the effect of tissue heterogeneity on correlations is directly estimated for a simple two component model. Then, a stochastic model of cell population kinetics is used to assess correlation effects for more complex mixtures. Finally, a mathematical model for correlation effects of subject-level heterogeneity is developed. Although decomposition of correlation into functional and nonfunctional sources will generally not be possible, since this depends on non-observable parameters, reasonable bounds on the size of such effects can be made using the methods proposed here.

email: linlin.chen@gmail.com

#### 7g. Making Computerized Adaptive Testing a Diagnostic Tool

**Hua-Hua Chang\***, University of Illinois, Urbana-Champaign  
**Ya-Hui Su**, National Chung Cheng University

Although CAT was originally developed by for high stakes testing, its findings have been beneficial to other domains such as quality of life measurement, patient report outcome, K-12 accountability assessment, survey research, media and information literacy measure, etc. The paper provides a survey of 10 years' progress about Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT). We start with a historical review of the establishment of a large sample foundation for CAT under a framework of martingale theory. Then, we address a number of issues that emerged from large scale implementation in educational testing and show that CAT can be used in the research of patient reported outcome. In addition we show the newest development in the research of multidimensional cognitive diagnostic adaptive assessment. Many issues concerning CAT in quality of life measurement will be discussed.

email: hhchang@illinois.edu

## 7h. Missing Value Imputation in High-Dimensional Phenomic Data: Imputable or Not? And How?

**Serena Liao\***, University of Pittsburgh  
**George C. Tseng**, University of Pittsburgh

In modern biomedical research of complex diseases, phenomic data are often collected for each patient and missing values are inevitable in the data collection process. Since many downstream statistical methods require complete data matrix in the implementation, imputation is a common and practical solution. In high-throughput experiments such as microarray data analysis, continuous intensities are measured and many mature missing value imputation methods have been developed and widely applied. Large phenomic data, however, contain continuous, nominal, binary and ordinal data types, which void application of most methods. In this paper, we developed four variations of K-nearest-neighbor (KNN) methods and compared with two existing methods, multivariate imputation by chained equations (MICE) and random forest (MissForest). We introduced a novel concept of imputability measure to characterize missing values that are fundamentally inadequate to impute. Simulations and applications showed that MICE often did not perform well; KNN-A, KNN-H and random forest were among the top performers although no method universally performed the best. Finally, we proposed a self-training selection scheme to select the best imputation method and provide a practical guideline for general applications.

email: liaoge.serena@gmail.com

## 7i. Age Prediction Using Supervised PCA

**Valerie J. Watkins\***, University of North Carolina, Wilmington  
**Yishi Wang**, University of North Carolina, Wilmington

Human age is among the most significant soft biometric features to extract from face images. Due to the complex nature of images and the high correlation among facial features, manifold learning is often a necessary step to reduce the feature dimension and/or decrease the correlation among the covariates. However, in soft biometric analysis the most widely used manifold learning methods such as PCA, CCA, ICA, and FLD are unsupervised. In this work, we will use the supervised PCA method proposed in Barshan et al., 2011 to analyze face age of the Productive Age Lab (PAL) database. Results of the experiment will be compared with previously published results of unsupervised methods.

email: vjw1275@uncw.edu

## 7j. Tensor Regression with Applications in Neuroimaging Data Analysis

**Xiaoshan Li\***, North Carolina State University  
**Hua Zhou**, North Carolina State University  
**Lexin Li**, North Carolina State University

Modern technologies are producing a wealth of data with complex structures. For instance in medical sciences, brain images of individuals are collected for large neuroimaging studies. These image data often take the form of 3D or even higher dimensional arrays, also known as tensors. To address the scientific questions arising from the data, new regression methods that take arrays as covariates are needed. Simply turning an image array into a vector leads to ultra-high dimensionality in classical regression methods and causes loss of valuable information. To keep the structural integrity of the data, we propose a flexible tensor regression model based on the Tucker decomposition of the regression coefficient array. It effectively exploits information in the tensor covariates, reduces the ultra-high dimensionality to a manageable level, and thus results in efficient estimation. We demonstrate the effectiveness of the method by applying it to identify brain regions associated with attention deficit hyperactivity disorder (ADHD) from magnetic resonance images (MRI). In addition, we compare our method with another type of tensor regression based on CANDECOMP/PARAFAC (CP) decomposition.

email: xli12@ncsu.edu

## 7k. Investigating Spatiotemporal Covariance Structures for Modeling Longitudinal Imaging Data

**Brandon J. George\***, University of Alabama, Birmingham  
**Inmaculada Aban**, University of Alabama, Birmingham

Longitudinal imaging studies allow great insight into how the structure and function of a subject's internal anatomy changes over time. Unfortunately, the analysis of longitudinal imaging data is complicated by inherent spatial and temporal correlation: the temporal from the repeated measures, and the spatial from the outcomes of interest being observed at multiple points in a patient's body. The issue is compounded by a small sample size characteristic of imaging studies, particularly ones employing MRI, which prevents the use of unstructured covariance functions. Previous work has utilized summary methods to avoid this problem, which results in a loss of information regarding the outcome in specific regions. We propose the use of a linear model with a parametric spatiotemporal error structure to analyze for analysis of repeated imaging data. This simulation study compared different information criteria for selecting a particular separable parametric spatiotemporal correlation structure as well as the effects on Type I and II error rates when the specified model is incorrect. Sample size, skewness, and degree of correlation (low vs. high) were examined for their effects on covariance structure selection and error rates. The simulation scenarios were inspired by a longitudinal cardiac imaging study on mitral regurgitation patients.

email: brgeorge@uab.edu

## 7l. Nonparametric Regression with Tree-Structured Response

**Yuan Wang\***, University of Texas  
MD Anderson Cancer Center  
**J. S. Marron**, University of North Carolina, Chapel Hill  
**Haonan Wang**, Colorado State University  
**Burcu Aydin**, University of North Carolina, Chapel Hill  
**Alim Ladha**, University of North Carolina, Chapel Hill  
**Elizabeth Bullitt**, University of North Carolina, Chapel Hill

Highly developed science and technology from the last two decades motivated the study of complex data objects. In this paper, we consider the topological properties of a population of tree-structured objects. Our interest centers on modeling the relationship between a tree-structured response and other covariates. For tree objects, this poses serious challenges since most regression methods rely on linear operations in Euclidean space. We generalize the notion of nonparametric regression to the case of a tree-structured response variable. In addition, a fast algorithm with theoretical justification is developed. We implement the proposed method to analyze a data set of human brain artery trees. An important lesson is that smoothing in the full tree space can reveal much deeper scientific insights than the simple smoothing of summary statistics.

email: ywang46@mdanderson.org

## 7m. Improving Scan-Rescan Reliability of Resting State fMRI Parcellation

**Amanda Mejia\***, Johns Hopkins School of Public Health  
**Mary Beth Nebel**, Kennedy Krieger Institute  
**Stewart Mostofsky**, Kennedy Krieger Institute  
**Brian Caffo**, Johns Hopkins School of Public Health  
**Martin Lindquist**, Johns Hopkins School of Public Health

Introduction: Resting-state functional connectivity (RSFC) is believed to measure intrinsic brain functioning, and differences in group-level RSFC network parcellations are associated with certain disorders, including ASD and ADHD. More recently, parcellation at the subject level has been attempted, but scan-rescan reliability of subject-level parcellations remains understudied. Methods: Scan-rescan resting-state fMRI scans for 20 healthy adults were obtained from the public Kirby-21 dataset. Two novel methods are proposed to improve scan-rescan reliability: first, an iterative clustering method draws subject-level parcellations towards the group; second, empirical Bayes shrinkage is applied to the correlation matrices used for clustering. A simulation is performed for validation. Results: Existing methods show poor scan-rescan reliability for subject-level parcellations. Our proposed methods improve scan-rescan reliability by approximately 20%. Conclusions: Resting-state fMRI data is highly variable, and parcellations created from subject-level data alone are unreliable and provide poor subject-level inference. However, subject- and group-level fMRI data may be combined to yield more reliable subject-level parcellations.

email: mandy.mejia@gmail.com

## 7n. SGPP: Spatial Gaussian Predictive Process Models for Neuroimaging Data

**Jung Won Hyun\***, St. Jude Children's Research Hospital  
**Yimei Li**, St. Jude Children's Research Hospital  
**John H. Gilmore**, University of North Carolina, Chapel Hill  
**Zhaohua Lu**, University of North Carolina, Chapel Hill  
**Martin Styner**, University of North Carolina, Chapel Hill  
**Hongtu Zhu**, University of North Carolina, Chapel Hill

We aim to develop a spatial Gaussian predictive process (SGPP) framework for accurately predicting neuroimaging data by using a set of covariates of interest, such as age and diagnostic status, and an existing neuroimaging data set. To achieve better prediction, we not only delineate spatial association between neuroimaging data and covariates, but also explicitly model spatial dependence in neuroimaging data. The SGPP model uses a functional principal component model to capture medium-to-long-range (or global) spatial dependence, while SGPP uses a multivariate simultaneous autoregressive model to capture short-range (or local) spatial dependence as well as cross-correlations of different imaging modalities. We propose a three-stage estimation procedure to simultaneously estimate varying regression coefficients across voxels and the global and local spatial dependence structures. Furthermore, we develop a predictive method to use the spatial correlations as well as the cross-correlations by employing a cokriging technique, which can be useful for the imputation of missing imaging data. Simulation studies and real data analysis are used to evaluate the prediction accuracy of SGPP and show that SGPP significantly outperforms several competing methods, such as voxel-wise linear model, in prediction.

email: jwhyun05@gmail.com

## 7o. Dimension Reduction Using Inverse Spline Regression

**Kijoeng Nam**, U.S. Food and Drug Administration  
**Paul J. Smith**, University of Maryland, College Park

In high-dimensional data analysis, we often want to reduce the number of predictors without eliminating variables which are related to the response of interest. Inverse regression methods use the response variable when performing dimension reduction so that information regarding the relation between the covariates and the response is not lost. However, it is common to assume that the inverse regression function is linear or to use some other ad hoc approach. Instead, we propose a new dimension reduction method which models the inverse regression function as a spline. We develop asymptotics for our approach and demonstrate its performance through simulations and several data sets commonly found in the Machine Learning literature. We show that its performance is better than existing inverse regression based methods, especially when the dimension reduction space is a nonlinear manifold such as the Swiss roll example of Roweis and Saul.

email: kijoeng@math.umd.edu

## 7p. Interpreting Large Dense (Scary) Linear Models Along Predictor Groups

**Yuval Benjamini\***, Stanford University  
**Julien Mairal**, INRIA, Grenoble  
**Bin Yu**, University of California, Berkeley

Linear predictive models are getting big; in neuroscience and genomics predictive models often use thousands or more non-orthogonal predictors. Model accuracy often benefits from the use of so many predictors. Still, it is unclear how scientists should interpret the parameter vector once predictive accuracy has been proven. In large dense models, looking at only few leading predictors may fail to capture the important effects in the model. We propose a method for summarizing such models along pre-specified predictor groups. We take into account the covariance structure of the design, and identify the impact of each group on the predictions. We use this method to explore - from predictive models - profiles of vision neuron in cortical area V4.

email: yuvalben@stanford.edu

## 8. POSTERS: ENVIRONMENTAL AND LONGITUDINAL DATA ANALYSIS

### 8a. Accounting for Complex Survey Design in Modeling Temporal Trends of Phthalate Metabolites in the U.S. Population

**Min Chen\***, ExxonMobil Biomedical Sciences, Inc.  
**Kevin Kransler**, ExxonMobil Biomedical Sciences, Inc.  
**Rosemary Zaleski**, ExxonMobil Biomedical Sciences, Inc.  
**Hua Qian**, ExxonMobil Biomedical Sciences, Inc.

We examined temporal trends in urinary levels of metabolites in the U.S. population for the six phthalates (di-butyl phthalate, DBP; di-isobutyl phthalate, DiBP; butyl-benzyl phthalate, BBP; bis (2-ethylhexyl) phthalate, DEHP; di-isononyl phthalate, DINP; and di-isodecyl phthalate, DIDP) from the National Health and Nutrition Examination Survey (NHANES) 1999-2010. The NHANES uses a complex, multistage probability sampling design with over-sampling of ethnic minorities and young children. Statistical analyses took into account unequal probabilities of selection resulting from the complex sample design, nonresponse, and non-coverage. Standard errors were estimated with Taylor series linearization. There were significant decreasing trends in urinary levels of Mono-n-butyl phthalate (DBP metabolite), Mono-benzyl phthalate (BBP metabolite), Mono-(2-ethyl)-hexyl phthalate (DEHP metabolite), Mono-(2-ethyl-5-hydroxyhexyl metabolite) phthalate (DEHP metabolite), Mono-(2-ethyl-5-oxohexyl) phthalate (DEHP metabolite), and Mono-2-ethyl-5-carboxypentyl phthalate (DEHP metabolite) in the United States. Urinary levels of Mono-isobutyl phthalate (DiBP metabolite) increased over the period of 2001-2010. Urinary levels of Mono (carboxyoctyl) phthalate (DINP metabolite) increased over the period of 2005-2010. Mono (carboxynonyl) phthalate (DIDP metabolite) showed no significant trend over the period of 2005-2010.

email: min.chen@exxonmobil.com

### 8b. Non-Stationary Covariance Functions Via Domain Segmentation

**Douglas C. Hom\***, University of Michigan  
**Timothy D. Johnson**, University of Michigan  
**Veronica J. Berrocal**, University of Michigan

Point-referenced spatial data is often modeled using stationary isotropic Gaussian processes, that is, processes where the spatial dependence is a function only of the distance between points. While the assumption of stationarity has computational advantages, for some environmental processes it might not be realistic. Several models have been proposed for non-stationary covariance function, both in a parametric and a non-parametric setting. In this paper, we present a model for non-stationary covariance functions by assuming that the spatial process in consideration is locally stationary, while globally it is not. Specifically, we assume that there exists a latent Gaussian process which segments the spatial domain into regions governed by local stationary covariance functions. As a consequence, the covariance between the spatial process of interest at any two locations does not depend simply on their distance but also on the position of the two sites within the domain. We illustrate the capability and flexibility of our model via simulation experiments, in which we compare the out-of-sample predictive performance of our model to that of other established models for both stationary and non-stationary covariance functions. As a real data application, we have applied our model to temperature data for the Pacific Northwest.

email: doughom@umich.edu

### 8c. The Effect of Exposure to Air Toxics on Age of Diagnosis and Subtype Of Childhood Leukemia - A Joint Modeling Approach

**Ting-Yu Chen\***, University of Texas School of Public Health  
**Elaine Symanski**, University of Texas  
School of Public Health  
**Wenyaw Chan**, University of Texas School of Public Health

A joint modeling approach was applied to assess tripartite associations among air pollution exposure, type of leukemia, and age at diagnosis in children with acute leukemia. This study focused on children who were diagnosed with leukemia at 4 years old or younger and identified from records in the Texas Cancer Registry (TCR). The individual-level confounder information was obtained by linking TCR records to birth certificates from the Texas Department of State Health Services. The residential exposure to air toxics of the mother at time of delivery was determined through modeled annual estimates of census tract ambient air levels for benzene, 1,3-butadiene, and polycyclic organic matter (POM), which were available from the U.S. EPA's National-Scale Air Toxics Assessment (NATA). For each pollutant, the proposed joint models involved simultaneously estimating a linear mixed model and a generalized linear mixed model. An algebraic expression of the correlation coefficient between the two outcomes was derived, but it does not have an explicit form. Monte Carlo integration was then applied to complete the computation. The confidence interval for this correlation coefficient was established by the Bootstrap method.

email: Ting-Yu.Chen@uth.tmc.edu

### 8d. Investigating the Health Risks Associated with Long Term Exposure to Coarse PM

**Helen L. Powell\***, Johns Hopkins Bloomberg School of Public Health  
**Roger D. Peng**, Johns Hopkins Bloomberg School of Public Health

In recent studies, consideration has been given to the health risks associated with the coarse fraction of particulate matter (PM), that is, particles which are between 2.5 and 10 microns in diameter. Studies have found that the acute effects of coarse PM are as strong if not stronger than those of fine PM (< 2.5 microns in diameter), which suggests that consideration should be given to the study and regulation of coarse particles. However, those who have investigated the long-term effects of this particular pollutant have found conflicting results. This may be due to differences with regards to the time periods and geographical boundaries under consideration. However, it may also be due to issues with regards to spatial confounding, which occurs when there is a lack of measurements on the key variables which may affect the relationship between the pollutant and the health outcome of interest. Therefore, using health data from the Medicare billings claims and pollution data from the EPA monitoring network we aim to investigate the need for a model which accounts for the potentiality of spatial confounding when investigating the health risks of chronic exposure to coarse PM.

email: hpowell@jhsph.edu

### 8e. Functional Data Analysis to Guide a Conditional Likelihood Regression in a Case-Crossover Study Investigating whether Social Characteristics Modify the Health Effects of Air Pollution

**Juana M. Herrera\***, University of Texas, El Paso  
**Joan Staniswalis**, University of Texas, El Paso  
**Sara E. Grineski**, University of Texas, El Paso

We are exploring whether social characteristics modify the relationship between air pollution and hospitalizations due to asthma or chronic pulmonary obstructive disease (COPD) in El Paso, TX. The case-crossover design with conditional regression analysis was used. Social characteristics are included in the models as interactions with the pollutants; variables used are age, sex, ethnicity and insurance status. The pollutants lags were chosen using the historical functional linear model to estimate the association between the response and pollutant at all lags simultaneously. The regression coefficient function was calculated by P-splines with the smoothing parameter chosen with a modified ridge trace method. We included single pollutant analyses for NO<sub>2</sub> and PM<sub>2.5</sub>, adjusting for apparent temperature and wind speed. The lags for low and high wind speed were chosen, in the case of asthma, based on previous literature and in the case of COPD based on odds ratios. We found that when PM<sub>2.5</sub> is equal to the 98th percentile of the daily values, Hispanics are more likely to be hospitalized due to asthma and COPD than non-Hispanics, but when NO<sub>2</sub> is equal to the 98th percentile for the daily values, contrary to PM<sub>2.5</sub>, Non-Hispanics are more likely to be hospitalized than Hispanics.

email: jmherrera4@miners.utep.edu

### 8f. Dependence Modeling of Spatio-Temporal Weather Extreme Events

**Whitney Huang\***, Purdue University  
**Hao Zhang**, Purdue University

There are two main objectives of spatial extreme modeling. The first one is to model the marginal behavior of extremes where to calculate return levels is of the main concern. The second one, the modeling of spatio-temporal extreme dependence is more of a challenge. The most widely used approach, max-stable processes, forms one useful characterization of extreme dependence of spatial processes. However, the lack of space-time modeling and the restricted tail dependence structure may lead to overestimation of the level of dependence in the extremes. In this talk, we analyze the extreme events in terms of temperature and precipitation based on observations at 750 weather stations across China to study the regional pattern of extreme events. Some thoughts beyond max-stability are discussed.

email: huang251@purdue.edu

### 8g. Identifying the Constellation of Emergency Health Conditions Most Sensitive to Extreme Heat

**Jennifer Bobb\***, Harvard School of Public Health

Extreme heat is currently the largest cause of severe weather fatalities in the US, and as climate change progresses the health impacts are likely to be profound. Although the adverse effects of extreme heat on broad classes of health outcome, such as cardiovascular and respiratory mortality, have been well established, the specific health conditions that are most sensitive to extreme heat have not been systematically identified. Here we develop methodology to identify the constellation of acute health conditions that are most sensitive to extreme heat. We consider a broad range of disease groupings classified by ICD-9 code, and we match extreme heat event (EHE) days to control days during the summer months. We estimate the number of excess cause-specific hospitalizations attributable to EHE in a cohort of approximately 12 million Medicare beneficiaries in 213 US counties during 1999-2010. Knowledge of the constellation of acute health conditions that occur during EHE will enable better preparedness to treat emerging conditions during major heat episodes.

email: jenniferfederbobb@gmail.com

## 8h. Statistical Strategies for Constructing Health Risk Models with Multiple Pollutants and their Interactions

**Zhichao Sun\***, University of Michigan

**Yebin Tao**, University of Michigan

**Shi Li**, University of Michigan

**Kelly K. Ferguson**, University of Michigan

**John D. Meeker**, University of Michigan

**Sung Kyun Park**, University of Michigan

**Stuart A. Batterman**, University of Michigan

**Bhramar Mukherjee**, University of Michigan

Estimating the adverse health effects due to simultaneous exposure to multiple pollutants is an important topic to explore, and its challenges reside in, but are not limited to: identification of the most critical components of the pollutant mixture, examination of potential interaction effects, and attribution of health effects to individual pollutants in the presence of multicollinearity. In this study, we reviewed five methods available in the statistical literature and conducted a simulation study evaluating their performances. We also proposed a two-step strategy employing an initial screening by a tree-based method followed by further dimension reduction/variable selection at the second step. From our investigation, there is no uniform dominance of one method across all simulation scenarios. Least absolute shrinkage and selection operator regression performs well for identifying important exposures, but will yield biased estimates and slightly larger model dimension given extensive collinearity and modest sample size. Bayesian model averaging and supervised principal component analysis are useful in variable selection under a strong exposure-response association. Substantial improvements on reducing model dimension and identifying important variables have been observed for the two-step modeling strategy when a large number of candidate variables exist, implying its potential under a multipollutant framework.

email: zcs@umich.edu

## 8i. Mixed Effects Models for Investigating Dietary Regimens Intended to Extend Lifespan in *Caenorhabditis Elegans*

**Jeffrey Burton\***, Pennington Biomedical Research Center

**Robbie Beyl**, Pennington Biomedical Research Center

**Jolene Zheng**, Pennington Biomedical Research Center

**William D. Johnson**, Pennington Biomedical Research Center

In preliminary aging studies, animal models are used to investigate effects of treatments designed to extend length of life. One such model involves *C. elegans* (nematodes) and has been used in testing the effect of dietary regimens on lifespan or healthspan. Healthspan can be assessed via repeated measurements of pharyngeal pumping rate (PPR) which is recorded as a count of pharyngeal muscle contractions per minute and decreases as the animal ages. Thus, maintenance of a healthy PPR is associated with a longer lifespan. Using this study design, aging data are generated to simulate lifespan in *C. elegans*. The analytic data are modeled using two generalized linear mixed models. The first assumes a Poisson distribution for the PPR

response and the second assumes a normal distribution. Hypothesis tests of equality of mean PPR between diet groups are performed via the two models and results are compared. *C. elegans* provide a novel, time- and cost-efficient experimental animal model for investigating aging. Here, we employ the generalized linear mixed model to promote a unified approach to researchers using *C. elegans* that is consistent with traditional experimental statistics.

email: Jeffrey.Burton@pbrc.edu

## 8j. Simulation from a Known Cox Msm Using Standard Parametric Models for the G-Formula

**Jessica G. Young\***, Harvard School of Public Health

**Eric J. Tchetgen Tchetgen**, Harvard School of Public Health

It is routinely argued that, unlike standard regression-based estimates, inverse probability weighted (IPW) estimates of the parameters of a correctly specified Cox marginal structural model (MSM) may remain unbiased in the presence of a time-varying confounder affected by prior treatment. Previously proposed methods for simulating from a known Cox MSM lack knowledge of the law of the observed outcome conditional on the measured past. Although IPW estimation does not require this knowledge, regression-based estimates rely on correct specification of this law. Such simulation methods, therefore, cannot generally isolate bias due to complex time-varying confounding as it may be conflated with bias due to model misspecification. Here we describe an approach to Cox MSM data generation that allows isolation of each type of bias. This approach involves simulating data from a parametrization of the likelihood and solving for the underlying Cox MSM. We prove that solutions exist and computations are tractable under many data generating mechanisms. We show analytically and confirm in simulations that, in the absence of model misspecification, the bias of regression-based estimates is indeed a function of the coefficients in observed data models quantifying the presence of a time-varying confounder affected by prior treatment.

email: jyoung@hsph.harvard.edu

## 8k. Reflecting the Orientation of Teeth in Random Effects Models for Periodontal Outcomes

**Rong Xia\***, University of Michigan

**Thomas M. Braun**, University of Michigan

Clinical attachment level (CAL) is a tooth-level measure that quantifies the severity of periodontal disease. Expressed in millimeters, CAL is the distance on a tooth from its crown to where its root is attached to the gingiva (gums). The within-mouth correlation of teeth is difficult to model due to the three-dimensional spatial geography of teeth and their functional similarity. Thus, traditional approaches have included (1) applying a t-test to mouth-level averages or (2) using generalized estimating equations (GEE) with simple correlation structures. As an alternative, we propose several linear mixed models that include random effects to better quantify the within-mouth correlation of teeth and lead to more efficient parameter estimation. Via simulation, we

compare the bias and efficiency of fixed effects estimates computed with our models to corresponding results produced with t-tests and GEE. We demonstrate that our mixed models give estimates that are unbiased and more efficient than other methods that fail to accurately model the within-mouth correlation of teeth. We also evaluate the performance of the approaches when data are missing under different biologically plausible mechanisms of missingness.

email: rongxia@umich.edu

### 8l. A Longitudinal Beta-Binomial Model for Over-Dispersed Binomial Data

**Hongqian Wu\***, University of Iowa

**Ying Zhang**, University of Iowa

Longitudinal binomial data are frequently generated from multiple questionnaires and assessments in various scientific researches. Generalized linear mixed-effects model (GLMM) is usually the standard approach for this type of data. However, GLMM may result in severe underestimation of standard error of estimated regression parameters in the presence of over-dispersed binomial data. In this paper, we propose a longitudinal beta-binomial model for over-dispersed binomial data and estimate the regression parameters under Probit model using Generalized Estimating Equation (GEE) method. A hybrid algorithm of the Fisher Scoring and the method of moment estimation is implemented for computing the model. An extensive simulation study is conducted to justify the validity of the proposed method. Finally the proposed method is applied to analyze functional impairment in subjects who are at risk of Huntington disease (HD) from a multi-site observational study of prodromal HD, PREDICT-HD.

email: hongqian-wu@uiowa.edu

## 9. POSTERS: EPIDEMIOLOGY AND CAUSAL INFERENCE

### 9a. Methods of Missing-Data Exploration that Reveal Potential Extrapolation

**Victoria Liublinska\***, Harvard University

Universal recommendations for reporting missing data and conducting sensitivity analyses in empirical studies are scarce. Both steps are often omitted by practitioners due to the lack of clear guidelines for summarizing missing data and systematic explorations of alternative assumptions. As a result, there is no common structure in reporting practices observed throughout the literature. We discuss recommendations on important missing data summaries that should appear in every study report. We especially focus on one of the most informative, but rarely reported, features of units in a study with missing data, namely, the extent to which the units with and without missing data look alike. Careful examination of the overlap between units' characteristics might reveal potential risk of extrapolation. Several graphical and analytical methods of assessing the extent of overlap are presented and discussed.

email: vliublin@fas.harvard.edu

### 9b. Exploring Mobile Technology to Enhance Birth Outcomes in Rural Mozambique: Pilot Study Results

**Manoj T. Rema\***, Georgia State University

**Ike Okosun**, Georgia State University

**Sheryl Strasser**, Georgia State University

Introduction: The World Vision Organization currently has a Mobile Health division. Mobile Technologies for Health (mHealth) is the term used for practicing medicine and public health, supported by mobile phones and other communication devices, such as tablets and personal digital assistants. This new field has emerged as a viable source to communicate health needs and collect community health data. It has been proven to help deliver healthcare information to community health workers (CHW), researchers, physicians and patients, in real-time. The goal was to compare two groups of prenatal mothers and see if mobile phone technologies provided a viable resource to better serve the health care needs of those in the rural area of Mozambique. Methods: The phones were used for health promotion, data collection, CHW training and emergency referral. The mobile phones were implemented into the intervention group and were compared to CHWs without the mHealth intervention. A survey was administered at the end of the study to women in both groups; data was analyzed to compare the experimental group with the control group to judge if the intervention led to more awareness of pregnancy and postpartum danger signs in women. Odds ratios, confidence intervals and p-values for each indicator were calculated and compared between groups. Results: The results above show, mothers who know at least 2 danger signs in pregnancy is significantly higher in the control area (68%, OR=0.4, p-value=0.009) than in the intervention group (51.6%). The proportion of mothers who know danger signs in the postpartum period is fairly low in both groups, but the intervention group (11.8%, OR=0.4, p-value=0.05) is significantly higher than the control group (5.3%). Discussion: Based on the findings, the interventions group was also more likely to know about pregnancy and postpartum danger signs than the control group. Because the difference in the two groups was the mHealth intervention modules, it can be proven that the cause of the improvements between the groups was the mobile phones; a self-selection bias could have accounted for the difference.

email: manojrema1@gmail.com

### 9c. Transformations to the Zero-Inflated Negative Binomial Model for Overall Exposure Effects: An Analysis of Blood Lead and Dental Caries in a Complex Survey

**D. Leann Long\***, West Virginia University

**R. Constance Wiener**, West Virginia University

Often dental health researchers are interested in investigating exposures of interest, such as blood lead levels, and exposure effects on caries experience or severity. One measure of caries severity is the number of decayed or filled teeth (dft), an index which frequently presents an excess of zeros to standard discrete distributions. Though zero-inflated regression models can be used to account for these excess zeros, these models often produce latent class

interpretations, with one set of parameters associated with the probability of being an excess zero and another set of parameters associated with the mean of the non-excess zero subpopulation. While the zero-inflated framework is a convenient tool for accounting for excess zeros, our research is ultimately interested in population exposure effect estimates. We present a zero-inflated negative binomial (ZINB) model using blood lead levels and dft index from the Third National Health and Nutrition Examination Survey (NHANES III). Transformations of the ZINB parameter estimates are calculated to estimate overall exposure effects and the delta method is employed to obtain variance estimates.

email: dllong@hsc.wvu.edu

#### **9d. Applying Multiple Imputation Using External Calibration to Propensity Score Methods**

**Yenny G. Webb-Vargas\***, Johns Hopkins Bloomberg School of Public Health  
**Elizabeth A. Stuart**, Johns Hopkins Bloomberg School of Public Health

The effect of using covariates with measurement error in propensity score methods has not been widely investigated. We assessed its implications when using inverse probability of treatment weighting to estimate treatment effects, and we developed a method based on the Multiple Imputation using External Calibration (MIEC) method by Guo et al (2012) that corrects for the bias when the calibration data is supplied externally to the main study data. We found that there is bias in the treatment effect estimate when using the covariate measured with error in the propensity score and outcome models. In contrast, our method estimates the treatment effect almost as well as we would if the confounding covariate was measured without error. We also found that the method does not correct for the bias if the final outcome is absent from the imputation procedure. We use a motivating example looking at the effects of early intensive intervention for young children with autism to illustrate our methods. These results show that estimating the propensity score using covariates measured with error leads to biased estimates of treatment effect, and MIEC can be used to correct for such bias.

email: ywebbvar@jhsph.edu

#### **9e. Efficient Estimation of Partial Rank-Based Correlation with Missing Data**

**Wei Ding\***, University of Michigan  
**Peter X.K Song**, University of Michigan

Rank-based correlation is widely used in practice to measure dependence between variables when their marginal distributions are skewed. Estimation of such correlation is challenged by both the presence of missing data and the need of adjusting for confounders. In this paper, we develop a unified framework of Gaussian copula regression that enables us to estimate both partial Pearson correlation and partial rank-based correlation (e.g. partial Kendall's tau or partial Spearman's rho), depending on the assumed marginal distributions. To adjust for confounding factors, we utilize marginal regression models with location-scale distributions for error terms. We establish the EM algorithm in this semi-

parametric modeling framework to handle the estimation with missing values. The semi-parametric efficiency of our estimation method is also discussed. We propose a peeling procedure to implement iterations required in the EM algorithm. We compare the performance of our proposed method to the traditional multiple imputation approach using simulation studies. For structured correlation, such as an exchangeable or a first-degree auto-regressive (AR1) correlation matrix, our method outperforms multiple imputation approach in both bias and efficiency.

email: dingwei@umich.edu

#### **9f. Data Analysis of Contributing Factors for Obesity in Low-Income Neighborhoods**

**Sujin Kim\***, Savannah State University  
**Rukmana Deden**, Savannah State University

Obesity is one of the leading factors contributing high risk for heart disease, diabetes, and cancer which can be prevented and direct and indirect cost for obesity has severe economic impact even though it can be prevented at low cost early. Obesity is related with poverty because low-income people tend to consume food with more calories at lower cost and low income people have limited outdoor exercise and physical activities due to lack of sidewalks and parks. The prevalence of obesity in Chatham County, GA has increased rapidly over the past decades. Especially, the study of obesity in low-income neighborhoods, Chatham County is worthwhile to set up the data base. This NIH-funded study has collected data relating obesity from 6 low-income neighborhoods in Savannah, Chatham County, GA by administering surveys with 50 questions. The collected data has been analyzed in each neighborhood, and analyzed whole data by using meta-analysis method.

email: kims@savannahstate.edu

### **10. POSTERS: NON PARAMETRIC AND NON LINEAR METHODS**

#### **10a. Comparison of Area Under the Curve and Mixed Effects Models Methodologies for Profile Analysis**

**Robbie A. Beyl\***, Pennington Biomedical Research Center  
**Jeffrey Burton**, Pennington Biomedical Research Center  
**William Johnson**, Pennington Biomedical Research Center

Assume that study subjects are randomly assigned to one of K treatments and assessed for a specific response at time 0 (baseline) and each of T post-treatment times that are not necessarily equally spaced. Further, assume the objective is to compare the treatments with respect to their response profiles across time. The first issue is to determine if there is significant treatment by time interaction. If there is convincing evidence that interaction is negligible, the next issue is to determine if the treatment and time main effects are significant. If there is evidence of interaction, the treatments typically are compared at each assessment time with adjustments for multiple comparisons. Currently, the analytical method of choice is to employ mixed effects models for repeated measures. Nevertheless, many analysts prefer comparing the treatments in terms of area under the

curve (AUC). Despite its long history and widespread use, there appear to be many misconceptions about the merits of using AUC for profile analysis. In this presentation, we use comparative studies of response profiles from an oral glucose tolerance test to show situations in which some analytical methods are more (or less) appropriate than others.

email: robbie.beyl@pbrc.edu

### 10b. Inferential Approaches to Relative Risk Regression

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

The relative risk is of natural interest to measure the treatment effect on a binary outcome, and can be directly interpreted as the ratio of the probability of the outcome event between the two arms of the exposure or intervention. In terms of regression estimations, the standard log-linear estimating equations suffer from boundary issues, to ensure the fitted probabilities not exceeding 1. The asymptotic distribution of the corresponding constrained estimators turns out to be rather complex (instead of simply normality) when the true parameters lie on the boundary of the parameter space. Alternatively, transformations of the parameter space into an unconstrained one are considered, but the convergence rate of the estimators depends on the degree of smoothness of the estimating equations after transformations. Here we propose to circumvent the boundary constraints in modeling the probabilities of the outcome by modeling the propensity score instead. We summarize the general form of such estimating equations and identify the most efficient regular and asymptotically linear (RAL) estimator within a restricted sub-class, which is numerically more feasible in real data application.

email: yilu@jhsph.edu

### 10c. Fractional Polynomial Regression with Multilevel Data

**Paul Kolm\***, Christiana Care Health System  
**Daniel Elliot**, Christiana Care Health System  
**Joann Brice**, Christiana Care Health System  
**Robert Young**, Northwestern University

There are a variety of methods available for building nonlinear models with linear model parameters (e.g., splines, polynomial regression). Fractional polynomial (FP) regression can be applied in almost any regression context including Cox proportional hazards regression. In many cases, FP regression is easier to implement and provides a better fit than ordinary polynomial regression and splines. In this study, we apply FP regression to multilevel data, hospitalization length of stay (LOS), where potential predictors of LOS, overall hospital occupancy and physician workload, vary from day to day. We compare the results to those of ordinary polynomial regression and splines.

email: pkolm@christianacare.org

### 10d. Blup Estimation in Unbalanced Mixed-Effects Models

**Samaradasa Weerahandi**, Pfizer Inc.  
**Peijin Xie**, Hershey's Company  
**Ching-Ray Yu**, Pfizer Inc.  
**Kelly H. Zou\***, Pfizer Inc.

Mixed models are now heavily employed in studies of public health and in clinical research. Widely-used likelihood-based methods for making inferences about the Best Linear Unbiased Predictor (BLUP) suffer from a number of drawbacks. For example, reasons include the non-convergence and lack of accuracy with small samples and few groups. The BLUP in mixed models is a function of the variance components, which are estimated using Maximum Likelihood (ML) or Restricted ML (REML) estimation. Unfortunately, ML and REML either do not provide any BLUPs or provide equal BLUPs for a fraction of possible samples due to zero variance estimates that they yield. To overcome this drawback, for unbalanced linear models, we derive a Generalized Estimator (GE) of the BLUP that does not suffer from the problem of negative or zero variance components. The superiority of the proposed method over ML and REML is established via a simulation study and are illustrated in dental research.

email: Kelly.Zou@pfizer.com

### 10e. Flexible Test for Interactions in Smoothing Spline Anova Models through the Use of Distance Correlation

**Sebastian J. Teran Hidalgo\***, University of North Carolina, Chapel Hill  
**Michael Wu**, University of North Carolina, Chapel Hill  
**Michael Kosorok**, University of North Carolina, Chapel Hill

In this work, we propose a flexible interaction test procedure in the Smoothing Splines ANOVA (SS-ANOVA) framework. This is done through the use of the Distance Correlation (DC) test statistic. We accomplish this by first fitting a SS-ANOVA model with  $p$  variables with main effects and a predefined set of interactions, and then, by testing the independence between the estimated residuals and the  $p$  variables using the DC. This procedure detects the signal left over from the interactions not included in the fitted model. Hence, we have a test that has as null-hypothesis no interactions exist beyond the ones already included in the model, and as alternative-hypothesis there exists some interaction not included in the model. Another possible setting would be to fit a SS-ANOVA with only main effects and run the test to check if any interactions exist at all. Many variations of these two settings can be tested. We use a model-based bootstrap procedure to estimate the distribution of the test statistic under the null. Extensive simulations were implemented to demonstrate that the procedure has correct Type-I error and good power performance.

email: shidalgo@email.unc.edu

## 10f. Optimal Global Test for Functional Linear Regression Models and its Applications

**Xiao Wang**, Purdue University  
**Simeng Qu\***, Purdue University

This paper studies minimax test of the nullity of the slope function in the framework of functional linear model and reproducing kernel Hilbert space. The quality of the test is measured by the minimal distance between the null and the alternative set for which such test is still possible. The lower bound for the minimax separation distance of the slope function is first derived. It is shown that the optimal rate is jointly determined by the reproducing kernel and the covariance kernel. However, this rate is different with the rate when studying prediction. We also propose the generalized likelihood ratio test statistic based an easily implementable data-driven roughness regularization estimate. It is shown that the generalized likelihood ratio test attains the optimal rate of convergence.

email: simengqu@gmail.com

## 10g. MODEL TUMOR PATTERN AND COMPARE TREATMENT EFFECTS USING SEMIPARAMETRIC LINEAR MIXED-EFFECTS MODELS

**Changming Xia\***, University of Rochester  
**Jianrong Wu**, St Jude Children's Research Hospital  
**Hua Liang**, The George Washington University

To analyze responses of solid tumor to treatments and to compare treatment effects with antitumor therapies, we applied semiparametric mixed-effects models to fit tumor volumes measured over a period. The population and individual nonparametric functions were approximated by smoothing spline. We also proposed an intuitive method for a comparison of the antitumor effects of two different treatments. Biological interpretation was also discussed.

email: c.xia@rochester.edu

## 10h. ROBUST VARIANCE COMPONENT ANALYSIS WITH APPLICATIONS IN BIOLOGICAL ASSAY VALIDATION

**Binbing Yu\***, MedImmune, LLC.

Biological assays (bioassays) are used to measure the effects of a biologically active substance using an intermediate in vivo or in vitro tissue or cell model under controlled conditions. The goal of a bioassay validation is to confirm that the assay is fit for its intended use. Particularly, the parameters in validation studies include relative accuracy, specificity, intermediate precision and range. Contributions of validation study factors to the overall intermediate precision of the bioassay can be determined using a variance component analysis, where the random variation can be estimated using the linear mixed model (LMM). However, the LMM often relies on the unrealistic assumption of

normality. Second, outliers may occur in bioassay. Third, the resulting estimates of variance components based on the standard statistical methods may be zero or even negative. To address these issues, we proposed a new robust variance component model (VCM) with focus on the biological assay validation. The proposed method is illustrated using two motivating examples from bioassay validation.

email: whybb@yahoo.com

## 10i. ORACLE INFERENCE FOR GMM MODELS

**Mihai C. Giurcanu\***, University of Florida  
**Brett D. Presnell**, University of Florida

We develop a new class of GMM estimators for stationary time series which have an oracle property: their asymptotic behavior is the same as of the efficient GMM estimators under the apriori information that the true moment conditions were known in advance. These oracle estimators are obtained as weighted efficient GMM estimators, where the weights are chosen so that the effects of the incorrect moment conditions become automatically negligible. We study the large sample properties of the Wald, likelihood-ratio type, score type, and J test statistics as well as the asymptotic properties of the standard block-bootstrap and centered block-bootstrap estimators of their null distributions. The results of an extensive simulation study show the finite sample properties of the analytic and block-bootstrap tests for a dynamic spatial process with correctly specified and misspecified random effects, respectively. Data analysis of a dynamic spatial data set shows practical aspects of our oracle GMM methodology.

email: giurcanu@ufl.edu

## 10j. COVARIATE-DEPENDENT FUNCTIONAL INFERENCE FOR THE LIFE-TIME CIRCADIAN RHYTHM OF PHYSICAL ACTIVITY

**Luo Xiao\***, Johns Hopkins  
Bloomberg School of Public Health  
**Lei Huang**, Johns Hopkins  
Bloomberg School of Public Health  
**Ciprian Crainiceanu**, Johns Hopkins  
Bloomberg School of Public Health

We introduce a covariate-dependent functional model to quantify the systematic and random changes of the circadian rhythm of physical activity as a function of subjects' age. The systematic changes are captured by a nonparametric smooth bivariate function of time-of-day and age. We propose to use a very fast bivariate spline smoother with the smoothing parameters selected by leave-one-subject-out cross-validation. To the best of our knowledge, this is the first time leave-one-subject-out cross validation is proposed for bivariate smoothing of two different variables. The age-dependent random changes are modeled by a covariate-dependent covariance operator. A fast trivariate smoother is proposed to estimate the covariance and its spectrum. Statistical methods are inspired by and applied to the Baltimore Longitudinal Study on Aging (BLSA). Physical activity is recorded as activity

counts per minute, which results in 1440 observations per day, with most subjects having data over multiple days. Our methods are computationally practical for dealing with the size and the correlation structure of the data. Results reveal several interesting, previously unknown, circadian patterns associated with human aging and gender.

email: lxiao@jhsph.edu

## 11. POSTERS: VARIABLE SELECTION, MACHINE LEARNING AND OTHER

### 11a. An Extended Beta Regression Model

**Min Yi\***, University of Missouri, Columbia  
**Nancy Flournoy**, University of Missouri, Columbia

The beta distribution demonstrates a simple and flexible model in which response is naturally confined to a finite interval. The parameters of the distribution can be related to covariates such as dose and gender through a regression model. However, the beta distribution is naturally restricted between known boundaries, 0 and 1. An extended beta regression model is developed to expand the response boundaries from (0, 1) to (L, U), where L and U are unknown parameters. Given bioassay data, we compare several extended beta regression models assuming that the mean of the beta function follows the four-parameter logistic function. Also, we compare the estimates of LD10 and LD50 from three different models: the four-parameter logistic model with normal errors, a bounded log-linear regression model that was recently characterized by Wang and Flournoy, and the extended beta regression model.

email: vincenty43@gmail.com

### 11b. Model-Adjusted Standardization to Account for Unmeasured Cluster-Level Covariates with Complex Survey Data

**Zhuangyu Cai\***, University of Florida  
**Babette Brumback**, University of Florida

Model-adjusted standardization relies on a statistical model to estimate an unconfounded population-averaged effect. We extend existing methodology to accommodate a high-dimensional categorical confounder, such as a cluster identifier. We use a generalized linear mixed model with a random effect that can be associated in an arbitrary manner with both individual-level confounders and the exposure of interest. We show how to consistently estimate the standardized effect, and we apply the new method to adjust for unmeasured zip-code level confounding in an analysis of the 2008 Florida Behavioral Risk Factor Surveillance System survey.

email: zycaiok@yahoo.com

### 11c. A New Multiple Comparisons with the Best Procedure

**Tianshuang Wu\***, University of Michigan  
**Susan Murphy**, University of Michigan

Among several available treatment options, we want to screen out those treatment options that have low expectations of a primary outcome, i.e. we only recommend to patients treatments with high mean outcomes. Multiple comparisons with the best (MCB) has been developed by S. Gupta and J. Hsu to achieve this goal. MCB aims to construct a confidence set that excludes some treatment options that are very likely to be not the best, i.e. inferior to other treatments. However, this procedure always fails to exclude treatment which yields high variance in the outcome, thus might end up including treatments with low mean outcomes, which are not screened out simply because of large variance in the outcome. We introduce a new MCB procedure that takes this issue into account. Instead of focusing on the expectation, we focus on the some quantile of the outcome associated to a treatment. Intuitively this reflects how bad a treatment can behave most of the time. This novel method penalizes the mean by the standard deviation, thus screening out treatment options with low mean or large variance. In particular, when the outcome in each treatment group has the same variance, this method boils down to traditional MCB.

email: wutiansh@umich.edu

### 11d. Mixture of D-Vine Copulas for Modeling Dependence

**Daeyoung Kim**, Sungkyunkwan University, Korea  
**Jong-Min Kim**, University of Minnesota, Morris  
**Shu-Min Liao**, Amherst College  
**Yoonsung Jung\***, Prairie View A&M University

The identification of an appropriate multivariate copula for capturing the dependence structure in multivariate data is not straightforward. The reason is because standard multivariate copulas (such as the multivariate Gaussian, Student-t, and exchangeable Archimedean copulas) lack flexibility to model dependence and have other limitations, such as parameter restrictions. To overcome these problems, vine copulas have been developed and applied to many applications. In order to reveal and fully understand the complex and hidden dependence patterns in multivariate data, a mixture of D-vine copulas is proposed incorporating D-vine copulas into a finite mixture model. As a D-vine copula has multiple parameters capturing the dependence through iterative construction of pair-copulas, the proposed model can facilitate a comprehensive study of complex and hidden dependence patterns in multivariate data. The proposed mixture of D-vine copulas is applied to simulated and real data to illustrate its performance and benefits.

email: yojung@pvamu.edu

### 11e. Ensemble Variable Selection and Estimation (EVE)

**Sunyoung Shin\***, University of North Carolina, Chapel Hill  
**Yufeng Liu**, University of North Carolina, Chapel Hill  
**Jason Fine**, University of North Carolina, Chapel Hill

The penalized maximum likelihood estimation has been extensively studied for simultaneous variable selection and estimation. However, the direct penalization on a certain full likelihood is computationally infeasible and requires model-specific theoretical work. To tackle the problems, we propose Ensemble variable selection and estimation (EVE) for a factorizable likelihood. EVE is a multi-layer procedure based on three methods: information combination across the factors, least squares approximation (LSA) method, and refitting. The full likelihood estimation can be obtained from likelihood factor estimation via weighted least squares. LSA is used to select important variables on the full likelihood. With the selected variables, we refit each factor and recombine the estimators. Our estimation has no asymptotic efficiency loss and is computationally efficient with existing software. Simulation studies and data analysis on HIV/AIDS studies confirm that EVE is competitive with other existing methods.

email: sunyoung@live.unc.edu

### 11f. Support Vector Classifiers and Missing Data: An Investigation of the Complete-Case Solution and a Proposal of an EM-Like Solution

**Thomas G. Stewart\***, University of North Carolina, Chapel Hill  
**Donglin Zeng**, University of North Carolina, Chapel Hill  
**Michael C. Wu**, University of North Carolina, Chapel Hill

Support Vector Classification, a statistical learning method developed in recent decades, has demonstrated utility for a wide variety of classification tasks. The method is well adapted to situations of very large sample size or for situations when the number of predictors is large. As a result, the classification method has gained popularity and is becoming a regular part of the researcher's tool-kit. In large part, the method is applied to research tasks for which accurate prediction is the primary goal. Missing data issues are common among health research tasks, including classification and prediction. In practice, most analysts building a support vector classifier will use imputation techniques when missing data is an issue. Other missing data methods specific to support vector classifiers are also available. Despite the availability of these missing data methods, we ask: Are they always necessary? Our presentation summarizes commonly used missing data methods. We provide guidance to help researchers identify if and when such missing data methods are needed. Further, we propose an EM-like missing data method for support vector classifiers. We demonstrate the performance of the proposed method along with other methods in both a simulation and a real data application.

email: tgs@live.unc.edu

### 11g. Evaluating Novel Intradialytic Sampling Designs for Individual Pharmacokinetic Analysis Using Monte Carlo Simulation

**Minchun Zhou\***, Vanderbilt University  
**William Henry Fissell**, Vanderbilt University  
**Matthew Stephen Shotwell**, Vanderbilt University

The goal of this study was to evaluate the utility of a novel intradialytic sampling design for the purpose of individual antibiotic pharmacokinetic (PK) analysis in patients receiving renal replacement therapy. The intradialytic sampling design involves halting and then resuming dialysis in quick succession, causing the blood concentration of antibiotic to fluctuate in a way that is especially informative about individual pharmacokinetics. This technique is contrasted with a conventional PK sampling design, where patients are only sampled while not receiving dialysis. A two-compartment PK model accommodating intermittent intravenous dosing (zero order) and dialysis (first order) is used to model the timecourse of drug concentration. Monte Carlo simulation is performed using estimates of PK variability available in the literature. The variance in estimators of the 24-hour area under the concentration-time curve (AUC) and percent time above a minimum inhibitory concentration in 24 hours (%T>MIC) are compared among sampling designs. We hypothesized that the intradialytic sampling method would yield less variable estimators of AUC and %T>MIC, and require fewer samples, relative to the conventional design. The intradialytic technique also eliminates the discomfort associated with repeated venipuncture, as the samples may be drawn directly from the dialysis machine, or dialysate.

email: minchun.zhou@vanderbilt.edu

### 11h. A Study on the Statistical Properties of the European Pharmacopoeia Test for Uniformity of Dosage Units Using Large Sample Sizes

**Meiyu Shen\***, U.S. Food and Drug Administration  
**Yi Tsong**, U.S. Food and Drug Administration  
**Xiaoyu Dong**, U.S. Food and Drug Administration

European Pharmacopoeia test for uniformity of dosage units using large sample sizes in European Pharmacopoeia 7.7 has two alternative tests. Option 1 is a parametric two-sided tolerance interval based method with an indifference zone and counting outside of (0.75M, 1.25M), here M is defined by sample mean,  $\bar{X}$ , as:  $M = 98.5\%$  if  $\bar{X} < 98.5\%$ ,  $M = 101.5\%$  if  $\bar{X} > 101.5\%$ , and  $M = \bar{X}$  otherwise. Option 2 is a nonparametric counting method with an additional indifference zone concept. We extend the parametric two one-sided tolerance interval based method originally proposed for dose content uniformity test for thirty tablets by Tsong to large sample sizes with restriction such that all operating characteristic (OC) curves of two one-sided tolerance intervals for any sample size intersect with that of the USP harmonized method with a sample size of 30

at the acceptance probability of 90% when the individual tablets are normal variables with on-target mean. We study the acceptance probabilities in relation to lot mean and lot standard deviation among two EU options and our method. Our simulations show that two EU options introduce larger acceptance probabilities than our method for off target lot mean.

email: meiyushen@yahoo.com

### 11i. Variable Selection when Some Predictors are Measured with Error

**Guangning Xu\***, North Carolina State University  
**Leonard A. Stefanski**, North Carolina State University

A fundamental problem in biomedical research is identifying key risk factors and determining their impact on health outcomes via statistical modeling. However, due to device limitations and within-subject variation, some risk factors are measured with error, e.g., blood pressure. Ignoring measurement error adversely impacts variable selection and model fitting. Thus it is desirable to develop a variable selection method that takes the measurement error into account. We propose a new method for variable selection in measurement error models by transforming the observed noisy data to a new dataset, so that a non-measurement error model analysis of the transformed data gives exactly the same measurement error model estimates of the original data. We use two strategies to transform the data, method of moments and conditional score. After transformation, existing non-measurement error variable selection methods can be applied to the transformed data directly. The key advantage of our new method is that it is conceptually simple and greatly eases computing by using existing algorithms.

email: gxu@ncsu.edu

### 11j. Variable Selection for Optimal Treatment Regimes

**Na Zhang\***, North Carolina State University  
**Eric Laber**, North Carolina State University  
**Howard Bondell**, North Carolina State University

Personalized medicine aims to tailor treatments to the individual based on measured characteristics. Choosing among the set of possible tailoring variables can be a daunting task. The goal is to determine the subset of variables that will yield the decision rule that results in the optimal treatment regime over the population. Unlike typical variable selection problems, variables that may be highly related to the clinical outcome may not be relevant to distinguish among competing treatments. We propose a two-stage approach by first obtaining a flexible fit to the potential outcomes under each treatment regime to determine an estimate of the optimal treatment for each subject. In the second step, we then perform a sparse classification method to perform variable selection for this optimal decision rule. We show that this approach can not only identify the relevant tailoring variables, but, in the process, the reduction in dimension yields better accuracy in classification of subjects to the optimal treatment.

email: nzhang@ncsu.edu

### 11k. Promoting Similarity of Model Sparsity Structures in Integrative Analysis

**Yuan Huang\***, The Pennsylvania State University  
**Runze Li**, The Pennsylvania State University  
**Jian Huang**, University of Iowa  
**Shuangge Ma**, Yale University

In high-throughput studies, single-dataset analysis often leads to unsatisfactory results because of the small sample sizes. In this study, we conduct integrative analysis and marker selection under the heterogeneity model, which postulates that different datasets have possibly overlapping but not necessarily identical sets of markers. Under certain scenarios, it is reasonable to expect similarity of identified marker sets - or equivalently, similarity of model sparsity structures - across multiple datasets. However, the existing methods do not have a mechanism to explicitly promote such similarity. To solve this problem, we develop a novel sparse boosting method. This method uses a BIC/HDBIC criterion to select weak learners and encourage sparsity. A new penalty is introduced to promote similarity of model sparsity structures across datasets. The proposed method has an intuitive formulation and is generically applicable and computationally affordable. Simulation shows that the proposed method outperforms alternatives with more accurate marker identification.

email: huangyuan.stat@gmail.com

## 12. MASSIVE ONLINE OPEN STATISTICS (MOOS): SHOULD WE BE TEACHING STATISTICS TO 100,000S OF THOUSANDS AT A TIME?

**MOOCs for Statistics and the Statistics of MOOCs**  
**Joseph Blitzstein\***, Harvard University

The advent of massive open online courses (MOOCs) raises many challenges and opportunities for both research and teaching in statistics. We will discuss this both in the context of personal experience teaching and designing such courses, and in the context of preliminary data from some courses on the edX platform.

email: blitz@fas.harvard.edu

**Can We Teach 100,000 People Data Analysis at Time?**

**Jeffrey T. Leek\***, Johns Hopkins  
Bloomberg School of Public Health

Massive online open courses (MOOCs) have received substantial attention in the popular press over the last two years (Chronicle, NYTimes, TED, Washington Post). A major source of this interest has been in statistical courses. The course I taught, Data Analysis, has attracted over 150,000 students. I will discuss the challenges of scaling data analysis education to thousands of students, the available tools for automated grading, peer grading, and discussion forums. I will also consider multiple definitions of participation and attrition, designed to capture the diverse backgrounds and

diverse goals of the students who take these courses. I will also present data on participation and attrition, and usage patterns in general for my course. A primary focus will be a discussion of the pros and cons of the MOOC platform for statistics education, particularly of health sciences students, and online education in general.

email: jtleek@gmail.com

### **Statistical Reasoning for the Masses**

**John McGready\***, Johns Hopkins Bloomberg School of Public Health

Statistical Reasoning in Public Health 1 is an eight week primer on the basics of statistical thinking as filtered through the concepts of EDA, estimation and inference. The target audience is first year public health graduate students and the content draws upon example from the current public health and medical literature. The course has been taught online for 13 years at Hopkins to audiences as large as 250 enrollees. As of January 2014, the ante will be (potentially) upped and the course will make its Coursera debut. At the time of this session the course will still be running. Up to the minute (or close to it) data on the enrollees, attrition thus far, and performance on the assessment items will be presented. As the other presentations will focus on more mathematically rigorous MOOS, this piece will shed light on the challenges and opportunities of teaching statistical reasoning to a potentially more heterogeneous audience in terms of mathematical skill sets, and familiarity with the basics of statistics.

e-mail: jmcgread@jhsph.edu

### **Massive Online Open Statistics (MOOS): Should We be Teaching Statistics to 100,000 at a Time?**

**Rebecca Nugent\***, Carnegie Mellon University

This discussion will tie together the themes heard in the previous talks on MOOCs with an emphasis on advantages, disadvantages, and next steps. It will also include an overview of the work in this area being done by several departments at Carnegie Mellon University and what we're hoping to tackle over the next few years in Statistics.

email: rnugent@andrew.cmu.edu

## **13. COUNCIL FOR EMERGING AND NEW STATISTICIANS (CENS) INVITED SESSION: SHOULD I DO A POSTDOC?**

### **My Experiences as a Postdoc in Biostatistics**

**Joshua Warren\***, University of North Carolina, Chapel Hill

Postdoctoral positions have long been a common career step in the physical sciences, often being required in certain fields before entry into a research oriented career. Within the past 25 years, postdoctoral positions in statistics and biostatistics have become more common and the level of interest from employers and potential applicants

continues to increase. In this talk, I discuss my experience as a postdoctoral researcher in the Department of Biostatistics at the University of North Carolina at Chapel Hill ranging in time from the application process as a graduate student to the end of the postdoc appointment. I also present an overview of reasons one might consider a postdoctoral position in statistics/biostatistics, how to get the most out of your time as a postdoc, and other general features of life as a postdoctoral researcher.

email: joshuawa@email.unc.edu

### **Is Post-Doctoral Fellowship Key to Academic Success?**

**Hemant K. Tiwari\***, University of Alabama, Birmingham

It is becoming essential to do a post-doctoral fellowship to be considered for a tenure track position in many disciplines. Most of the tenure-track positions have three requirements for tenure: (1) Scholarship, (2) Teaching, and (3) Service. Postdoctoral fellowships usually vary from 2 to 3 years in most of the institutions, and provide ample time to develop as a researcher without the pressure of a ticking tenure clock. The tenure clock starts the day you start your job as a tenure track Assistant Professor. There are several additional advantages of being a post-doctoral fellow including the following: learning how to evaluate others research, time to develop new research areas of expertise or extend your existing areas of research, time to transition from student to professional life, time to create your own thematic research, networking, and build your CV with more publications, etc. Also, the post-doctoral fellows have better chance of getting a good academic position compared to fresh PhDs, since employers are usually hesitant to hire students with very little experience. In this presentation, I will draw from my own experiences as a post-doctoral fellow and also as a mentor and Director of the Post-Doctoral Program in Statistical Genetics/Genomics funded by NHLBI at UAB.

email: htiwari@uab.edu

### **Should I do a Post-Doctoral Fellowship? The NICHD Experience**

**Paul S. Albert\***, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

The Biostatistics & Bioinformatics branch at the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) has an active post-doctoral fellowship program. We currently have four post-doctoral fellows working in areas ranging from statistical genetics, Bayesian methodology, longitudinal data, and innovative clinical trial methodology. Most fellows work primarily on developing new statistical methodology related to substantive problems in population health as it relates to pregnant women, fetal development, child health, and the life course. Our fellows have recently published in leading journals of biostatistics including: JRSS-C, JASA, Biometrics, Biostatistics, Statistical Methods in Medical Research and Statistics in Medicine. Our fellowship program provides recent graduates the opportunity to begin their career in

a stimulating data rich research environment where their primary responsibility is methodological research. We believe that such an experience will provide young statistical scientists a good start to their research career.

email: albertp@mail.nih.gov

## 14. ADAPTIVE RANDOMIZED TRIAL DESIGNS AND IMPROVED ANALYSIS METHODS TO LEARN WHICH SUBPOPULATIONS BENEFIT FROM WHICH TREATMENTS

### Impacts of Predictive Genomic Classifier Performance on Subpopulation-Specific Treatment Effects Assessment

**Sue-Jane Wang\***, U.S. Food and Drug Administration  
**Ming-Chung Li**, National Cancer Institute, National Institutes of Health

We consider predictive biomarker scenarios with varying prevalence as such there is no treatment effect in the biomarker negative patient subpopulation. Using a Breiman machine learning voting algorithm (1996) via a k-fold cross-validated approach applied by Freidlin et al. (2010), a predictive biomarker may be developed. We consider development or discovery of a genomic biomarker using microarray gene expressions data in randomized controlled trials and validate the biomarker's predictive performance in an independent dataset. We investigate the classification performance characteristics of a binary genomic composite biomarker expected to be predictive of treatment effects and evaluate approaches to improve sensitivity when a biomarker is highly specific but poorly sensitive. We explore when it will be beneficial to develop a binary predictive biomarker. In addition, we compare the predictive performance of a biomarker classifier between use of direct selection and selection from a candidate pool shedding favorable lights of direct selection approach where biological or mechanistic plausibility can be relied upon. Further research is needed if accurate classifier is required irrespective of prevalence level.

email: suejane.wang@fda.hhs.gov

### Shine Shadow: A Bayesian Adaptive Trial Vs. a Group Sequential Trial in Stroke

**Jason T. Connor\***, Berry Consultants  
**Kristine R. Broglio**, Berry Consultants  
**Valerie L. Durkalski**, Medical University of South Carolina

We compare an NIH-sponsored group sequential design for the treatment of stroke with an innovative Bayesian adaptive Goldilocks trial. The Adaptive Designs Accelerating Promising Trials into Treatments (ADAPT-IT) project is a collaborative effort supported by the NIH and FDA to explore how adaptive clinical trial design might improve the evaluation of drugs and medical devices. We use the NINDS-supported Neurological Emergencies Treatment Trials network as a 'laboratory' in which to study the development of adaptive clinical trial designs. The Stroke Hyperglycemia Insulin Network Effort (SHINE) trial was fully funded by the NIH-

NINDS at the start of ADAPT-IT and is a currently ongoing phase III trial of tight glucose control in hyperglycemic acute ischemic stroke patients. Within ADAPT-IT, a Bayesian adaptive alternative design was developed. The primary endpoint is a severity-adjusted, dichotomized 90-day modified Rankin scale (mRS). We present both designs and compare their operating characteristics. The alternative design will be retrospectively executed upon completion of SHINE to later compare the designs based on their use of patient resources, time, and strength of conclusions in a real world setting.

email: jason@berryconsultants.com

### Adaptive Enrichment Designs for Clinical Trials

**Noah Simon\***, University of Washington  
**Richard Simon**, National Cancer Institute, National Institutes of Health

With the advent of new high-throughput biotechnologies, we are better able to understand the mechanisms which cause and characterize different diseases. These technologies have led to the discovery of vast heterogeneity in the mechanisms and molecular pathology of previously considered homogeneous diseases. Leveraging these heterogeneities, we have graduated from broad spectrum treatments to targeted therapeutics. These therapeutics target specific pathways and are generally only beneficial to a small subset of a heterogeneous diseased population. Often this subset is poorly understood until well into large-scale clinical trials. As such, standard practice has been to broadly enroll patients and run post-hoc subset analysis to determine those patients who particularly benefit. This unnecessarily exposes many patients to hazardous side-effects and may vastly decrease the efficiency of the trial. In this talk I will propose a simple class of adaptive enrichment designs which allow the eligibility criteria of a clinical trial to be adaptively updated during the trial, restricting entry to patients likely to benefit. These designs both preserve type 1 error and in many cases provide a substantial increase in power. I will give some specific designs and use recommendations. I will also discuss challenges inherent in this approach, and some potential solutions.

email: nrsimon@uw.edu

### Constructing Confidence Sets for the Optimal Regime

**Sherri Rose\***, Harvard University  
**Tuo Zhao**, Johns Hopkins University  
**Han Liu**, Princeton University  
**Michael Rosenblum**, Johns Hopkins University

It is frequently beneficial to define treatment "rules" over time (often referred to as treatment regimes) in order to identify optimal patient outcomes. In this presentation, we will introduce a new algorithm for the construction of valid confidence sets for the optimal regime among a set of finite regimes. We accomplish this by using convex optimization methods and present simulation studies based on previously collected data.

email: rose@hcp.med.harvard.edu

## 15. STATISTICAL METHODS FOR COMPLEX STRUCTURED BIOMEDICAL OBJECT DATA

### Object Oriented Data Analysis: Backwards PCA

**J. S. Marron\***, University of North Carolina, Chapel Hill

Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Challenges in modern medical image analysis motivate the statistical analysis of populations of more complex data objects which lie in a curved manifold. In such contexts the backwards approach to PCA has proven to be very successful in a wide variety of setting. The reason for this is explained by representing analogs of PCA in terms of constraints.

email: marron@unc.edu

### Additive and Interaction Models for Nonparametric Functional and Object Regression, with Application to Ophthalmological Multi-Level Functional Data on Spherical Domains

**Jeffrey S. Morris\***, University of Texas MD Anderson Cancer Center

**Veera Baladandayuthapani**, University of Texas MD Anderson Cancer Center

**Massimo Fazio**, University of Alabama, Birmingham

Glaucoma is a condition involving optic nerve damage from interocular pressure (IOP). Our collaborators have developed a novel system for inducing IOP and measuring displacement and maximal principal strain (MPS) on a fine grid of the outer scleral surface of the eye. This yields multi-level functional data on the sphere, with functions from each of 9 IOPs per eye, two eyes per subject. We introduce new methods for these data that allow a nonparametric effect of age on MPS freely varying over the scleral surface, borrowing strength from nearby positions on the sphere in terms of estimation, variability, and degrees of freedom of the nonparametric fit. Further, our model also includes an IOP-age interaction that allows the nonparametric age effect to vary over IOP, and a growth curve component to capture the longitudinal correlations across different IOPs for a given eye. Our analysis demonstrates MPS is greatest near the optic nerve, where there is a clear decreasing trend in MPS with age that accelerates around age 65. This work greatly extends the functional mixed model framework to incorporate additive model and growth curve ideas for regression analyses involving responses that are functions or objects on some fixed domain.

email: jefmorris@mdanderson.org

### On Synergy Between Statistical Shape Analysis (SSA) and Functional Data Analysis (FDA)

**Anuj Srivastava\***, Florida State University

**Sebastian Kurtek**, The Ohio State University

**Eric Klassen**, Florida State University

**Jingyong Su**, Texas Tech University

The statistical shape analysis (SSA) has traditionally been formulated as analysis of a set of landmarks (registered points) modulo certain similarity transformations (rotation, translation, and scale). More recently SSA has been extended to include shapes of continuous objects -- parameterized curves, surfaces, and their temporal evolutions -- by treating them as elements of Hilbert spaces. Functional data analysis (FDA) also deals with generating inferences on certain Hilbert spaces and shares some common issues and solutions with SSA. Specifically, the problem of phase-amplitude separation in FDA is similar to optimal registration of points in SSA. An elegant solution to these problems, both in SSA and FDA, comes from extending Fisher-Rao metrics and the use of a family of square-root transforms of the original functions or objects. The most important property of this setup is that simultaneous warping, or re-parameterizations, of functions/objects do not change the L2 distance between their square-root transforms. I will describe this framework using examples from FDA and SSA of curves in Euclidean and non-Euclidean spaces.

email: anuj@stat.fsu.edu

### Bayesian Spatial Functional Models for High-Dimensional Genomics Data

**Veerabhadran Baladandayuthapani\***, University of Texas MD Anderson Cancer Center

**Lin Zhang**, University of Texas

MD Anderson Cancer Center

**Jeffrey Morris**, University of Texas

MD Anderson Cancer Center

**Keith Baggerly**, University of Texas

MD Anderson Cancer Center

Many scientific applications generate correlated functional data, where it is often of interest to flexibility model the dependence patterns. We present methods that focus on spatial functional data analysis where spatial dependence is present in high-dimensional functional data. Our methods allow for simultaneous characterization of these high-dimensional functions using non-parametric basis functions, joint modeling of spatially correlated functional data and detection of local features in spatially heterogeneous functional data to answer several important biological questions. Our methods are motivated by and applied to a high-throughput copy number dataset generated through whole-organ histologic genomics maps of bladder cancer development. Our model identifies several genetic markers with plaque-like copy number alterations that are potentially associated with development of bladder cancer, which were not discovered by methods that ignore the dependence.

email: veera@mdanderson.org

## 16. MULTIVARIATE ANALYSIS IN HIGH DIMENSIONS

### Laplacian Shrinkage for Estimation of Inverse Covariance Matrices from Heterogenous Samples

**Takumi Saegusa**, University of Washington  
**Ali Shojaie\***, University of Washington

We introduce a general framework, using a Laplacian shrinkage penalty, for estimation of inverse covariance, or precision matrices from heterogenous samples. The proposed framework encourages similarity among disparate, but related, subpopulations while allowing for differences among estimated matrices. We propose an efficient alternating direction method of multiplier (ADMM) algorithm for parameter estimation, and establish both variable selection and norm consistency of the estimator for distributions with exponential or polynomial tails. Finally, we discuss the selection of the Laplacian shrinkage penalty based on hierarchical clustering among samples, and discuss conditions under which this data driven choice results in consistent estimation of precision matrices. Extensive numerical studies and applications to gene expression data from subtypes of cancer with distinct clinical outcomes indicate the potential advantages of the proposed method over existing approaches.

email: ashojaie@uw.edu

### Joint Mean-Covariance Models for Incomplete Multivariate Longitudinal Data

**Mohsen Pourahmadi\***, Texas A&M University

Multivariate longitudinal studies are common in biological, medical and social sciences. We deal with two main challenges in this area: modeling the correlations and handling the missing data. The unconstrained parameterization of the covariance matrix for multivariate longitudinal data based on its block Cholesky decomposition is not directly applicable in the presence of missing values, because a coherent Cholesky factorization may not exist for all subjects. We develop a framework to avoid this complication and employ a generalized EM algorithm for estimating the parameters of the generalized linear mean-covariance models. We derive the score functions and Fisher information matrices for the parameters. The performance of the proposed estimators is illustrated through extensive simulation studies and an application to real data. (Joint work with Priya Kohli and Tanya Garcia).

email: pourahm@stat.tamu.edu

### Prediction in Abundant High-Dimensional Linear Regression

**Dennis Cook\***, University of Minnesota  
**Liliana Forzani**, Instituto de Matemática Aplicada del Litoral and Facultad de Ingeniería Química CONICET and UNL  
**Adam J. Rothman**, University of Minnesota

An abundant regression is one in which most of the predictors contribute information about the response, which is contrary to the common notion of a sparse regression where few of the predictors are relevant. We discuss asymptotic characteristics of methodology for prediction in abundant linear regressions as the sample size and number of predictors increase in various alignments. We show that some of the estimators can perform well for the purpose of prediction in abundant high-dimensional regressions.

email: rdcook@umn.edu

### Properties of Optimizations Used in Penalized Gaussian Likelihood Inverse Covariance Matrix Estimation

**Adam J. Rothman\***, University of Minnesota  
**Liliana Forzani**, Instituto de Matemática Aplicada del Litoral and Facultad de Ingeniería Química CONICET and UNL

We establish necessary and sufficient conditions for the existence of inverse covariance matrix estimates obtained by minimizing the negative Normal log-likelihood plus a weighted ridge or weighted L1 penalty. A new algorithm to solve this optimization with the weighted ridge penalty is developed and its convergence is established. This algorithm combines the majorize minimize principle with minorize minimize acceleration attempts. Numerical experiments show this algorithm is superior to its only competitor and that ridge penalization is useful within quadratic discriminant analysis.

email: arothman@umn.edu

## 17. RECENT ADVANCES IN LIFETIME DATA ANALYSIS

### Bayesian Threshold Regression for Informatively Censored Current Status Data

**Tao Xiao**, University of Maryland, College Park and The Ohio State University  
**Michael L. Pennell\***, The Ohio State University

In some biomedical studies, there is interest in making inferences about a time to event distribution but the exact time of the event is unknown. For instance in carcinogenicity studies in animals, tumors are not discovered until the time of examination and hence time to tumor is left censored; this is known as current status data. Sometimes, the examination time is not independent of the event time. For example, in an animal study, an exam may have occurred because the animal died due to a cause related to tumor development. In this case, survival analysis methods which assume

independent censoring would result in biased inferences. To address this issue, we propose a Bayesian approach which jointly models time to event and time to censoring using latent Wiener processes which fail once they hit a boundary value. Using data augmentation, we sample the unobserved event time and values of the latent processes for those subjects who do not experience an event. Informative censoring is accounted for by modelling time to censoring using two latent health processes: one is independent of the event of interest and the other dependent. In addition to being a conceptually appealing model, our model does not require the assumption of proportional hazards of some standard methods. We demonstrate our method using time to lung tumor data from a National Toxicology Program study.

email: mpennell@cph.osu.edu

### **Semiparametric Estimation for the Additive Hazards Model with Left-Truncated and Right-Censored Data**

**Chiung-Yu Huang\***, Johns Hopkins University  
**Jing Qin**, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Survival data from prevalent cases collected under a cross-sectional sampling scheme are subject to left-truncation. When fitting an additive hazards model to left-truncated data, the conditional estimating equation method (Lin & Ying, 1994), obtained by modifying the risk sets to account for left-truncation, can be very inefficient, as the marginal likelihood of the truncation times is not used in the estimation procedure. In this paper, we use a pairwise pseudolikelihood to eliminate nuisance parameters from the marginal likelihood, and by combining the marginal pairwise pseudo-score function and the conditional estimating function, propose an efficient estimator for the additive hazards model. The proposed estimator is shown to be consistent and asymptotically normally distributed with a sandwich-type covariance matrix that can be consistently estimated. Simulation studies show that the proposed estimator is more efficient than its competitors. A data analysis illustrates the method.

email: cyhuang@jhmi.edu

### **Semiparametric Inference on the Absolute Risk Reduction and the Restricted Mean Survival Difference**

**Song Yang\***, National Heart, Lung, and Blood Institute, National Institutes of Health

For time-to-event data, when the hazards may be non-proportional, in addition to the hazard ratio, the absolute risk reduction and the restricted mean survival difference can be used to describe the time-dependent treatment effect. The absolute risk reduction measures the direct impact of the treatment on event rate or survival, and the restricted

mean survival difference provides a way to evaluate the cumulative treatment effect. However, in the literature, available methods are limited for flexibly estimating these measures and making inference on them. We propose to study these measures under a semiparametric model that can be used in a sufficiently wide range of applications. Point estimates, point-wise confidence intervals and simultaneous confidence bands of the absolute risk reduction and the restricted mean survival difference are established. These methods are motivated by and illustrated for data from the Women's Health Initiative estrogen plus progestin clinical trial.

email: yangso@nhlbi.nih.gov

### **Evaluating Calibration of Risk Prediction Models**

**Ruth Pfeiffer\***, National Cancer Institute, National Institutes of Health

Statistical models that predict disease incidence, disease recurrence or mortality following disease onset have broad public health and clinical applications. Before a model can be recommended for practical use, its performance characteristics need to be understood. General criteria to evaluate prediction models for dichotomous outcomes include predictive accuracy, proportion of variation explained, calibration and discrimination. Most recent validation studies have emphasized calibration and discrimination. A model is called well calibrated (or unbiased) when the predicted probabilities agree with observed risk in subsets of the population and overall. I propose and study novel criteria to assess the calibration of models that predict risk of disease incidence and compare their performance to standard methods to assess model calibration. I illustrate the methods with models that predict incidence of endometrial and breast cancer.

email: pfeiffer@mail.nih.gov

## **18. EPIDEMIOLOGIC METHODS**

### **Modeling Epidemiological Features of Disease Outbreaks**

**Manasi Sheth-Chandra\***, Booz Allen Hamilton  
**N. Rao Chaganty**, Old Dominion University

The purpose of monitoring disease outbreaks in non-hostile situations is to identify important medical events or trends early so preventive medicine investigation or countermeasures can be adjudicated. Large obvious disease outbreaks can indicate a problem exists which could negatively impact public health management and clinical outcomes. Usually, over-dispersed counts of disease outbreaks consists of certain values occurring more than frequently allowed by common parametric family of distributions. Lambert (Technometrics, 1992, pp. 1-14) proposed the zero-inflated Poisson (ZIP) regression model for dealing with zero-inflated count data. We introduce Doubly Inflated Poisson (DIP) models for count data situations where there is another inflated value  $k > 0$  along with their distribution properties. For the disease outbreaks data consisting of un-grouped as well as grouped frequencies,

with and without covariates, we discuss parameter estimation using maximum likelihood (ML) and method of moments. Asymptotic and small sample comparisons show that the ML estimators are far superior than the moment estimators. Parameter estimation and analysis of count data with a negative binomial data-generating process will also be presented using a Doubly Inflated Negative Binomial (DINB) Model.

e-mail: manasi2580@hotmail.com

### **A Stochastic Model for Explicit Estimation of Effect Modification In Finite Sample**

**Xiaoshan Wang\***, Forsyth Institute  
**Jacqueline Starr**, Forsyth Institute

In the conditional likelihood scenario, effect modification is usually modeled by adding an interaction term. But our analysis suggests that conditional likelihood leads to information loss. And such information loss results in less efficiency, and thus reduced sensitivity in detecting effect modification when we have a finite sample. In this study, we propose a stochastic logistic model based on multivariate Bernoulli distribution, through which the effect modification can be explicitly estimated without a need of interaction term. Simulation indicates an improved capability to detect effect modification in finite sample.

e-mail: xwang@forsyth.org

### **Modeling the Effects of Climate Change and Air Quality on Asthma, Accounting for Uncertainty**

**Stacey E. Alexeeff\***, National Center for Atmospheric Research  
**Stephan R. Sain**, National Center for Atmospheric Research  
**Doug Nychka**, National Center for Atmospheric Research

Climate change is expected to have many impacts on public health and the environment, including changes in air pollution. High-resolution regional climate change projections and air quality models are used to quantify the subsequent impacts on asthma-related health effects. Current relationships between ozone concentrations and asthma exacerbations and hospitalizations are assessed by meta-analysis. Two key sources of uncertainty are in the climate projections themselves and in the relationship between air quality and asthma. Bayesian hierarchical models provide a statistical relationship between asthma and future air pollution levels, and naturally allow the propagation of uncertainty through to public health outcomes. We estimate the national change in asthma-related health effects attributable to the change in future ground-level ozone levels from the 2000's to the 2050's.

e-mail: salxeeff@ucar.edu

### **One Novel Approach to Handle Random Measurement Error Using Hidden Markov Models**

**Lola Luo\***, University of Pennsylvania  
**Dylan Small**, University of Pennsylvania  
**Jason A. Roy**, University of Pennsylvania

A major goal in studying the progression of incurable chronic diseases such as chronic kidney disease (CKD) is to estimate the transition rates between disease stages. One way this can be estimated is with a hidden Markov model (HMM). In a HMM, the observed disease stages are linked with the true disease states using a state-dependent probability matrix. The transition rates between the true states is typically of primary interest. However, these models depend on strong assumptions about the observed data. For example, observed disease stages should not differ much for a given subject in a short period of time. However, this assumption will sometimes be violated in healthcare databases, which tend to have higher error rates than do data that were collected as part of a research study. In this paper, we extend HMMs to accommodate data that are contaminated with some unusual values. We assume that the observed stages are from a contaminated distribution, where there is probability ( $\eta$ ) of being accurate and  $(1-\eta)$  of being an error. The method allows estimation both of the contamination rate and of the transition probabilities among the true disease states. Simulation studies have shown that failure to account for the random measurement error when it exists in the data will produce bias and affect the coverage probability of the model.

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

### **Quantifying Circadian Trajectory of Fatigability Using the Proportional Intensity Model**

**Jiawei Bai\***, Johns Hopkins University  
**Jennifer Schrack**, Johns Hopkins University  
**Mei-Cheng Wang**, Johns Hopkins University  
**Luigi Ferrucci**, National Institute of Aging, National Institutes of Health  
**Ciprian M. Crainiceanu**, Johns Hopkins University

Fatigue is a subjective description of lack of physical and/or mental energy or will required to perform activities of daily living. Thus, fatigue is both hard to define and/or estimate in spite of its importance for public health. In this paper we introduced a new model-based definition of subject-specific fatigability. More precisely, we define "fatigability" as the instantaneous inverse intensity of transitioning from non-movement to movement given the subject's specific history of movement and covariates. Methods are motivated by and applied to the Baltimore Longitudinal Study of Aging (BLSA), which collected minute-by-minute accelerometry data on more than 600 subjects for up to 7 days. Our results indicate that there is a gradient of fatigability with respect to age and time of day.

e-mail: jbai@jhsph.edu

## **A Comparison of Methods for Biomarker Associations with Endogenous Treatment**

**Andrew J. Spieker\***, University of Washington  
**Joseph AC Delaney**, University of Washington  
**Robyn L. McClelland**, University of Washington

One challenge in cross-sectional observational data is determining biomarker associations (e.g., age and blood pressure) when the biomarker may be altered by a treatment or therapy. Part of the difficulty in quantifying these associations is that treatment is endogenous to risk factors and the untreated biomarker value itself. Naïve methods, although widely used, do not take this into account: these include (i) ignoring the issue, (ii) excluding treated participants from analysis, and (iii) including a treatment covariate. Propensity score adjustment can modestly improve bias, but increases the variability of the estimates. A censored normal approach may be useful, but can often overestimate associations in the presence of only modest treatment effects. A hybrid model, developed by Heckman in 1978, is a likelihood-based approach which jointly models the outcome and the treatment. In a simulation study, we show that when model assumptions are met, the Heckman model provides valid estimates and inference, while the alternative models frequently fail. In practice, the assumptions of a correctly specified model, normally distributed errors, and linear treatment effect are often violated. We show with various sensitivity analyses that the Heckman model fares well even in the presence of some moderate assumption violations.

e-mail: ajspiek@uw.edu

## **Modeling Temporal Patterns in Exposure/Response Relationships with Change Points, with an Application to Incident Obstructive Airway Disease in Firefighters Exposed to the World Trade Center Rescue/Recovery Effort**

**Charles B. Hall\***, Albert Einstein College of Medicine of Yeshiva University  
**Michelle Glaser**, Montefiore Medical Center  
**Mayris Webber**, Montefiore Medical Center  
**Xiaoxue Liu**, Montefiore Medical Center  
**Rachel Zeig-Owens**, Montefiore Medical Center  
**David Prezant**, Fire Department of the City of New York

Adverse respiratory effects of exposure to the World Trade Center (WTC) disaster site have consistently been demonstrated. We model relative incidence with respect to exposure intensity over the first five years using change points. Obstructive airway disease (OAD) was diagnosed by Fire Department of the City of New York physicians. Exposure was categorized by time of arrival for work at the WTC site as follows: (high) morning 9/11/2001 (n=1,451); (moderate) afternoon 9/11- 9/12/2001 (n=6,506); (low) 9/13-24/2001 (n=1,083). Exponential survival models fit by maximum likelihood were used to estimate relative incidences by exposure group, with change points in the relative incidences estimated via profile likelihood. One change point at 15 months was observed, with relative incidences (high vs. low exposure) of 3.90 (95% CI 2.49-6.10) prior to 15 months and 1.79 (95% CI 1.28-2.48) thereafter. Trend tests before and after the change point were significant ( $p < 0.001$ ).

Analyses by OAD subtypes (asthma and chronic bronchitis/COPD) were similar. These results show that change point models are a valuable method for evaluating changes in exposure/response relationships over time.

e-mail: charles.hall@einstein.yu.edu

## **19. COMPUTATIONAL METHODS AND IMPLEMENTATION**

### **Performance of Shannon's Maximum Entropy Distribution Under Some Restrictions**

**Sinan Saracli\***, Afyon Kocatepe University  
**Hatice Cicek**, Afyon Kocatepe University

Entropy has a very important role in Statistics. In recent studies it can be seen that entropy started to take place nearly in every branches of sciences. In information theory, entropy is a measure of the uncertainty in a random variable. While there are different kinds of methods in entropy, the most common MaxEnt method maximizes the Shannon's entropy according to the restrictions which are obtained from the random variables. MaxEnt distribution is the distribution which is obtained by this method. While calculating the maximum entropy, Claude Shannon, the inventor of information theory, used logarithmic transformations to make it as a cumulative quality, concordantly in this study first of all, entropy distributions are obtained by the help of Lagrange multipliers, under some restrictions like moments of the distribution, then among these entropy distributions, the distribution which has the max entropy value is used as an entropy distribution. The accordance between probability distribution of a data set and this distribution is examined via Chi-Square test. The results are given in related tables.

email: ssaracli@aku.edu.tr

### **Propensity Score Matching with Survival Outcomes: Critical Considerations in the Choice of the Caliper Size**

**Adin-Cristian Andrei\***, Bluhm Cardiovascular Institute, Northwestern University  
**Zhi Li**, S. Bluhm Cardiovascular Institute, Northwestern University  
**Chris Malaisrie**, Bluhm Cardiovascular Institute, Northwestern University  
**Edwin McGee**, Bluhm Cardiovascular Institute, Northwestern University  
**Jane Kruse**, Bluhm Cardiovascular Institute, Northwestern University  
**Patrick M. McCarthy**, Bluhm Cardiovascular Institute, Northwestern University

Propensity score (PS) matching methods have become extremely popular in observational studies. Areas of application are continually expanding and include cardiac surgery outcomes, epidemiology or A/B testing in online marketing and advertising. Of high importance for the quality of the PS matching process are several aspects, including the appropriateness of the PS model and the

choice of the caliper size. Based on both simulated and actual data, we discuss several in-depth considerations involving the bias/variance trade-off in studies with survival outcomes. We explore how the robustness of the results is affected by these choices.

email: aandreim@nmh.org

### **Model Free Variable Rank Using Randomized Decision Tree, an Ensemble of Trees**

**Bong-Jin Choi\***, University of South Florida  
**Chris P. Tsokos**, University of South Florida

Tree-based methods have become popular for performing survival analysis with complex data structures. Within the Random Forest (RF), we applied decision tree analysis (DTA) conjunction with bootstrapping, counting method, and weighted factor method to identify and rank variables that significantly contribute to estimating the survival time of given cancer patient. The present study is concerned with difficult data which we cannot find a good statistical model, but we want to know the significant attributable variables and rank of variables for a subject study. The weighted factor method is quite effective method in comparison to the counting method. In addition, we performed to compare our result and classical counting method.

email: behappygene@gmail.com

### **Optimal Computational and Statistical Rates of Convergence for Sparse Nonconvex Learning Problems**

**Zhaoran Wang\***, Princeton University  
**Han Liu**, Princeton University  
**Tong Zhang**, Rutgers University

We provide statistical and computational analysis of nonconvex penalized  $\ell_1$ -estimators. For nonconvex problems, it is intractable to compute the global solution. In this paper, we propose an approximate regularization path following algorithm for solving a variety of nonconvex learning problems. Under a unified analytic framework, we simultaneously provide explicit statistical and computational rates of convergence of any local solution obtained by the algorithm. Computationally, our algorithm attains a global geometric rate of convergence for calculating the full regularization path, which is optimal among all first-order algorithms. Unlike most existing methods which only attain geometric rates of convergence for one single regularization parameter, our algorithm calculates the full regularization path with the same iteration complexity. In particular, we provide a refined iteration complexity bound to sharply characterize the performance of each stage along the regularization path. Statistically, we provide sharp sample complexity analysis for all the approximate local solutions along the regularization path. In particular, our analysis improves upon existing results by showing a more refined sample complexity bound for the final estimator. This result shows that the final estimator attains an oracle statistical property due to the usage of nonconvex penalty. Thorough numerical results are provided to back up our theoretical analysis.

email: zhaoran@princeton.edu

### **A Modified EM Algorithm for Regression Analysis of Data with Non-Ignorable Non-Response**

**Yang Zhang\***, University of Pittsburgh  
**Gong Tang**, University of Pittsburgh

Missing data are prevalent in biomedical studies especially in large clinical trials and longitudinal studies where some study subjects are subject to loss of follow-up for various reasons. The often unknown mechanism for the missing data process may be associated with the underlying values. We consider regression analysis of data with missing response values. Standard statistical methods, including likelihood-based methods and weighted estimating equations, require a model for the missing-data mechanism and incorporate it in the estimation and inference. Misspecification of the missing-data model often causes biased estimates and wrongful conclusions. The expectation-maximization (EM) algorithm is an iterative algorithm that is often used to find the maximum likelihood estimate for the likelihood-based methods. In the E-steps, under the premise that the current estimate is consistent, we found that those conditional expectations could be approximated from the empirical data without the need for modeling the missing-data mechanism. Subsequently we proposed a modified EM algorithm regardless of the potential missing-data mechanism. Our simulation studies showed that the parameter estimates had negligible bias and were more efficient than the initial values obtained from external data.

email: zhangyang412@hotmail.com

### **A Computationally Fast and Asymptotically Efficient Approach for the Broken-Stick Model**

**Ritabrata Das\***, University of Michigan  
**Moulinath Banerjee**, University of Michigan  
**Bin Nan**, University of Michigan

The existence of one or more change-points in linear regression problems has significant applications in climate data, economic time series and for modeling biological processes, where the change-points mostly pertain to the onset of biologically important phenomena. Estimation of change-point(s) in a broken-stick model using the exact likelihood has been discussed in some depth in the literature, but most of the methods are computationally quite expensive: the non-differentiability at the kink(s) necessitates an exhaustive search across tuples of order statistics. In this article, we present a local smoothing approach to address this difficulty. We smooth the broken-stick in a shrinking neighborhood of the kinks by quadratic functions and use this as our working model, which allows the use of Newton-Raphson type methods for the working likelihood function. Asymptotic properties of our estimates are presented. We find that our estimates converge at root- $n$  rate and are fully efficient. Simulations clearly vindicate the computational economy of our approach with quite remarkable gains in computation times for the two change-points problem. We implement our method on a data set from a crop research experiment.

email: ritob@umich.edu

## 20. NON-PARAMETRIC AND SEMIPARAMETRIC METHODS IN FUNCTIONAL DATA ANALYSIS

### Restricted Likelihood Ratio Tests for Linearity in Scalar-On-Function Regression

**Mathew W. McLean\***, Texas A&M University  
**Giles Hooker**, Cornell University  
**David Ruppert**, Cornell University

We propose a procedure for testing the linearity of a scalar-on function regression relationship. To do so, we use the functional generalized additive model (FGAM), a recently developed extension of the functional linear model. For a functional covariate,  $X(t)$ , the FGAM models the mean response as the integral with respect to  $t$  of  $F\{X(t), t\}$  where  $F(\cdot, \cdot)$  is an unknown bivariate function. The FGAM can be viewed as the natural functional extension of generalized additive models. We show how the functional linear model can be represented as a simple mixed model nested within the FGAM. Using this representation, we then consider restricted likelihood ratio tests for zero variance components in mixed models to test the null hypothesis that the functional linear model holds. The methods are general and can also be applied to testing for interactions in a multivariate additive model or for testing for no effect in the functional linear model. The performance of the proposed tests is assessed on simulated data and in an application to measuring diesel truck emissions, where strong evidence of nonlinearities in the relationship between the functional predictor and the response are found.

email: mathew.w.mclean@gmail.com

### Incorporating Covariates in Skewed Functional Data Models

**Meng Li\***, North Carolina State University  
**Ana-Maria Staicu**, North Carolina State University  
**Howard D. Bondell**, North Carolina State University

Most existing functional data analysis (FDA) models are moment-based, with the focus mostly on the mean function and covariance, without allowing for covariates. However, there are important cases where both higher order moments and the covariate information are of interest. We introduce a semiparametric covariate-adjusted Skewed Functional Model (cSFM) to incorporate covariates within functional data in the presence of skewed distributions. We propose a computationally feasible and stable algorithm to implement the estimation and prediction. Our methodology uses a copula approach to allow the separation of pointwise distributions and the dependence structure. The proposed cSFM provides a unifying platform to estimate model components nonparametrically, obtain pointwise quantiles and make prediction, benefiting from the semiparametric assumptions and the usage of copulas. Through simulation studies, we illustrate the flexibility and the efficiency of the proposed approach. We apply the proposed method to a

tractography study of multiple sclerosis, and figure out a significant interaction effect between the locations along the tract and the selected covariate, as well as a 2-dimensional critical region to reflect key changes between the patients and healthy subjects. An R package has been developed to implement the proposed methodology.

email: mli9@ncsu.edu

### Simultaneous Inference for Repeated Functional Data

**Guanqun Cao\***, Auburn University  
**Lily Wang**, University of Georgia

We develop a new procedure to construct simultaneous confidence bands for mean function(s) of repeatedly observed functional data. In this situation, curves are recorded repeatedly for each subject in a sample and thus they are dependent functional data. The proposed spline simultaneous confidence bands are shown to be asymptotically correct by taking into account the correlation of trajectories within subjects. {The procedure is extended to the two-sample case in which we focus on comparing the mean functions from two populations of functional data}. In addition, we propose a resampling-based method to select an appropriate correlation structure from a given set of candidates. We show the finite sample properties of the proposed confidence bands by simulation studies, and compare the performance of our approach with the "naïve" method which assumes the independence within the repeatedly observed trajectories. This comparison is also illustrated through the analysis of mortality data from period life-tables that are repeatedly collected over years for various countries. Our results indicate that the proposed confidence bands not only achieve the nominal level as suggested by numerical analysis, but also have the desired semiparametric efficiency as shown by our theoretical investigations.

email: gzc0009@auburn.edu

### Generalized Functional Concurrent Model

**Janet S. Kim\***, North Carolina State University  
**Arnab Maity**, North Carolina State University  
**Ana-Maria Staicu**, North Carolina State University

We consider the generalized functional model, where both the response and the covariate are functional data and are observed on the same domain. In contrast to typical functional linear concurrent models, we allow the relationship between the response and covariate to be nonlinear, depending on both the value of the covariate at a specific time point as well as the time point itself. In this framework we develop methodology for estimation of the unknown relationship and construction of point-wise confidence bands, allowing for correlated error structure as well as sparse and/or irregular design. We investigate this approach in finite sample size through simulations and a real data application.

email: jskim3@ncsu.edu

**Variable-Domain Functional Regression****Jonathan E. Gellar\***, Johns Hopkins

Bloomberg School of Public Health

**Elizabeth Colantuoni**, Johns Hopkins

Bloomberg School of Public Health

**Dale M. Needham**, Johns Hopkins School of Medicine**Ciprian M. Crainiceanu**, Johns Hopkins

Bloomberg School of Public Health

We introduce a class of scalar-on-function regression models with subject-specific functional predictor domains. The fundamental idea is to consider a bivariate functional parameter that depends both on the functional argument and on the width of the functional predictor domain. Both parametric and nonparametric models are introduced to fit the functional coefficient. The nonparametric model is theoretically invariant to functional support transformation, or support registration. Methods were motivated by and applied to a study of association between daily measures of the Intensive Care Unit (ICU) Sequential Organ Failure Assessment (SOFA) score and in-hospital mortality. Methods are generally applicable to a large number of new studies that record a continuous variables over unequal domains.

email: jgellar@jhspsh.edu

**Interaction Models for Functional Data****Joseph Usset\***, North Carolina State University**Ana-Maria Staicu**, North Carolina State University**Arnab Maity**, North Carolina State University

We consider a functional regression model with a scalar response and multiple functional predictors that accommodates two-way interactions in addition to their main effects. We develop an estimation procedure where the main effects are modeled using penalized regression splines, and the interaction effect by a tensor product basis. Extensions to generalized linear models and data observed on sparse grids or with error are also presented. Our proposed method can be easily implemented through existing software. Through numerical study we find that fitting an additive model in the presence of interaction leads to both poor estimation performance and lost prediction power, while fitting an interaction model where there is in fact no interaction leads to negligible losses. We illustrate our methodology by analyzing the brain tractography data and the AneuRisk65 study data.

email: jlusset@ncsu.edu

**A Novel Statistical Method Based on Dynamic Models for Classification****Lerong Li\***, University of Texas School of Public Health, Houston**Momiao Xiong**, University of Texas School of Public Health, Houston

Realizations of stochastic processes are often observed temporal data or functional data. There is growing interest in the classification of these types of data. The basic feature of functional data is that the functional data have infinite dimensions and are highly correlated. An essential issue for classifying dynamic data is how to effectively reduce the

number of dimensions and explore dynamic features. We propose to use second-order ordinary differential equations (ODE) to model dynamic processes and to use principal differential analysis to estimate constant or time-varying parameters in the ODE. We examine differential dynamic properties of dynamic systems across different conditions, including stability and transient-response, which determine how the dynamic systems maintain their functions and performance under a broad range of random internal and external perturbations. We use the parameters in the ODE as features for classifiers. As a proof of principle, the proposed methods are applied to classifying normal and abnormal QRS complexes in the electrocardiogram (ECG) data analysis, which is of great clinical value in the diagnosis of cardiovascular disease. We show that the ODE-based classification methods in QRS complex classification outperform the currently widely-used neural networks with Fourier expansion coefficients of the functional data as their features.

email: lerong.li@uth.tmc.edu

**21. STATISTICAL METHODS FOR MICROARRAY AND BIOMARKER DATA****Modeling qRT-PCR Dynamics with Application to Cancer Biomarkers Quantification****Inna Chervoneva\***, Thomas Jefferson University

Quantitative RT-PCR technology is a valuable tool for identification of molecular variations in specific biomarkers between normal, precancerous, cancerous, and metastatic cells that can serve as targets for detection, diagnosis, therapy, and prevention of cancer. The utility of quantifiable mRNA expression of cancer biomarker relies heavily on the accuracy and precision of relative qRT-PCR quantification, which is still challenging for low abundance transcripts. Relative qRT-PCR quantification measures a target biomarker expression normalized to endogenous reference genes. Due to variable amplification efficiency among the reactions, reliable estimates of amplification efficiency are needed for accurate relative quantification. In this work, we consider new models for dynamics of qRT-PCR efficiency. The model provides an estimate of amplification efficiency, which is used for efficiency-adjusted relative quantification. Using various kinetic qRT-PCR data from serial dilution experiments, we demonstrate that new efficiency estimates improve the accuracy of relative qRT-PCR quantification. Further, the new relative quantification method is applied to compute the normalized expressions of GUCY2C mRNA in the blood of colorectal cancer patients.

email: Inna.Chervoneva@jefferson.edu

**Evaluation Drug Efficacy in the Presence of the Imperfect Companion Diagnostic Device****Meijuan Li\***, U.S. Food and Drug Administration

Statistical methodologies have been proposed for evaluating drug efficacy in the presence of a biomarker. One major limitation of the existing approaches is that the biomarker is unrealistically assumed to have perfect diagnostic accuracy i.e. a perfect 100-percent sensitivity and a perfect 100-percent specificity. Biomarker accuracy,

including sensitivity and specificity, relies on the chosen discriminatory cut-off on a continuous test scale and its analytical performance. This paper discusses statistical issues and challenges for evaluating the drug efficacy in the presence of imperfect biomarker.

email: meijuan.li@fda.hhs.gov

### **Joint Graphical Models for Relational Structures In Multi-Dimensional Phenotypic Data**

**Vivian H. Shih\***, Novartis Pharmaceuticals  
**Catherine A. Sugar**, University of California, Los Angeles

The explosion of phenotypic research, the study of dimensional patterns of deficits characterizing specific disorders, has yielded more data than conventional statistical tools can digest. Traditional dimension reduction techniques such as means, correlations, principal component analysis, and factor analysis collapse across phenotypes before analysis, possibly leading to loss of information. Alternatively, graphical models extract the underlying structure of the data and provide a holistic view of the interrelationships as well as specific hotspots within phenotypes through the use of sparse covariance estimation. Joint graphical models further allow for comparisons of these structures across multiple groups. We uncover key phenotypes for childhood ADHD and chronic tic disorder using the conventional algorithm and suggest modifications to reflect covariate adjustment and longitudinal patterns across time.

email: vivianhshih@gmail.com

### **Sample Size Methods for Training Classifiers Developed from Regularized Logistic Regression**

**Sandra Safo\***, University of Georgia  
**Xiao Song**, University of Georgia  
**Kevin K. Dobbin**, University of Georgia

Classification of new patients into one of two classes is a common objective in high dimensional studies. The classification method is developed on a training set and the validity is determined on a testing set or by resampling. Determining the number of samples required in the training set is important from the perspective of both cost and classification accuracy. Few sample size methods have been proposed and most of these methods either use parametric models or are designed for low dimensional studies. We present a new nonparametric training sample size method for regularized logistic regression. The model permits inclusion of clinical covariates with high dimensional data, an approach that has not been considered in the existing sample size methods, to better classify observations into one of two existing classes. The sample size is chosen so that the expected performance of the classifier is within a user-specified tolerance value of the optimum performance. We apply the sample size method to simulated data, microarray data, and next generation sequencing data. We also develop user friendly software to guide clinicians and practitioners in the use of the proposed method.

email: dobbinke@uga.edu

### **Bilaterally Contaminated Normal Model with Nuisance Parameter and its Applications**

**Qian Fan\***, University of Kentucky  
**Hongying Dai**, Children's Mercy Hospital  
**Richard J. Charnigo**, University of Kentucky

Dai and Charnigo (2008, 2010) have proposed various contaminated density models which can be applied to microarray data and studied the MLRT and D-test statistics for these models under an omnibus null hypothesis of no differential expression. Charnigo et al (2013) subsequently proposed the bilateral contaminated normal model (BCN) without nuisance parameter, which is able to describe both gene under- and over-expression simultaneously. They proved the testing procedure is consistent and the test statistic has a limiting normal distribution under the unilateral null hypothesis of differential expression in one direction only. This presentation extends the previous work and puts forward tests of contamination in the BCN model when the common within-component variance is unknown. This model is more flexible compared to the BCN model without nuisance parameter. Yet the derivation and justification of the testing procedures are challenging. The first step is to do an omnibus test of no contamination, for which we propose a union-intersection test. The next stage is testing unilateral contamination against bilateral contamination. Both test procedures are shown to be asymptotically unbiased and consistent. We investigate their empirical performances via simulations and application to real data.

email: qfa222@uky.edu

### **Correlation Coefficient Inference for Left-Censored Biomarker Data with Known Detection Limits**

**Courtney E. McCracken**, Emory University  
**Stephen W. Looney\***, Georgia Regents University

Researchers are often interested in the association between the concentrations of two different analytes, both of which may be biomarkers. Despite the continuing advances in biotechnology, the value of a particular analyte may fall below some known limit of detection (LOD) of the measuring device. Data values such as these are referred to as non-detects (NDs), which are usually treated as left-censored in statistical analyses. When attempting to measure the association between two left-censored variables, serious complications can arise. The most commonly used method consists of simply replacing each ND with some representative value such as the LOD or LOD/2. Spearman correlation, in which all NDs are all assumed to be tied at some value smaller than the LOD, has also been used. In this presentation, we compare the performance of several methods for measuring the association between two variables, both of which are subject to NDs. Using simulation, we consider several scenarios, including small to moderate sample size, moderate to large censoring proportions,

extreme imbalance in censoring, and non-bivariate normal data. A maximum likelihood approach based on the bivariate normality assumption has acceptable performance under most scenarios, even when the data do not follow a bivariate normal distribution. Spearman's rho also performs adequately in many situations.

email: slooney@gru.edu

### **A Semi-Parametric Model for Time-Dependent Predictive Accuracy Curves of Biomarkers**

**Weining Shen\***, University of Texas

MD Anderson Cancer Center

**Jing Ning**, University of Texas MD Anderson Cancer Center

**Ying Yuan**, University of Texas

MD Anderson Cancer Center

A major goal in biomedical studies is to develop a score-evaluation model for candidate biomarkers based on their outcomes throughout the entire study period. Time-dependent receiver operating characteristic (ROC) curves have been widely studied in this context. Existing methods require modeling true positive rates and true negative rates and specifying a model of the event time and the marker. In this paper, we propose a semi-parametric regression model that directly estimates the time-dependent area under the curve (AUC) using fractional spline approximations. We establish the asymptotic properties of the proposed estimator and obtain the confidence bands. Numerical results suggest that our method works reasonably well comparing to existing approaches.

email: wshen@mdanderson.org

## **22. MACHINE LEARNING**

### **Joint Estimation of Multiple Graphical Models from High Dimensional Dependent Data**

**Huitong Qiu\***, Johns Hopkins University

**Fang Han**, Johns Hopkins University

**Han Liu**, Princeton University

**Brian S. Caffo**, Johns Hopkins University

In this manuscript the problem of jointly estimating multiple graphical models in high dimensions is considered. It is assumed that the data are collected from  $n$  subjects, each of which consists of non-independent observations. The graphical models of subjects vary, but are assumed to change smoothly corresponding to a measure of the closeness between subjects. A new kernel based method for jointly estimating all graphical models is proposed. Theoretically, under a double asymptotic framework, where both  $(m,n)$  and the dimension  $d$  can increase, the explicit rate of convergence in parameter estimation is provided, thus characterizing the strength one can borrow across different individuals and impact of data dependence on parameter estimation. Empirically, experiments on both synthetic and real world resting state functional magnetic resonance imaging (rs-fMRI) data illustrate the effectiveness of the proposed method.

email: qht19881226@gmail.com

### **MBACT - Multiclass Bayesian Additive Classification Trees**

**Bereket P. Kindo\***, University of South Carolina

**Hao Wang**, University of South Carolina

**Edsel A. Pena**, University of South Carolina

We propose Multiclass Bayesian Additive Classification Trees (MBACT) as a nonparametric procedure to deal with multiclass classification problems. MBACT is a multiclass extension of BART: Bayesian Additive Regression Trees [Chipman et al., 2010]. In a range of data generating schemes and real data applications, MBACT is shown to have good predictive performance, competitive to existing procedures, and in particular it outperforms most procedures when the relationship between the response and predictors is nonlinear.

email: kindo@email.sc.edu

### **Random Forest Importance Scores: Significance Testing and Conditional Importance**

**Eric Bair\***, University of North Carolina, Chapel Hill

**Lira Pi**, University of North Carolina, Chapel Hill

Random forests are a data mining method that can be used to evaluate the association between a response variable and a large number of predictors. In particular, random forests can be used to calculate variable importance scores, which measure how much the predictive accuracy of the model is decreased when a given variable is measured with error. Although importance scores are a useful tool, they have certain shortcomings. There is no simple method to test the null hypothesis that an importance score is greater than 0, and predictors that are not associated with the outcome variable can have a high importance score if they are strongly correlated with another predictor. We derive an approximation for the distribution of importance scores under the null hypothesis of no association between a predictor variable and the outcome. This null distribution can be used to test the null hypothesis of no association between the predictor and the outcome. We also show how one may calculate this distribution after conditioning on the other predictor variables. The method is applied to several simulated data sets and used to identify the most important predictors of TMD in the OPPERA study.

email: ebair@email.unc.edu

### **Large-Margin Classifier Selection Via Decision Boundary Instability**

**Wei Sun\***, Purdue University

**Guang Cheng**, Purdue University

**Yufeng Liu**, University of North Carolina, Chapel Hill

Large-margin methods have been widely used for classification problems. However, the question of which classifier should be chosen for a given problem is far less studied. The existing criteria for classifier comparison are mainly based on their generalization errors (GEs) or excess risks, and do not take into account the prediction/classification (in)stability. In this paper, we attempt to address this important selection issue from a new perspective by introducing a new measure of classification instability: decision boundary instability (DBI). This new measure

universally applies to all types of linear classifiers. Specifically, we propose a two-stage algorithm for identifying the most {em accurate and stable} classifier: (i) we initially select a subset of classifiers whose GEs are not significantly different from the minimal GE among all the candidate classifiers; (ii) the optimal classifier is chosen as the one with the minimal DBI among the subset selected in stage (i). The large-margin unified machine (Liu et al., 2011) is used as a prototypical example to illustrate the above idea. Our selection method is shown to be consistent in the sense that the optimal classifier simultaneously achieves the minimal GE and the minimal DBI. The empirical performance is demonstrated using various simulated examples and real data sets.

email: sun244@purdue.edu

### **Bias Correction for Selecting the Minimal-Error Classifier from many Machine Learning Models**

**Ying Ding\***, University of Pittsburgh

**Shaowu Tang**, University of Pittsburgh

**Ge Liao**, University of Pittsburgh

**Jia Jia**, University of Pittsburgh

**Yan Lin**, University of Pittsburgh

**George C. Tseng**, University of Pittsburgh

Supervised machine learning is commonly encountered in genomic data analysis. The goal is to construct a classifier from training data that is generalizable to predict independent testing data. When an independent testing data set is not available, cross-validation is commonly used to evaluate an unbiased error rate estimate. It has been a common practice that many machine learning methods are applied to one data set and the method that produces the smallest cross-validation error rate is selected and reported. Theoretically such a minimal-error classifier selection produces bias with an optimistically smaller error rate, especially when the sample size is small and many classifiers are examined. In this paper, we illustrated this problem and explored the statistical and asymptotic properties. We compared two existing methods developed to correct this bias: nested cross validation and Tibshirani's procedure. We showed that nested cross validation had an upward bias and Tibshirani's procedure also produced significant and fluctuating biases. We proposed another correction method based on learning curve fitting by inverse power law and showed its advantage to extrapolate the estimates to a larger sample size. These three methods were applied to five small real datasets and one large TCGA breast cancer dataset.

email: dingying85@gmail.com

### **Ensemble Learning of Inverse Probability Weights for Marginal Structural Modeling in Large Observational Datasets**

**Susan Gruber\***, Harvard School of Public Health

**Roger W. Logan**, Harvard School of Public Health

**Inmaculada Jarrin**, Instituto de Salud Carlos III, Madrid, Spain

**Susana Monge**, Instituto de Salud Carlos III, Madrid, Spain

**Miguel Hernan**, Harvard School of Public Health

Inverse probability weights used to fit marginal structural models are typically estimated using logistic regression, however a data-adaptive procedure may be able to better

exploit information available in measured covariates. By combining predictions from multiple algorithms, ensemble learning offers an alternative that may further reduce bias in estimated marginal structural model parameters. Two ensemble learning approaches to estimating stabilized weights were used to fit a weighted marginal structural Cox model: 1) an ensemble learner (EL) that creates a single partition of the data into training and validation sets, and 2) super learning (SL), an ensemble approach that relies on V-fold cross-validation. Longitudinal data from two multicenter cohort studies (CoRIS and CoRIS-MD) were analyzed to estimate the average hazard ratio of mortality for initiating combined anti-retroviral therapy versus non-initiation among HIV positive subjects. Results were in agreement, and computation time for EL was less than half that of SL. Conclusion: Ensemble learning using a library of diverse algorithms is a worthwhile alternative to parametric modeling of inverse probability weights when fitting marginal structural models. With large datasets, EL provides a rich search over the solution space in less time than SL with comparable results.

email: sgruber@hsph.harvard.edu

### **Ordinal Logic Forest: Discovering Interactions Among Binary Predictors for Classifying Ordinal Responses**

**Bethany J. Wolf\***, Medical University of South Carolina

**Elizabeth G. Hill**, Medical University of South Carolina

**Elizabeth H. Slate**, Florida State University

Predicting patients' disease risk, disease severity, or response to treatment often necessitates modeling complex interactions among genetic and environmental factors. Identifying higher-order interactions such as those that might describe disease status can be difficult using traditional statistical methods. Ordinal Logic Regression (OLR), a nonparametric tree-based method based on logic regression, models an ordinal response as a series of Boolean combinations of binary predictors associated with the levels of the ordinal response. However, OLR is unstable when data are noisy or in instances where data fail to include an important predictor truly associated with the response. We implement an ensemble adaptation of OLR called Ordinal Logic Forest (OLF). Additionally, we develop a measure of importance for the Boolean combinations identified by an OLF model. We compare the ability of Ordinal Logic Regression and Ordinal Logic Forest to identify interactions predictive of each ordinal response level. Our findings indicate Ordinal Logic Forest is superior to Ordinal Logic Regression for identifying important predictors. We also apply our method determine association between genetic and health factors and severity of adult periodontitis in diabetic African Americans.

email: wolfb@musc.edu

## 23. MULTIPLE TESTING

### Sizing Clinical Trials that Compare Two Interventions Using Two Time-to-Event Outcomes

**Yuki Ando**, Pharmaceuticals and Medical Devices Agency  
**Toshimitsu Hamasaki\***, Osaka University Graduate School of Medicine and National Cerebral and Cardiovascular Center

**Tomoyuki Sugimoto**, Hirosaki University Graduate School of Science & Technology

**Scott R. Evans**, Harvard School of Public Health  
**Yuko Ohno**, Osaka University Graduate School of Medicine

The use of two primary time-to-event endpoints has become common in clinical trials evaluating interventions in many disease areas such as infectious disease, oncology, or cardiovascular disease. We previously developed methods for sizing clinical trials with two co-primary time-to-event outcomes when both events are non-terminal and are not censored by the other event. In this presentation, we describe methods that extend these results to accommodate two situations, i.e., (i) when one event is terminal, and (ii) when both are terminal. For (i), the non-terminal event-time (e.g., disease progression, MI, or stroke) may be censored by the terminal event (e.g., death), but for (ii), each event-time may be censored by other event (e.g., disease specific death and all cause death). We describe a sample size formula with the aim being to detect: (a) effects on all endpoints (referred as multiple co-primary endpoints), and (b) effects on at least one endpoint with a prespecified non-ordering of endpoints (referred as multiple primary endpoints). We evaluate the performance of the methods and investigate the behavior of the required sample sizes via simulation.

email: hamasakt@medstat.med.osaka-u.ac.jp

### Multiple Simultaneous Tests for Noninferiority and Superiority: A Graphical Approach

**Heng Li**, U.S. Food and Drug Administration  
**Vandana Mukhi\***, U.S. Food and Drug Administration

More and more often, clinical studies are designed to use active control to evaluate novel treatment regimens involving new medical products, and test hypotheses that the latter are non-inferior to the former in certain aspects. It is natural in such a design to anticipate that a superiority claim will also be of interest if non-inferiority is established. We discuss a graphical approach to controlling the family-wise type I error rate in this setting involving multiple simultaneous tests of non-inferiority and superiority.

email: vandana.mukhi@fda.hhs.gov

### Multiple Testing that Considers Assumptions and Network

**Demba Fofana\***, University of Memphis  
**E. O. George**, University of Memphis  
**Dale Bowman**, University of Memphis

Analyzing gene expression data rigorously requires taking assumptions into consideration but also relies on using information about network relations that exist among genes. Combining these different elements cannot only improve

statistical power, but also provide a better framework through which gene expression can be properly analyzed. We propose a novel statistical model that combines assumptions and gene network information into the analysis. Assumptions are important since every test statistic is valid only when required assumptions hold. We incorporate gene network information into the analysis because neighboring genes share biological functions. This correlation factor is taken into account via similar prior probabilities for neighboring genes. With a series of simulations our approach is compared with other approaches. Our method that combines assumptions and network information into the analysis is shown to be more powerful.

email: dfofana@yahoo.com

### Testing the Disjunction Hypothesis Using Voronoi Diagrams, with Applications to Genetics

**Daisy Phillips\***, The Pennsylvania State University  
**Debashis Ghosh**, The Pennsylvania State University

Testing of the disjunction hypothesis is appropriate when each gene or location studied is associated with multiple p-values, each of which is of individual interest. This situation can occur when more than one aspect of an underlying process is measured. For example, cancer researchers may hope to detect genes that are both differentially expressed on a transcriptomic level and show evidence of copy number aberration. Currently used methods of p-value combination for this setting are overly conservative, resulting in very low power for detection. In this work, we introduce a method to test the disjunction hypothesis by using cumulative areas from the Voronoi diagram of two-dimensional vectors of p-values. Our method offers much improved power over existing methods, even in challenging situations, while maintaining appropriate error control. We apply the approach to data from a published study of high-throughput gene expression and copy number data to identify genes associated with prostate cancer.

email: dlp245@psu.edu

### A Class of Improved Hybrid Hochberg-Hommel Type Step-Up Multiple Test Procedures

**Jiangtao Gou\***, Northwestern University  
**Ajit C. Tamhane**, Northwestern University  
**Dong Xi**, Novartis Pharmaceuticals Corporation  
**Dror Rom**, Prosoft Software, Inc.

In this paper we derive a new procedure which improves upon the Hommel (1988) procedure by gaining power as well as having a simpler step-up structure similar to the Hochberg (1988) procedure. Therefore, it is the recommended choice over the other commonly used p-value based stepwise multiple test procedures. The key to this improvement is that the Hommel procedure is a nonconsonant procedure and can be improved by a consonant procedure (Romano et al., 2011). Exact critical constants of this new procedure can be numerically determined and tabled. The 0th order approximations to the exact critical constants, albeit slightly conservative, are simple to use and need no tabling, and hence are

recommended in practice. The resulting procedure is shown to strongly control the familywise error rate (FWER) both under independence (analytically) and under positive dependence (via simulation) among the  $p$ -values, and is also shown to be more powerful (via simulation) than competing procedures. Illustrative examples are given.

email: jgou@u.northwestern.edu

### Identifying Multiple Regulation Across a Diverse Set of Outcomes

**Denis M. Agniel\***, Harvard University  
**Tianxi Cai**, Harvard University

In recent years, considerable interest has been focused on studying multiple phenotypes simultaneously in both epidemiological and genomic studies, either to capture the multidimensionality of complex disorders or to understand shared etiology of related disorders. We seek to identify 'multiple regulators' or predictors that are associated with multiple phenotypes, or more generally outcomes, when these outcomes may be measured on very different scales. We employ a two-stage technique to both estimate all effects and identify multiple regulation while controlling error rates. In the first stage, we use regularization to induce sparsity in the estimated effects, and in the second stage we use a resampling-based multiple testing procedure to identify multiple regulation while controlling the familywise error rate. Simulation results indicate that our estimation method can improve over unregularized methods, our resampling method estimates the variability in our estimator accurately, our testing method identifies multiple regulation at a higher rate than marginal testing methods, and using sparsity increases power to detect all non-null effects. We apply our method to a genetic study of autoantibodies.

email: dagniel@g.harvard.edu

### Dorfman Testing with Correlated Responses

**Elena K. Bordonali\***, University of North Carolina, Chapel Hill  
**Michael G. Hudgens**, University of North Carolina, Chapel Hill  
**Bahjat F. Qaqish**, University of North Carolina, Chapel Hill

The efficiency, or expected number of tests per unit, is derived for the Dorfman group testing algorithm when units within clusters exhibit first-order autoregressive or moving average pairwise correlation. An example is presented demonstrating that the efficiency can be substantially over- or underestimated when incorrectly assuming either an independent or exchangeable correlation structure. A general result is given showing that the efficiencies assuming independence, exchangeable, autoregressive and moving average correlation always follow a particular ordering.

email: ekb@unc.edu

## 24. METHODS FOR STATISTICAL GENETICS

### Fitting Generalized Linear Mixed Models to Family Data in Genetic Association Studies

**Tao Wang\***, Medical College of Wisconsin  
**Peng He**, Medical College of Wisconsin  
**Kwang Woo Ahn**, Medical College of Wisconsin  
**Xujing Wang**, University of Alabama, Birmingham  
**Soumitra Ghosh**, GlaxoSmithKline  
**Purushottam Laud**, Medical College of Wisconsin

In family-based genetic association studies, the generalized linear mixed model (GLMM) provides a useful tool for controlling or assessing the unobserved genetic effects. However, fitting this type of GLMM using existing statistical software has been a challenge due to varying family sizes and genetic covariance structures across families. In this study, we explore a Cholesky decomposition based method to re-formulate the GLMM into a standard regression model with random regression coefficients. This Cholesky decomposition based approach is flexible to handle varied family sizes with different degrees of relatedness among family members. It also allows us to fit the GLMM using existing statistical software conveniently. We provide detailed code on fitting this type of GLMM using either "proc nlmixed" or "proc glimmix" procedures in SAS, and OpenBUGS with R. Performance of these model fitting procedures is assessed through simulation studies. Application to a real data set is also included.

email: taowang@mcw.edu

### Kernel Methods for Regression Analysis of Microbiome Compositional Data

**Jun Chen\***, Harvard School of Public Health  
**Hongzhe Li**, University of Pennsylvania

The human microbiome can now be studied using next generation sequencing. Many human diseases have been shown to be associated with the disorder of the human microbiome. Previous statistical methods for associating the microbiome composition with an outcome such as disease status focus on the association of the abundance of individual taxon or their abundance ratios with the outcome variable. However, the problem of multiple testing leads to loss of power to detect the association. When individual taxon-level association test fails, an overall test, which pools the individually weak association signal, can be applied to test the significance of the effect of the overall microbiome composition on an outcome variable. In this paper, we propose a kernel-based semi-parametric regression method for testing the significance of the effect of the microbiome composition on a continuous or binary outcome. Our method provides the flexibility to incorporate the phylogenetic information into the kernels as well as the ability to naturally adjust for the covariate effects. We evaluate our methods using simulations as well as a real data set on testing the significance of the human gut microbiome composition on body mass index (BMI) while adjusting for total fat intake. Our result suggests that the gut microbiome has a strong effect on BMI and this effect is independent of total fat intake.

email: jchen1981@gmail.com

**Latent Class Quantitative Trait LOCI (QTL) Mapping****Shuyun Ye\***, University of Wisconsin, Madison**Xiaomao Li**, University of Wisconsin, Madison**Mark Keller**, University of Wisconsin, Madison**Alan Attie**, University of Wisconsin, Madison**Christina Kendzierski**, University of Wisconsin, Madison

Identifying the genetic basis of complex traits is an important problem with the potential to impact a broad range of biological endeavors. A number of good statistical methods are available for quantitative trait loci (QTL) mapping that allow for the efficient identification of multiple, potentially interacting, loci under a variety of experimental conditions. Although proven useful in hundreds of studies, the majority of these methods assume a single model common to each subject and consequently sacrifice power and accuracy when genetically distinct subgroups exist. To address this, we have developed an approach to enable latent class QTL mapping. The approach combines ideas from latent class regression and QTL mapping to estimate the number of subgroups in a population, and to identify the genetic model that best describes each subgroup. Simulations demonstrate that the method performs well under a variety of situations, and does not sacrifice power and efficiency when only a single group is present. An application of the method to a study of diabetes in mouse illustrates advantages of the approach in practice.

email: yeshuyun721@gmail.com

**Using Gene Expression to Improve the Power of Genome-Wide Association Analysis****Yen-Yi Ho\***, University of Minnesota**Emily C. Baechler**, University of Minnesota**Ward Ortmann**, Genentech, Inc.**Timothy W. Behrens**, Genentech, Inc.**Robert R. Graham**, Genentech, Inc.**Tushar R. Bhangale**, Genentech, Inc.**Wei Pan**, University of Minnesota

Motivation: Genome-wide association (GWA) studies have reported susceptible regions in the human genome for many common diseases and traits, however, these loci only explain a minority of trait heritability. To boost the power of a GWA study, substantial research endeavors have been focused on integrating other available genomic information in the analysis. Results: Advances in high through-put technologies have generated a wealth of genomic data, and made combining SNP and gene expression data become feasible. In this paper we propose a novel procedure to incorporate gene expression information into GWA analysis. The procedure utilizes weights constructed by gene expression measurements to adjust p values from a GWA analysis. Results from simulation analyses indicate that the proposed procedure achieves substantial power gains while controlling family-wise type I error rate (FWER) at the nominal level. We demonstrate the implementation of our proposed approach in a GWA analysis for serum interferon-regulated chemokine levels in systemic lupus erythematosus (SLE) patients. The study results can provide valuable insights for the functional interpretations of GWA signals.

email: yho@umn.edu

**Extending Linear Predictors to Impute Genotypes In Pedigrees****Wenan Chen\***, Mayo Clinic**Daniel J. Schaid**, Mayo Clinic

Recently, Wen & Stephens proposed using conditional multivariate normal moments to impute genotypes with accuracy similar to current state-of-the-art methods. One novelty is that it regularized the estimated covariance matrix based on a model from population genetics. We extended multivariate normal moments to impute genotypes in pedigrees. Our proposed method utilizes both the linkage disequilibrium (LD) information estimated from external panel data (such as CEU for European populations) and the pedigree information through the kinship matrix. We found that incorporating the pedigree information can improve imputation accuracy. Furthermore, because rare variants usually have low LD with other SNPs, using kinship information gave the greatest imputation improvement for rare variants. We also found that when the observed identity by descent (IBD) of SNPs is available, using the observed IBD instead of the expected kinship can further improve the accuracy. In this presentation, we will give an intuitive explanation of the theoretical background, and present numerical results on imputation accuracy for both simulated pedigree data and real pedigree data from a prostate cancer study.

email: Chen.Wenan@mayo.edu

**Inferring Rare Disease Risk Variants Based on Exact Probabilities of Sharing by Multiple Affected Relatives****Alexandre Bureau\***, Institut universitaire en santé mentale de Québec**Samuel Younkin**, University of Wisconsin, Madison**Margaret M. Parker**, Johns Hopkins

Bloomberg School of Public Health

**Joan E. Bailey-Wilson**, National Human Genome

Research Institute, National Institutes of Health

**Mary L. Marazita**, University of Pittsburgh**Jeffrey C. Murray**, University of Iowa**Elisabeth Mangold**, University of Bonn**Hasan Albacha-Hejazi**, Dr. Hejazi Clinic**Terri H. Beaty**, Johns Hopkins

Bloomberg School of Public Health

**Ingo Ruczinski**, Johns Hopkins

Bloomberg School of Public Health

Family based designs are regaining popularity for genomic sequencing studies because they provide a way to test co-segregation with disease of variants too rare in the population to be tested individually in a conventional case-control study. Where only a few affected subjects per family are sequenced, the probability any variant would be shared by all affected relatives given it occurred in any one family member provides evidence against the null hypothesis of a complete absence of linkage and association. A p-value can be obtained as the sum of the probabilities of sharing

events as (or more) extreme in one or more families. Highly penetrant variants observed in ten families are detected with good power. We generalized an existing closed-form expression for exact sharing probabilities to more than two relatives per family. We also examined the impact of unknown relationships and proposed an approximation of sharing probabilities based on empirical estimates of kinship among founders obtained from genome-wide marker data, approximation shown to be accurate for low levels of kinship. We applied this method to a study of 55 multiplex families with apparent non-syndromic forms of oral clefts, with whole exome sequences available for two or three affected members per family.

email: alexandre.bureau@msp.ulaval.ca

### **People Can't See Statistical Significance: A Massive Randomized Trial on the Visual Perception of Relationships**

**Aaron Fisher\***, Johns Hopkins

Bloomberg School of Public Health

**Georgiana B. Anderson**, Johns Hopkins

Bloomberg School of Public Health

**Jeff Leek**, Johns Hopkins

Bloomberg School of Public Health

Visually inspecting a bivariate scatter-plot is a central technique of exploratory data analysis (EDA). However, it is not actually known how accurate humans are when estimating the significance of relationships after observing scatter-plots. To address this question we conducted an online survey of 2039 statistically literate Coursera students. Each student was shown a random set of scatter-plots and asked to guess if the underlying relationships had significant p-values. Users were only able to correctly classify 47.4% (95% CI: 45.1%-49.7%) of truly significant relationships, and 74.5% (95% CI: 72.5%-76.6%) of non-significant relationships. Adding the visual aid of a best-fit line increased the probability that a user would guess that a relationship was significant, regardless of whether the relationship was actually significant. Classification of truly significant relationships improved after attempting the survey more than once, although classification of non-significant relationships did not. Our results imply a disconnect between the rigorous definition of statistical significance, and the intuitive standards many researchers have for "strong evidence" of a relationship.

email: afisher@jhsph.edu

## **25. STATISTICAL INNOVATIONS FOR STUDYING THE HUMAN BRAIN FUNCTION**

### **A New Method for Estimating Changes in Granger Causality in Eeg Data**

**Ivor Cribben\***, University of Alberta

Recently in brain imaging there has been an increased interest in quantifying changes in connectivity between brain regions over the experimental time course to provide a deeper insight into the fundamental properties of brain

networks. The application of network science and graphical modelling has been instrumental in these analyses and enabled the examination of the brain as an integrated system. In this work, we propose a new method that detects changes in Granger Causality (GC) between brain regions over time for single-subject and multi-subject data where the number of changes is unknown a priori. This new method makes use of graphical models to detect the changes in GC, to easily visualize the relationships between brain regions and to generate biological hypotheses. In brain imaging studies, variability between subjects makes the use of multi-subject data sets challenging. Hence, in the multi-subject case, the new method jointly estimates multiple graphical models by borrowing strength across the subjects and by exploiting the similarity between subjects' brain networks. We evaluate the effectiveness of our proposed method on simulated data sets as well as on an Electroencephalography (EEG) data set collected while the subjects watched several hours of movie clips.

email: cribben@ualberta.ca

### **Genome-Wide Scan of Brain Phenotypes Discovers Common Genetic Variants Influencing Cortical Surface Area**

**Chi-Hua Chen\***, University of California, San Diego

**Andrew Schork**, University of California, San Diego

**Wesley K. Thompson**, University of California, San Diego

**Ole Andreassen**, University of Oslo

**Anders Dale**, University of California, San Diego

Although it is well-recognized that human brain structures are heritable, little is known about genetic underpinnings of these traits. With development of neuroimaging and genotype-wide sequencing techniques, we have an unprecedented opportunity to study common genetic variants associating with brain phenotypes. It has been shown that the stringent Bonferroni derived cut-off used in genome-wide association studies (GWAS) is insufficient to recover the vast majority of true genetic signals with small effects for most complex traits. This problem increases with high dimensional neuroimaging data where correction for multiple testing is an even more daunting challenge. We used data driven clustering techniques to identify parcels of the human cortex that are maximally genetically correlated based on twin imaging data (N = 406). We then extracted average measures of cortical surface area in these parcels in our multi-study cohort (N = 1,803) to reduce the dimension in imaging data, and then performed a GWAS to identify common polymorphisms that may contribute to the area size of these parcels. Although our sample is modest by GWAS standards, we have found some significant hits. Further, we employed a new analytical paradigm such as empirical Bayesian mixture modeling of Efron 2010 to enhance gene discovery.

email: chc101@ucsd.edu

### Comparison of Parametric and Semiparametric Statistical Methods and Signal Processing Methodology for fMRI Signal Analysis Illustrated Using a Gustatory Experiment

**Jaroslav Harezlak\***, Indiana University Fairbanks

School of Public Health

**Mario Dzemidzic**, Indiana University School of Medicine

**Maria A. Kudela**, Indiana University Fairbanks

School of Public Health

**Jacek Urbanek**, Indiana University Fairbanks

School of Public Health

**Brandon G. Oberlin**, Indiana University

School of Medicine

**David A. Kareken**, Indiana University School of Medicine

The parametric regression framework to analyze the blood oxygen level dependent (BOLD) responses in fMRI studies frequently employs a canonical hemodynamic response function (HRF). Some stimuli, however, do not evoke the canonical HRF. To deal with this difficulty, recently proposed statistical methods estimate HRF in either nonparametric or semiparametric way in the first stage and use parametric statistical methods to model voxel activation in the second stage. Yet another approach utilizes signal processing methodology to transform the observed fMRI time series using frequency-based methods and summarizes the strength of the signal in the frequency domain. In our research, we contrast and compare the abovementioned methods and apply them in the gustatory stimuli brain activation study. Compared to the classical canonical HRF analysis, we find that both semiparametric and signal processing methods give us larger sensitivity in detecting differentially activated clusters which contain more voxels in the left caudate and posterior cingulate areas.

email: harezlak@iupui.edu

### A Semi-Parametric Quadratic Inference Approach for Longitudinal fMRI Data

**Yu Chen\***, University of Michigan

**Timothy D. Johnson**, University of Michigan

**Min Zhang**, University of Michigan

We present a semi parametric quadratic inference approach to find spatially smoothed activated regions in longitudinal fMRI data analysis. Due to the complexity of high dimensional correlations (both spatial and temporal) in data, very few statistical methods are well developed to address problems in longitudinal fMRI data analysis. We propose to model the data using both a semi-parametric generalized estimation equation (GEE) and a quadratic inference function (QIF) approach to study the temporal trend of fMRI measurements for each brain voxel. Due to its greater efficiency, we are able to discover more activation regions under QIF, especially under misspecification of the working correlation structure, than under GEE. In addition to investigating the temporal trend, we also account for spatial correlation across the brain by extending the voxel-wise QIF method to a locally adaptive kernel quadratic inference function. This approach provides a set of spatially

smoothed estimators for each voxel, increasing the efficiency by borrowing strength across neighboring sites. We apply our method to a longitudinal fMRI of adolescents at risk for substance abuse. We will also investigate the relationships between activated brain regions and several covariates collected including IQ, age, gender, behavioral and personality variables.

email: cheyu@umich.edu

## 26. META-ANALYSIS OF GENE-ENVIRONMENT INTERACTION IN THE POST-GWAS ERA

### Testing G x E in Genome-Wide Association Studies

**Li Hsu\***, Fred Hutchinson Cancer Research Center

Identifying gene and environment interaction (G x E) can provide insights into biological mechanisms of complex diseases, identify novel genes that act synergistically with environmental factors, and inform risk prediction. Notable examples of GxE from genome-wide associations studies (GWAS) are smoking and GSTM1 and NAT2 in bladder cancer and alcohol and ADH1B and ALDH2 in esophageal squamous-cell carcinoma. However, it remains challenging to identify GxE genome-wide, mainly due to measurement error in the exposure assessment, heterogeneity across studies, and inadequate power. Among these, power is a critical issue, as even in a perfect situation with no measurement error or heterogeneity, the detection of an interaction needs at least 4 times as many subjects to detect a main effect of comparable effect size. Hence, large consortia are formed to increase sample size. In this talk, I will present challenges that arise from consortia including data harmonization and intensive computing. I will also present statistical approaches that harness various aspects of GxE testing ranging from single SNP to genome-wide and from testing one SNP at a time to test a set of SNPs.

email: lih@fhcrc.org

### Bayesian Meta-Analysis Methods for Detecting G-E Interactions in Genomic Data

**Xiaoquan Wen\***, University of Michigan

The interactions between genetic variants and cellular environments is a critical component of gene regulatory processes. Motivated by meta-analysis methods, we develop a set of flexible Bayesian analysis tools to detect such interactions by analyzing high-throughput genomic data. In particular, we formulate a model comparison problem and derive Bayes factors to identify both consistent and inconsistent genetic effects across different cellular environments. We demonstrate our methods by mapping expression Quantitative Trait Loci (eQTLs) across multiple cell types using data from NIH ENCODE and GTEx projects.

email: xwen@umich.edu

## **The Role of Covariate Heterogeneity in Meta-Analysis of Gene-Environment Interactions with Quantitative Traits**

**Bhramar Mukherjee\***, University of Michigan  
**Shi Li**, Eli Lilly and Company

With heterogeneity in environmental covariate distributions across cohorts and obvious challenges with data harmonization involving various data sources, meta-analysis of studies of gene-environment interaction can often involve subtle statistical issues. In this paper, we study the effect of environmental covariate heterogeneity (within and between cohorts) on two approaches for fixed-effects meta-analysis: the standard inverse-variance weighted meta-analysis and a meta-regression approach. Akin to the results obtained in Simmonds and Higgins (2007), we obtain analytical efficiency expressions for both methods under the assumption of gene-environment independence. The relative efficiency of the two methods depend on the ratio of within versus between cohort variance of the environmental covariate. We propose to use an adaptively weighted estimator (AWE), as a combination of meta-analysis and meta-regression estimators, that can be used as a default choice, retaining full efficiency of the gold standard joint analysis for the interaction parameter using individual level data under certain natural assumptions. Lin and Zeng (2010) showed that a multivariate inverse-variance weighted estimator is asymptotically equivalent to the estimator using individual level data, given that all the common parameters with full covariance matrix are pooled across all studies. We show connection of our work to Lin and Zeng (2010). The AWE improves efficiency by combining meta-analysis and meta-regression based on only univariate summary statistics from each study, and bypasses issues with sharing of individual level data or multivariate information matrices across studies without sacrificing efficiency. We compared the performance of the methods under a variety of scenarios through a comprehensive simulation study. The methods were illustrated through meta-analysis of interaction between Single Nucleotide Polymorphisms in the FTO gene and body mass index on high-density lipoprotein cholesterol data from a set of eight studies of type 2 diabetes.

email: bhramar@umich.edu

## **Meta and Mega Analysis of G x E Interactions with Complex Disease Outcomes: Experience and Insights from the Charge Consortium**

**Kenneth Rice\***, University of Washington  
**Colleen Sitlani**, University of Washington

Gene-by-environment interactions invariably rely on smaller numbers of samples than corresponding main effects analyses. As a consequence, asymptotically-justified results that may work well for main effects analyses can perform badly when extended to interactions. In this talk we describe some methods that can improve on the asymptotic properties, particularly for interaction analyses and meta-analyses. These methods will be illustrated using high-throughput pharmacogenetic analyses from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium.

email: kenrice@uw.edu

## **27. STATISTICS METHODS FOR HIGH-THROUGHPUT GENOMICS**

### **Statistical Issues with RNAseq Data**

**Rafael Irizarry\***, Dana-Farber Cancer Institute and Harvard School of Public Health

RNA-sequencing technology is on pace to replace microarrays as the most common approach to measure gene expression. This transition is being driven by the rapid decline in cost of sequencing and the flexibility of sequencing technology to measure genomic events such as transcription of novel regions, allele specific transcription, and alternative transcription. In this talk, I will give a brief introduction to gene expression and related measurement technologies, describe some of the statistical issues with a focus on assessing and dealing with unwanted variability, and give a short summary of our ongoing work related to detecting differentially expressed regions of the genome. I will show some examples from a clinical datasets.

email: rafa@jimmy.harvard.edu

### **Model-Based Estimation of Abundances of Species, Microbial Genes and Pathways in Metagenomic Data**

**Hongzhe Li\***, University of Pennsylvania  
**Eric Chen**, University of Pennsylvania

Metagenomics is the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms and by sequencing the mixture of the genomes of microorganisms. The first step of statistical analysis of these short reads is to quantify the relative abundances of species, genes and pathways coded by these microbes. We present model-based approaches to address this problem where all samples and all species are analyzed simultaneously. Such multi-sample approaches can account for systematic biases and non-uniformity in read counts observed in the data and provide more precise estimates of the relative abundances at the species-, gene- and pathway-level. The models also provide a way of normalizing the data count data across samples. To account for the sparse nature of the high-dimensional compositional data, we propose regularization estimation methods to estimate the model parameters. We illustrate these methods using an on-going longitudinal metagenomic study of pediatric crohn disease patients at the University of Pennsylvania.

email: hongzhe@upenn.edu

### **Statistical Analysis of Time Course ChIP-seq Data**

**Xuekui Zhang**, Johns Hopkins  
Bloomberg School of Public Health  
**Hongkai Ji\***, Johns Hopkins  
Bloomberg School of Public Health

Transcription factor binding sites (TFBS) and chromatin modifications provide key information for understanding gene regulation. Coupling chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) allows one to map TFBS and chromatin states genome-wide. The first

generation ChIP-seq experiments are mostly designed to survey TFBS or chromatin states in one biological condition. More recently, time course ChIP-seq experiments are increasingly used to study gene regulation dynamics in cell cycles, circadian rhythms, and developmental processes. In this talk, we will discuss numerous new challenges that arise in the analyses of time course ChIP-seq data. We will introduce a statistical and computational framework to model the spatial and temporal structures of the data, and an algorithm for discovering and characterizing the temporal dynamics. The method will be illustrated using data from both yeast and mammals. We will demonstrate how this approach can be applied to study the dynamic interplay between transcription and chromatin modifications.

email: hjj@jhsph.edu

### **Sequencing Thousands of Human Genomes** Goncalo R. Abecasis, University of Michigan School of Public Health

Identifying and characterizing the genetic variants that affect human traits is one of the central objectives of human genetics. Ultimately, this aim will be achieved by examining the relationship between interesting traits and the complete genome sequences of many individuals. Whole genome re-sequencing of thousands of individuals remains challenging, but advances in laboratory methods and in statistical methodology have resulted in substantial progress in our understanding of complex disease biology. Here, we discuss results of the first generation of large scale sequencing studies. I illustrate findings from analysis of sequence data in 1,000s of individuals. Along the way, I will highlight analytical challenges and opportunities posed by next-generation sequence data - ranging from methods for the analysis of low-coverage data, to strategies for estimating individual ancestry using sequence data, to methods for combining the results of sequencing studies across samples.

email: goncalo@umich.edu

## **28. PANEL DISCUSSION: PERSONALIZED MEDICINE: BETTER TREATMENT FOR THE PATIENT OR THE RIGHT PATIENT FOR THE TREATMENT?**

### **Personalized Medicine: Better Treatment for the Patient or the Right Patient for the Treatment?** Anastasios A. Tsiatis, North Carolina State University

We speak a lot about personalized medicine, but what is the meaning of this term? Even though the ultimate goal is to have healthier society, the individual goals are different, and, therefore, statistical designs applied to address these goals in studies are different. Pharmaceutical companies embrace and propagate the idea of personalized medicine through biomarker discovery designs and enrichment strategies that help to build a clinical development program providing substantial evidence for the risk benefit of the treatment/ drug in a specific subpopulation of patients. Sequential, multiple assignment, randomized trial (SMART) designs are being now increasingly used by investigators and academic centers to address the individual treatment selection by

using Adaptive Treatment Strategies, individually tailored sequences of treatments, with treatment type and dosage adapted to the patient. The SMART designs were originally introduced by to address the treatment options in cancer are being used now to address a broad array of problems not only in oncology but in behavioral medicine.

email: tsiatis@ncsu.edu

### **Personalized Medicine: Better Treatment for the Patient or the Right Patient for the Treatment?** Keaven M. Anderson, Merck & Company, Inc.

Personalized medicine in the context of drug development today generally involves characterizing the biology of each patient's disease using a diagnostic test. Neither pharmaceutical companies nor regulators are used to approving new treatments using this paradigm. This talk will address some of the complications that are created and how actual practice seems to be evolving in this area. Experience based both on imaging and tissue diagnostics will be discussed. Important considerations include 1) when diagnostics should reach different stages of readiness, 2) ensuring that results generalize to the intended population, and 3) tradeoffs in delaying diagnostic approval until after an overall approval of the therapeutic entity. Note: this is intended as a joint submission for a session proposed by Olga Marchenko.

email: keaven\_anderson@merck.com

### **Overview of Study Designs for Personalized Medicine Approach** Sandeep M. Menon, Pfizer Inc. and Boston University

Personalized medicine is a relatively young but rapidly evolving field of clinical research. It involves identifying genetic, genomic, and clinical characteristics that have the potential to accurately predict patient's susceptibility of developing a certain disease and its response to treatment. Personalized medicine is the translation of this knowledge to patient care. However, this "translation" can be very challenging in the phase of limited knowledge of the biomarker and /or appropriate diagnostics. Hence, the appropriate selection of the study design is important to critically determine biomarker performance, reliability and eventually regulatory acceptance. This presentation will discuss various designs including adaptive designs available at our disposal and its merits and limitations.

email: smenon@bu.edu

### **Personalized Medicine: Better Treatment for the Patient or the Right Patient for the Treatment?** Ilya Lipkovich, Quintiles

Pharmaceutical companies embrace and propagate the idea of personalized medicine through biomarker discovery designs and enrichment strategies that help to build a clinical development program providing substantial evidence for the risk benefit of the treatment/ drug in a specific subpopulation of patients. Sequential, multiple

assignment, randomized trial (SMART) designs are being now increasingly used by investigators and academic centers to address the individual treatment selection by using Adaptive Treatment Strategies, individually tailored sequences of treatments, with treatment type and dosage adapted to the patient. The SMART designs were originally introduced by to address the treatment options in cancer are being used now to address a broad array of problems not only in oncology but in behavioral medicine. In this session, two experts, one from an academic center and one from a pharmaceutical company, will make a presentation on designs used in personalized medicine. The presentation will be followed by a panel discussion from academia, pharmaceutical industry, and regulatory agency experts on the topic.

email: [ilya.lipkovich@gmail.com](mailto:ilya.lipkovich@gmail.com)

## 29. RECENT ADVANCES IN STATISTICAL METHODS FOR META-ANALYSIS

### Bayesian Network Meta-Analysis for Categorical Outcomes

**Christopher H. Schmid\***, Brown University  
**Thomas A. Trikalinos**, Brown University

We develop a Bayesian multinomial network meta-analysis model for ordered and unordered categorical outcomes that allows for partially observed data in which exact event counts may not be known for each category. This model properly accounts for correlations of counts in mutually exclusive categories and enables proper comparison and ranking of treatment effects across multiple treatments and multiple outcome categories. We apply the model to analyze data from sets of clinical trials with categorical outcomes.

email: [christopher\\_schmid@brown.edu](mailto:christopher_schmid@brown.edu)

### Incorporation of Mixed Bivariate Outcomes and Individual Patient Data in Network Meta Analysis

**Bradley P. Carlin\***, University of Minnesota  
**Hwanhee Hong**, University of Minnesota  
**Haoda Fu**, Eli Lilly and Company  
**Karen L. Price**, Eli Lilly and Company

Bayesian methods have been shown to be useful in network meta-analysis (NMA), as they facilitate borrowing of strength across treatments, trials, and outcomes (say, safety and efficacy) as well as provide a natural framework for filling in missing data values that respect the underlying correlation structure in the data. In this talk, we first extend the aggregate data framework to mixed bivariate outcomes (say, a continuous efficacy measure and a binary safety measure), indicating how such models are readily coded and analyzed in BUGS, the most popular Bayesian software package. We then move on to models that incorporate individual patient data, perhaps available only on a subset of the trials in the NMA. Both contrast- and arm-based models will be considered, and both in the presence of covariate adjustment. We will also explore the use of such approaches in concert with Bayesian adaptive trial designs. We close by

speculating on further broadening of the range of external information that may one day be incorporated into NMAs, including expert opinion, user-reported observations (say, from handheld devices), and other unstructured 'big data' historically thought of as unsuitable for inclusion in rigorous scientific investigations.

email: [brad@biostat.umn.edu](mailto:brad@biostat.umn.edu)

### Meta-Analysis of Diagnostic Test Accuracy Comparisons: Network Methods

**Wei Cheng**, Brown University  
**Constantine Gatsonis\***, Brown University  
**Christopher Schmid**, Brown University  
**Thomas Trikalinos**, Brown University

The methodology and practice in research synthesis of diagnostic accuracy studies have grown substantially during the past decade. In particular, comparisons of diagnostic test accuracy provide important information for clinical and health policy decision making. However comparative accuracy assessments are typically made on the basis of indirect evidence. In this presentation we will discuss recent progress in the development of network methods for comparative meta-analysis of diagnostic test studies. We will examine the technical and conceptual challenges in this area of meta-analysis and describe approaches that take into account correlation in test results conducted on the same subjects in a network of studies involving several tests.

email: [gatsonis@stat.brown.edu](mailto:gatsonis@stat.brown.edu)

## 30. SUBGROUP ANALYSIS AND PERSONALIZED PREDICTION

### Personalized Prediction

**Yunzhang Zhu**, University of Minnesota  
**Xiaotong Shen\***, University of Minnesota  
**Changqing Ye**, University of Minnesota

Consider personalized prediction in classification and regression, where a person's preference over a large number of items is predicted. In this situation, partial latent models are used to integrate additional user-specific and content-specific predictors, for higher predictive accuracy. In particular, we factorize a user-over-item preference matrix into a product of two matrices, each representing a user's preference and an item preference by users. Then we propose a likelihood method to seek a sparsest latent factorization, from a class of overcomplete factorizations, possibly with a high percentage of missing values. This promotes additional sparsity beyond rank reduction. Computationally, we design methods based on a "decomposition and combination" strategy, to break large-scale optimization into many small subproblems to solve in a recursive and parallel manner. Several numerical examples will be given, in addition to some theoretical results.

email: [xshen@stat.umn.edu](mailto:xshen@stat.umn.edu)

**Personalized Treatment for Longitudinal Data**

**Hyunkeun Cho\***, Western Michigan University  
**Peng Wang**, Bowling Green State University  
**Annie Qu**, University of Illinois, Urbana-Champaign

We develop new modeling and estimation for personalized treatment for individuals with high heterogeneity. Incorporating subject-specific information into treatment subgroup is critical since individuals could react to the same treatment quite differently. We propose to identify subgroups with longitudinal observations through random-effects estimation where the random effects are not necessarily normal distributed. The advantage of this approach is that we can quantify intrinsic associations between unobserved subject-specific effects and observed treatment outcomes, and therefore provide optimal treatment assignments for different individuals. In contrast, traditional mixed-effects models assuming normal distribution cannot effectively distinguish different patterns of treatment effects. We develop asymptotic consistency theory for individual treatment effect estimation, and show that the new estimator is more efficient than the random effect estimator which ignores correlation information from longitudinal data. Simulation studies and a data example from an AIDS clinical trial group confirm that the proposed method is quite efficient in identifying an effective treatment strategy for subgroups in finite samples.

e-mail: hyunkeun.cho@wmich.edu

**Multiway Clustering with Hidden Structure**

**Bruce G. Lindsay\***, The Pennsylvania State University  
**Francesco Bartolucci**, University of Perugia  
**Francesca Chiaromonte**, The Pennsylvania State University

We consider arrays of data with two or more dimensions, starting with two dimensions. For each row there is a hidden state variable. The same is true for each column. For example, we might have a hidden state variable for rows that is randomly sampled for each row, describing different row types, while for columns we might suppose the hidden states are following a Markov chain, representing the transition between states. The resulting model might be called "mixture x hidden Markov chain." One output of an analysis would be clustering of the rows by most likely state while simultaneously segmenting the columns into intervals of constant state. Such models create unmanageable likelihoods. We consider the statistical issues that arise when one uses composite likelihood or likelihood simulation to create estimators.

e-mail: bgl@psu.edu

**Model-Based Inference in Subgroup Analysis**

**Xuming He\***, University of Michigan  
**Juan Shen**, University of Michigan

The need for subgroup identification arises in clinical trials and in market segmentation analysis. Subgroups can always be found, but can we distinguish them from data artifacts? In this talk we introduce a structured logistic-normal mixture model, which enables us to perform a confirmatory statistical test for the existence of subgroups, and at the

same time, to construct predictive scores for the subgroup membership. The inferential procedure proposed in the paper is built on the recent literature on hypothesis testing for Gaussian mixtures, but the structured logistic-normal mixture model enjoys some distinctive properties that are unavailable to the simpler Gaussian mixture models. With the bootstrap approximations, the proposed tests are shown to be powerful, and equally importantly, insensitive to the choice of tuning parameters.

e-mail: xmhe@umich.edu

**31. LATENT VARIABLE MODELING FOR MULTIPLE OUTCOMES AND GROWTH MODELS IN PSYCHIATRIC STUDIES****Shared Versus Specific Effects of Treatment on Multiple Outcomes in Clinical Trials Using Latent Variable Modeling**

**Melanie Wall\***, Columbia University

It is common in clinical trials to have multiple correlated outcome variables, yet typically a single primary outcome is chosen and traditional statistical analysis of treatment is performed. Further analyses taking account of the multiple variables may include analyzing each outcome separately or performing MANOVA. However, in some types of studies particularly in psychiatry it may be reasonable to consider the multiple outcome variables as indicators of a shared latent variable construct which itself is being affected by the treatment intervention. In such settings it may even be argued that the latent construct itself is the primary outcome of interest, rather than any one of the specific outcome variables measuring it. The current talk addresses this setting and describes a method to quantify the treatment effect on the latent variable and additionally identify how much of the treatment effect on the multiple observed variables is attributable to shared effects through the latent variable versus specific effects for each variable. We will demonstrate the method on a set of several anxiety scales measured in a clinical trial of a new cognitive-behavioral therapy (CBT) vs control designed to treat alcohol dependence patients with a broad spectrum of internalizing psychopathology.

email: mmwall@columbia.edu

**Using Multiple Imputation to Harmonize Data Across Multiple Trials that Use Different Outcome Measures**

**Juned Siddique\***, Northwestern University  
**Ahnalee Brinks**, University of Miami  
**Charles H. Brown**, Northwestern University  
**Jerome P. Reiter**, Duke University

There are many advantages to individual participant data meta-analysis for combining data from multiple studies. These advantages include greater power to detect effects, increased sample heterogeneity, and the ability to perform more sophisticated analyses. However, a fundamental challenge is that is unlikely that all the studies to be combined use the same measure for the construct of

interest. We propose that this situation can be viewed as a missing data problem and use multiple imputation to fill in missing measurements. We apply our method to 5 longitudinal adolescent depression trials where 4 studies used the Childrens Depression Rating Scale and the fifth study used the Hamilton Depression Rating Scale. None of the 5 studies used both depression measures. We make use of external information in order to produce more accurate imputations.

email: siddique@northwestern.edu

### **Simultaneous Estimation of Mixture Model for Multilevel Data**

**Haiqun Lin\***, Yale University

**Shu-xia Li**, Yale University

**Xiao Xu**, Yale University

**Harlan M. Krumholz**, Yale University

We develop a multilevel mixture model to identify distinct hospital patterns of practice change over time. We propose a simultaneous modeling framework using latent class model with each class representing a unique hospital practice pattern over time. Because patient composition varies over time even within the same hospital, growth mixture models cannot be directly fitted to such multilevel data. The currently available analytic approach is a two-stage approach where the patient-level data is aggregated within each hospital at a given time in the first stage and the growth mixture model is fitted to the aggregated hospital-level data in the second stage. This two stage approach does not account for variance in the aggregated measure. In this paper, we combine the two respective stages into one simultaneous model specification and estimation. The major advantage of our approach is to directly incorporate patient-level data in the evaluation of hospital performance over time and hence appropriately handle the differential patient variabilities. Our simultaneous model vs. the two-stage approach are evaluated in their interpretation and performance with data analysis of 30-day mortality rates following acute myocardial infarction during 2005-2010 in CMS and with simulation studies.

email: haiqun.lin@yale.edu

### **Three Novel Applications of Latent Variable Modeling: A Discussion**

**Samprit Banerjee\***, Cornell University

In this discussion, I will present an overview of the three methods to be presented by Dr. Wall, Dr. Siddique and Dr. Lin. I hope to highlight the strengths and weaknesses of the approaches and suggest future research directions. I would also emphasize the potential of latent variable modeling in a wide range of applications in psychiatry through a few illustrative examples.

email: sab2028@med.cornell.edu

## **32. BAYESIAN ANALYSIS OF HIGH DIMENSIONAL DATA**

### **Constrained Priors and X-Inactivation**

**Alan B. Lenarcic\***, University of North Carolina, Chapel Hill

**John Calaway**, University of North Carolina, Chapel Hill

**Fernando Pardo**, University of North Carolina, Chapel Hill

**William Valdar**, University of North Carolina, Chapel Hill

In every female mammalian cell, one X-chromosome is deactivated through a random selection process occurring in early development. Hybrids of inbred mouse strains can favor X-inactivation toward certain parental strains, leading to the over-representation of one parental X. A multi-allelic X-located locus Xce (X-Controlling Element) has been posited to tip the balance, believed to be within a 1.9 Mb region on the X. To better locate Xce, a study of F1 hybrids measured skew in crosses of previously uncategorized strains. Measurement of allele specific expression in multiple tissues was conducted with pyrosequencing and RNASeq. We posit a beta-distribution hierarchical model to convert pyrosequencing data to an aggregate whole-body statistic, using Bayesian shrinkage to estimate and compute credibility for differences in mean X-inactivation among multiple crosses. Typical iid priors would result in an unidentifiable model. Thus we use this experiment to use manifold-restricted priors, which we accommodate through reparametrized slice-sampling. Modeling X-inactivation using i.i.d. priors or a "weight-biased coin" regression model gives comparable predictive accuracy, supporting that Xce alleles differ through copy number variation (CNV), and that laboratory strains descended from similar sub-species carry the same allele.

email: alenarc@med.unc.edu

### **Bayesian Approach for Predicting Protein Secondary Structure**

**David B. Dahl**, Brigham Young University

**Qiwei Li\***, Rice University

**Marina Vannucci**, Rice University

**Hyun Joo**, University of the Pacific

**Jerry W. Tsai**, University of the Pacific

Determining the primary structure (i.e., amino acid sequence) of a protein is cheaper, faster, and more accurate than determining its secondary structure. The secondary structure, however, provides much more insight in the function of the protein. Therefore, a number of computational prediction methods have been developed, most of which are based on similarity to amino acid sequences with known secondary structure. We propose a Bayesian model to translate the given primary structure sequence to an unknown linear sequence of helix, strand, turn, and coil blocks, considering the packing influence of residues on secondary structure determination. Both sampling model and prior are constructed using the Protein Data Bank (PDB). The results are assessed using cross-validation. We find the predictive accuracy is comparable to other popular methods such as PROFphd on average. Our modeling allows insights into the rules governing packing, filling a substantial gap in the current understanding of protein structure.

email: ql6@rice.edu

## A Hierarchical Bayesian Model for Inference of Copy Number Variants and their Association to Gene Expression

**Alberto Cassese\***, Rice University  
**Michele Guindani**, University of Texas  
 MD Anderson Cancer Center  
**Mahlet G. Tadesse**, Georgetown University  
**Francesco Falciani**, University of Liverpool  
**Marina Vannucci**, Rice University

We describe a method for the integration of high-throughput data from different sources. More specifically, we focus on combining transcriptomics data (e.g. gene expression profiling) with genomic data, collected on the same subjects. At DNA level we focus on measuring copy number variation (CNV) using comparative genomic hybridization (CGH) arrays. Through the specification of a measurement error model we relate the gene expression levels to latent copy number states which, in turn, are related to the observed surrogate CGH measurement via a hidden Markov model. We develop selection priors to explicitly incorporate dependencies information across adjacent copy number states and we use MCMC stochastic search techniques for posterior inference. Our approach simultaneously infers CNV and identifies their significant associations with mRNA transcripts abundance. The performance of our method is shown on simulated data and we also illustrate an application to data from a genomic study on human cancer cell lines.

email: ac43@rice.edu

## Sampling Designs for Multi-Species Assemblage with Unknown Heterogeneity

**Hongmei Zhang**, University of South Carolina  
**Kaushik Ghosh\***, University of Nevada, Las Vegas  
**Pulak Ghosh**, Indian Institute of Management, Bangalore

In a sample of mRNA species counts, sequences without duplicates or with small numbers of copies are likely to carry information related to mutations or diseases and can be of great interest. However, in some situations, sequence abundance is unknown and sequencing the whole sample to find the rare sequences is not practically possible. To collect mRNA sequences of interest, we propose a two-phase Bayesian sampling method that addresses these concerns. The first phase of the design is used to infer sequence (species) abundance through a cluster analysis applied to a pilot data set. The clustering method is built upon a multivariate hypergeometric model with a Dirichlet process prior for species relative frequencies. The second phase, through Monte Carlo simulations, infers the sample size needed to collect a certain number of species of particular interest. Efficient posterior computing schemes are proposed. The developed approach is demonstrated and evaluated via simulations. An mRNA segment data set is used to illustrate and motivate the proposed method.

email: kaushik.ghosh@unlv.edu

## Bayes Multiple Classification Function in Logic Regression Models

**Wensong Wu\***, Florida International University  
**Tan Li**, Florida International University

In this presentation we consider a two-class classification problem, where the goal is to predict the class membership of  $M$  units based on the values of high-dimensional binary predictor variables as well as both the values of the binary predictor variables and the class membership of other  $N$  independent units. We focus on applying Logic Regression models (LRM), where in general linear regression models the predictors are Boolean expressions of original binary predictors. We consider a Bayesian and decision-theoretic framework, and develop a general form of Bayes multiple classification function (BMCF) with respect to a class of cost-weighted loss functions. In particular, the loss function pairs such as the proportions of false positives and false negatives, and (1-sensitivity) and (1-specificity), are considered, and the cost weights are pre-specified. The best Boolean expressions in LRM are selected using Apriori Algorithm, an efficient algorithm for detecting association rules. The results will be illustrated via simulations and on a Lupus diagnose data set.

email: wenswu@fiu.edu

## Using Informative Priors Obtained from Historical Data Significantly Improves Detection of Differentially Expressed Genes Using Microarray Data

**Ben Li\***, Emory University  
**Qing He**, Emory University  
**Zhaonan Sun**, Purdue University  
**Yu Zhu**, Purdue University  
**Zhaohui Qin**, Emory University

Modern high throughput biotechnologies such as microarray produce massive amount of information for each sample assayed. However, in a typical high throughput experiment, only limited amount of data are observed for each individual feature, thus the classical large  $p$ , small  $n$  problem. On the other hand, rapid propagation of these high throughput technologies has resulted in massive collection of data, often carried out on the same platform and using the same protocol. It is highly desirable to utilize the massive amount of existing data when performing analysis and inference on a new dataset. One possibility is to use the massive repository of historical data to build informative priors under a Bayesian framework and then use them in the downstream inference. In this work, using microarray data, we investigate the feasibility and effectiveness of deriving informative priors from historical data and using them in the problem of detecting differentially expressed genes. Through simulation and real data analysis, we showed that the proposed strategy significantly outperforms competing methods including the popular and state-of-art Bayesian hierarchical model-based approaches. Our work illustrates the feasibility and benefits of exploiting the increasingly available genomics big data in statistical inference, and present a promising strategy for dealing with the large  $p$ , small  $n$  problem.

email: ben.li@emory.edu

### Smoothing Functional Data with a Hierarchical Bayesian Model

**Jingjing Yang\***, Rice University  
**Hongxiao Zhu**, Virginia Tech  
**Dennis D. Cox**, Rice University

A hierarchical Bayesian model is developed for smoothing functional data. Functional data analysis has become popular because of the power of treating particular data as evaluations of smooth functions. However, much research tends to ignore the fact that observed data are originally collected with noise, thus paying no attention to the noise in the process of converting discretized observations to smooth functions. Moreover, most current smoothing methods are usually applied independently on each replication of the observed data, which do not use information from the other replications. We propose a hierarchical Bayesian model with data-driven hyper-priors to smooth all observations simultaneously, possessing the property of borrowing strength across all replications. Furthermore, the model gives smoother Bayesian estimates for the mean and covariance functions. Case studies of simulated and real data demonstrate that this Bayesian method produces more accurate signal estimates and smoother covariance function estimate.

email: jy13@rice.edu

## 33. GENETICS AND EPIDEMIOLOGIC STUDY DESIGN

### Control Function Assisted IPW Estimation with a Secondary Outcome in Case-Control Studies

**Tamar Sofer\***, Harvard School of Public Health  
**Eric J. Tchetgen Tchetgen**, Harvard School of Public Health

Case-control studies are designed towards studying associations between risk factors and a single, primary outcome. Information about additional, secondary outcomes is also collected, but association studies targeting such secondary outcomes should account for the case-control sampling scheme, or otherwise results may be biased. Often, one uses inverse probability weighted (IPW) estimators to estimate population effects in such studies. However, these estimators are known to be inefficient relative to estimators that make additional assumptions about the data generating mechanism. We propose a class of estimators for the effect of risk factors on a secondary outcome in case control studies, when the mean is modelled using either the identity or the log link. The proposed estimator combines IPW with a mean zero control function that depends explicitly on a model for the primary disease outcome. The efficient estimator in our class of estimators reduces to standard IPW when the model for the primary disease outcome is left unrestricted, but more efficient inference than standard IPW is possible by assuming the latter model is either parametric or semiparametric.

email: tsofer@hsph.harvard.edu

### Prediction of Cancer Drugs' Sensitivities Using High-Dimensional Genomic Features

**Ting-Huei Chen\***, University of North Carolina, Chapel Hill  
**Wei Sun**, University of North Carolina, Chapel Hill

A large number of cancer drugs have been developed to target particular genes/pathways that are crucial for cancer growth. Several drugs that share a molecular target may also share some predictive genomic features for drugs' sensitivity. Therefore, it is desirable to analyze these drugs as a group to identify the associated genomic features, which may provide biological insights underlying drug responses. Furthermore, those identified genomic features may be robust predictors for any drug sharing the same target. High-dimensionality and strong correlations among genomic features are main challenges of this task. Motivated by this problem, we develop a new method for high-dimensional bi-level feature selection using a group of response variables that may share a common set of predictors in addition to their individual ones. Simulation results show that our method can have substantially higher sensitivity and specificity than existing methods. We apply our method to two large-scale drug sensitivity studies in cancer cell lines. Within study cross-validation demonstrates the high prediction power of our method. Between study validation shows that the genomic features selected for a drug target can be good predictors for other drugs designed for the same target.

email: thchen@live.unc.edu

### Enhancing Genetic Case-Control Studies Using Sample Surveys

**Parichoy Pal Choudhury\***, Johns Hopkins University  
**Daniel Scharfstein**, Johns Hopkins University  
**Joshua Galanter**, University of California, San Francisco  
**Chris Gignoux**, University of California, San Francisco  
**Lindsey Roth**, Kaiser Permanente  
**Sam Oh**, University of California, San Francisco  
**Esteban Burchard**, University of California, San Francisco  
**Saunak Sen**, University of California, San Francisco

Genetic case-control studies are useful for characterizing the association between a gene and case-control status, the so-called primary phenotype. In these studies, it is common for investigators to collect additional health outcomes, referred to as secondary phenotypes. It is often natural to ask about the association between a gene of interest and a secondary phenotype. Since case-control data are not a random sample from the target population, the observed association between the gene and secondary phenotype may be biased. External information is required to correct for this bias. We propose an inferential framework for learning about the association between a gene and a secondary phenotype, utilizing information from a case-control and a representative sample survey from a target population. By way of illustration, we study the relationship between

a candidate gene (linked to asthma) and obesity and how this relationship differs by ethnicity. We use data from the GALA II study (an asthma case-control study in Latino American children) and the NCHIS study (a national sample survey of children). Information from these two distinct data sources are combined to estimate standardized associations between the gene and obesity within ethnicity strata, which are then compared across the different ethnicities.

email: ppalchou@jhsph.edu

### **The Effect of FTO Gene Variants and Physical Activity Interaction on Trunk Fat Percentage Among the Population of Newfoundland**

**Anthony Payne**, Memorial University  
**Taraneh Abarin\***, Memorial University  
**Farrell Cahill**, Memorial University  
**Guang Sun**, Memorial University  
**J Concepción Loredo-Osti**, Memorial University

Overweight and obesity is a major epidemic in Newfoundland and Labrador (NL), as well as throughout Canada. Data from the 2012 Canadian census shows that NL has had the highest percentage of self-reported overweight/obese residents in Canada since 2008 with 63.2% of adults reporting being overweight or obese that year. It is widely acknowledged that there is a significant genetic contribution to obesity-related traits. However, the genetic contribution to trunk fat percentage, which is more closely associated with obesity, is not completely understood. We studied 3,008 individuals from the Newfoundland population with trunk fat percentage measured by dual-energy x-ray absorptiometry and their physical activities. We genotyped eleven single-nucleotide polymorphisms in the fat mass and obesity associated (FTO) gene. We are currently investigating that whether FTO variants and the corresponding haplotypes are associated with PTF, and if so, whether the detrimental associations of FTO gene variants and haplotypes can be lessened by increased physical activity.

email: tabarin@mun.ca

### **On The Underlying Assumptions of Threshold Boolean Networks as a Model for Genetic Regulatory Network Behavior**

**Van Tran\***, University of Rochester Medical Center  
**Mathew N. McCall**, University of Rochester Medical Center  
**Helene McMurray**, University of Rochester Medical Center  
**Anthony Almudevar**, University of Rochester Medical Center

Boolean networks (BoN) are simple and interpretable models of gene regulatory networks (GRN). Specifying these models with fewer parameters while retaining their ability to describe complex regulatory relationships is a methodological challenge. We explore a previously-proposed class of BoNs characterized by linear threshold functions, which we refer to as threshold Boolean networks

(TBN). Compared to traditional BoNs, these models require fewer parameters and a more direct interpretation. However, the functional form of a TBN results in a reduction of regulatory relationships that can be modeled. We show that TBNs can be extended with a state augmentation procedure to permit Markovian variable self-degradation. Next, we propose two theorems relating self-degradation and regulatory feedback to the steady state behavior of a TBN. Finally, we show how TBNs can be extended to relax the assumptions of synchronous gene response and asynergistic regulation. Applying our methods to the budding yeast cell-cycle network revealed that although the network is complex, its steady state is simplified by the presence of self-degradation and lack of purely positive regulatory cycles.

email: thanh\_tran@urmc.rochester.edu

### **Evaluation of Illumina Infinium 450K Methylation Chip Using Technical Replicates**

**Maitreyee Bose\***, University of Minnesota  
**Weihua Guan**, University of Minnesota  
**Chong Wu**, University of Minnesota  
**James Pankow**, University of Minnesota  
**Ellen Demerath**, University of Minnesota  
**Jan Bressler**, University of Minnesota

DNA methylation is the widely studied epigenetic mechanism that alterations in methylation patterns may be involved in the development of common diseases. As part of the Atherosclerosis Risk in Communities (ARIC) study, the Illumina Infinium HumanMethylation450 (HM450) BeadChip was used to measure DNA methylation in peripheral blood obtained from ~3000 participants. Over 480,000 cytosine-guanine (CpG) dinucleotide sites were surveyed on the HM450 BeadChip. For many of the CpG sites, significant difference was observed on methylation levels between samples on different plates and chips, largely due to technical error. To evaluate the impact of technical errors, 135 technical replicates were included in the study. For each CpG site, we calculated the intraclass correlation coefficient (ICC) to compare variation of methylation levels within- and between-replicate pairs. Given a large proportion of observed ICC values at 0, we modeled the distribution of ICC as a mixture of censored or truncated normal and normal distributions using an EM algorithm. The CpG sites are clustered into low- and high-reproducibility groups, according to the calculated posterior probabilities. We also demonstrate the performance of this clustering when applied to the study of association between methylation levels and smoking habits of individuals.

email: bosex020@umn.edu

### **Leveraging Family History in Genetic Association Studies**

**Arpita Ghosh\***, Public Health Foundation of India

**Patricia Hartge**, National Cancer Institute,  
National Institutes of Health

**Peter Kraft**, Harvard School of Public Health

**Amit D. Joshi**, Harvard School of Public Health

**Regina G. Ziegler**, National Cancer Institute,  
National Institutes of Health

**Myrto Barrdahl**, German Cancer Research Center

**Stephen J. Chanock**, National Cancer Institute,  
National Institutes of Health

**Sholom Wacholder**, National Cancer Institute,  
National Institutes of Health

**Nilanjan Chatterjee**, National Cancer Institute,  
National Institutes of Health

Population-based epidemiologic studies often gather information from study participants on disease history among their family members. Although investigators widely recognize that family history will be associated with genotypes of the participants at disease susceptibility loci, they commonly ignore such information in primary genetic association analyses. In this report, we propose a simple approach to association testing by incorporating family history information as a phenotype. We account for the expected attenuation in strength of association of the genotype of study participants with family history under Mendelian transmission. The proposed analysis can be performed using standard statistical software adopting either a meta- or pooled-analysis framework. Re-analysis of a total of 115 known susceptibility SNPs, discovered through genome-wide association studies for several disease traits, indicates that incorporation of family history information can increase efficiency by as much as 40%. Efficiency gain depends on the type of design used for conducting the primary study, extent of family history and accuracy and completeness of reporting.

email: arpita.ghosh@phfi.org

## **34. NON-LINEAR MODELS**

### **Single Index Change Point Model with an Application of Environmental Health Study on Mortality and Temperature**

**Hamdy Mahmoud\***, Virginia Tech

**Inyoung Kim**, Virginia Tech

**Ho Kim**, Seoul National University

Environmental health studies are of great interest in human research to evaluate the relationship between daily/weekly mortality and temperature. It has been shown that there is a nonlinear relationship between these two with a fixed number of change points for temperature. The current available methods consist of two steps: they first estimate the models and then detect change points. However, the methods for simultaneously identifying the nonlinear relationship and detecting the number of change point(s) are quite limited. Therefore, in this paper we propose a unified approach to simultaneously estimate the nonlinear relationships and detect the change points. We propose a

single index change point model as our unified approach by adjusting for several other covariates. We also provide a permutation based testing procedure to detect multiple change points. Our proposed model is compared with the generalized linear model and generalized additive model using simulation and a real application. Our simulation results suggest that our approach performs better than other two in terms of type I error and power. We also show the asymptotic properties of the permutation test in single index change point model, suggesting that the number of change points is consistent.

email: ehamdy@vt.edu

### **A Model for Extreme Stacking of Data Censored at Endpoints of a Distribution with a Continuous Interior: Illustration with W-Shaped Data**

**Robert Gallop\***, West Chester University

**Randall H. Rieger**, West Chester University

**Scott McClintock**, West Chester University

**David C. Atkins**, University of Washington

Regression analyses will only provide correct inferences when certain assumptions about the data are met. When data violate these assumptions we could do nothing and rely on the Central Limit Theorem (CLT). However, it is rarely clear how big a sample size need be for the CLT to protect against Type I errors. Another common remedy is to seek a transformation of the outcome variable that leads to normally distributed residuals. No transformation can spread out a stack of data. When data is stacked and censored at two extremes with a near normal distribution in between, resulting in a bimodal W-shape distribution, a new modeling structure must be used. Medical devices may have two extremes scales. Whether the response could be lower the low extreme or higher than the high extreme is unknown. In this presentation, we propose a mixture model in which the complete distribution of the outcome is approximated by mixing three component distributions: the two censored extreme responses and the normally spread responses in between. We illustrate this modeling structure for biomarkers quantifying health status in a sample of cows. We compare our results to the standard modeling approaches with respect to goodness of fit.

email: rgallop@wcupa.edu

### **Estimating a Dengue Ordinary Differential Equation Model with the Mesh Adaptive Direct Search Method**

**Yu-Ting Weng\***, University of Pittsburgh

**Shawn T. Brown**, Pittsburgh Supercomputing Center

**Nathan Stone**, Pittsburgh Supercomputing Center

**Abdus S. Wahed**, University of Pittsburgh

Modeling dengue incidence over time is challenging. Ordinary Differential Equation (ODE) models are standard approaches to describe complex systems involving interactions between various populations. In this paper, we propose a set of ODEs to model the incidence of dengue infection that elucidates the interaction between humans and mosquitos throughout the

life cycle of *Aedes aegypti*. Estimation of ODE parameters from real world data is usually done via discretization methods and collocation methods. However, few can provide satisfactory results in the class for which explicit information about the analytical or numerically approximated gradient or the Hessian matrix of objective function is unavailable. Here, we utilize mesh adaptive direct search (MADS) for finding a global minimized estimator for the dengue model, and combine this with a sieve bootstrap for non-stationary time series to estimate its confidence interval. This method is one of the derivative free discretization methods, and is derived from a generalized pattern search method. The simulation studies show that MADS provides unbiased estimators and the Monte Carlo standard errors of the estimators are similar to the mean of the bootstrap standard errors. The method is demonstrated by fitting our ODE dengue model to a real world dengue fever incidence data.

email: yuw22@pitt.edu

### Parametric and Nonparametric Spherical Regression

**Michael M. Rosenthal\***, Florida State University

**Wei Wu**, Florida State University

**Eric Klassen**, Florida State University

**Anuj Srivastava**, Florida State University

The potential applications of regression analysis restricted to spherical variables are useful in many branches of science. The challenge of developing methods for spherical regression comes from the non-linearity of the sphere. We present two approaches for spherical regression - one parametric and one non-parametric - that utilize the geometry of the sphere to develop estimation procedures. We derive algorithms for maximum likelihood estimation for parametric and penalized maximum likelihood for non-parametric models. The performances of the proposed models are compared using simulated data and real data from vector cardiograms and shape based image analysis. Additionally, we provide an asymptotic analysis for the MLE of the parametric model.

email: michaelr@stat.fsu.edu

### Non-Parametric Tests for One-Sided Interaction in Shape Restricted Models

**Mingyu Xi\***, University of Maryland, Baltimore County

In chemical-chemical or drug-drug interaction problems, people are often interested in the direction of interaction such as whether the agents interact in a synergistic manner or whether they have antagonistic behavior. Thus, the hypothesis of interest is often that of a one-sided interaction. The response model is usually in the form of a regression where a response  $y$  is observed for a mixture of covariates  $(x_1, x_2)$ , and the mean function is shape restricted. We want to test for interaction of  $x_1$  and  $x_2$  under a general nonparametric model for the response. In particular we are interested in the direction of interaction relative to some null interaction model such as Bliss independence. Based on the available work on one-sided tests for multivariate normal mean, we propose a test for directional interaction

in the context of a nonparametric family of models defined by Bernstein functions. We investigate properties of the proposed test and also demonstrate its performance via simulation.

email: mxi1@umbc.edu

### Sparse Kernel Machine Regression for Ordinal Outcomes

**Yuanyuan Shen\***, Harvard School of Public Health

**Katherine Liao**, Brigham and Women's Hospital

**Tianxi Cai**, Harvard School of Public Health

Ordinal outcomes arise frequently in clinical studies when each subject is assigned to a category and the categories have a natural order. Classification rules for ordinal outcomes may be developed using commonly used regression models such as the continuation ratio model (CR) and the proportional odds (PO) model. For settings where the covariate effects differ between some categories but not all, fitting a full CR model may be inefficient due to overfitting while fitting a PO model may lead to classification rules with poor performance due to underfitting. In addition, these standard models do not allow for non-linear covariate effects. In this paper, we propose a sparse CR kernel machine (KM) regression method for ordinal outcomes where we use the KM framework to incorporate non-linearity and impose sparsity on the overall differences between adjacent categories in the covariate effects to control for overfitting. In addition, we provide data driven rule to select the optimal kernels to maximize the prediction accuracy. Simulation results show that our proposed procedures perform well under both linear and non-linear settings, especially when the true underlying model is in-between full CR model and PO model. We apply our procedures to an autoantibody dataset to illustrate the advantage of our method over other commonly used methods.

email: yshen@g.harvard.edu

### Regression Models on Riemannian Symmetric Spaces

**Emil A. Cornea\***, University of North Carolina, Chapel Hill

**Hongtu Zhu**, University of North Carolina, Chapel Hill

**Joseph G. Ibrahim**, University of North Carolina, Chapel Hill

The aim of this paper is to develop a general regression framework for the analysis of manifold-valued response in a Riemannian symmetric space (RSS) and its association with covariates of interest, such as age, in Euclidean space. Such RSS-valued data arises frequently in medical imaging, computational biology, molecular imaging, surface modeling, and computer vision, among many others. Little has been done when the response is in a general RSS. We develop an intrinsic regression model solely based on an intrinsic conditional moment assumption, avoiding specifying any parametric distribution in RSS. We propose

various link functions to map from the Euclidean space of covariates to the the RSS of responses. We develop a two-stage procedure to calculate the parameter estimates, and determine their asymptotic distributions. We construct the Wald and geodesic test statistics to test hypotheses of unknown parameters. We systematically investigate the geometric invariant property of these estimates and test statistics. Simulation studies are used to evaluate the finite sample properties of our methods and a real data set is analyzed to illustrate the use of our test statistics.

email: ecornea@bios.unc.edu

## 35. SURVIVAL ANALYSIS FOR CLINICAL TRIAL DATA

### Sample Size Calculation Based on Efficient Unconditional Tests for Clinical Trials with Historical Controls

**Guogen Shan\***, University of Nevada, Las Vegas  
**Sheniz Moonie**, University of Nevada, Las Vegas

In historical clinical trials, the sample size and the number of success in the control group are often considered as fixed. The traditional method for sample size calculation is based on the asymptotic approach developed by Makuch and Simon (1980). Exact unconditional approaches may be considered as alternative to control for the type I error rate where the asymptotic approach may fail to do so. We propose the sample size calculation using an efficient exact unconditional testing procedure based on estimation and maximization, and compare this with the sample size calculation using asymptotic and exact approaches. The sample size calculation based on the new exact unconditional approach is generally smaller than those based on the other approaches, especially when the sample size in the control group is not too small. The procedure is recommended for use due to the substantial sample size savings experienced.

email: guogen.shan@unlv.edu

### Sieve Estimation in a Markov Illness-Death Process Under Dual Censoring

**Audrey Boruvka\***, University of Waterloo  
**Richard J. Cook**, University of Waterloo

Estimation of the proportional hazards model is considered for a progressive Markov illness-death process under two censoring mechanisms acting separately on intermediate and terminal events, a scheme we refer to as dual censoring. The proposed estimator globally converges to the truth slower than the parametric rate. However, under certain conditions, its component for the regression coefficient is asymptotically normal and achieves the semiparametric information bound. A simulation study shows that the estimator performs well compared to common alternatives and is robust to some forms of dependent censoring. The new methods are illustrated using data from cancer trials.

email: ajboruvka@uwaterloo.ca

### A Simple Locally Efficient Estimator for Relative Risk in Case-Cohort Studies

**Emmanuel Sampene\***, University of Pittsburgh  
**Abdus S. Wahed**, University of Pittsburgh

A case-cohort study is a two-phase study where at the first phase a representative sample, referred to as the study cohort, is selected from the target population. At the second phase, a subsample is selected from the cohort based on the case status. All cases are included in the subsample whereas only a random sample of controls is included. The endpoint of interest in such studies is usually the failure time. Several methods have been proposed to estimate the relative risk or hazard ratio from a case-cohort study. These methods almost always disregard the covariate information that is not included in the sampled study sub-cohort, and therefore, results in the loss of efficiency. While there have been attempts to derive the most efficient estimators, the resulting estimators were challenging from the data analysis point of view. We propose a locally efficient estimator (LEE) based on Robins et al. (1994, J. AM Stat. Assoc. 89,846-866) by restricting the estimator to a class of regular asymptotically linear estimators. The properties of this estimator are investigated through simulation and application to the Wilm's tumor study.

email: ems120@pitt.edu

### Generation of Virtual Control Groups for Single Arm Prostate Cancer Adjuvant Trials

**Zhenyu Jia\***, University of Akron and  
Northeast Ohio Medical University  
**Michael B. Lilly**, Medical University of South Carolina  
**Dan A. Mercola**, University of California, Irvine

It is difficult to construct a control group for trials of adjuvant therapy (Rx) of prostate cancer after radical prostatectomy (RP) due to ethical issues and patient acceptance. We utilized 8 curve-fitting models to estimate the time to 60%, 65%, ... 95% chance of progression free survival (PFS) based on the data derived from Kattan post-RP nomogram. The 8 models were systematically applied to a training set of 153 post-RP cases without adjuvant Rx to develop 8 subsets of cases (reference case sets) whose observed PFS times were most accurately predicted by each model. To prepare a virtual control group for a single-arm adjuvant Rx trial, we first select the optimal model for the trial cases based on the minimum weighted Euclidean distance between the trial case set and the reference case set in terms of clinical features, and then compare the virtual PFS times calculated by the optimum model with the observed PFSs of the trial cases by the logrank test. The method was validated using an independent dataset of 155 post-RP patients without adjuvant Rx. We then applied the method to patients on a Phase II trial of adjuvant chemo-hormonal Rx post RP, which indicated that the adjuvant Rx is highly effective in prolonging PFS after RP in patients at high risk for prostate cancer recurrence.

email: zjia@uakron.edu

### Imbalanced Randomization in Non-Inferiority Trials can be Highly Efficient

**Rick Chappell\***, University of Wisconsin, Madison

Non-inferiority (equivalence) trials are clinical experiments which attempt to show that one intervention is not too much inferior to another on some quantitative scale. One of their interesting features is that balanced randomization may not constitute the most efficient approach; in fact, strongly imbalanced allocation can greatly increase efficiency in many circumstances. These include when outcomes have nonconstant variance, such as in binary and survival cases, and when the non-inferiority margins are on a relative rather than absolute scale. My presentation gives details and examples of this phenomenon.

email: chappell@stat.wisc.edu

### Estimating Survival Benefit in Randomized Clinical Trials with Treatment Arm Switching after Disease Progression

**Shan Kang\***, University of Michigan

**Thomas M. Braun**, University of Michigan

Phase III clinical trials in oncology commonly examine if there is an increased survival benefit for a new treatment relative to an existing treatment or standard-of-care. Although patients are randomized to their treatment assignments, ethical motivations dictate that patients who experience disease recurrence or other event indicating increased likelihood of death may be offered the option to switch to the other treatment arm and continue to be followed for survival. Standard statistical methods that ignore this non-random treatment arm switching can lead to biased estimation of the survival benefit attributed to the new treatment. Although methods do exist to account for the effect of treatment arm switching, several of these methods focus on quantifying an overall switching effect, which can still lead to biased results if the benefit derived from switching varies among patients. We propose a new parametric method to address this limitation that factorizes the likelihood into two parts in order to evaluate the individual benefit of switching. Via simulation, we examine the performance of our method and compare the performance with existing methods.

email: shankang@umich.edu

### Semiparametric Proportional Rates Regression for the Composite Endpoint of Recurrent and Terminal Events

**Lu Mao\***, University of North Carolina, Chapel Hill

**Danyu Lin**, University of North Carolina, Chapel Hill

Recurrent event data are commonly encountered in clinical and epidemiological studies. A major complication arises when recurrent events are terminated by death. To assess the overall covariate effects on the two types of events, we define the composite endpoint as the cumulative number of recurrent and terminal events over time and propose a semiparametric proportional rates model which specifies that the (possibly time-varying) covariates have multiplicative effects on the rate function of the composite endpoint while leaving the form of the rate function and

the dependence among recurrent and terminal events completely unspecified. We construct appropriate estimators for the regression parameters and the cumulative frequency function. We show that the estimators are consistent and asymptotically normal with variances that can be consistently estimated. Simulation studies demonstrate that the proposed methods perform well in realistic situations. An application to a cancer clinical trial is provided.

email: lmao@unc.edu

## 36. CLUSTERED DATA METHODS

### A New Semiparametric Approach to Finite Mixture of Regressions Using Penalized Regression Via Fusion

**Erin Austin\***, University of Minnesota

**Wei Pan**, University of Minnesota

**Xiaotong Shen**, University of Minnesota

For some modeling problems a population may be better assessed as an aggregate of unknown subpopulations, each with a distinct relationship between a response and associated variables. The finite mixture of regressions (FMR) model, where an outcome is derived from one of a finite number of linear regression models, is a natural tool in this setting. In this article we first propose a novel penalized regression approach, then we demonstrate how it can, in some types of problems, better identify subpopulations and their corresponding models than a semiparametric FMR method. Our new method fits models for each person via grouping pursuit, utilizing a new group truncated L1-penalty (gTLP) that shrinks differences between estimated parameter vectors. The methodology causes the subjects' regression coefficients to cluster into a few common models, in turn revealing previously unknown subpopulations. In fact, by varying the penalty strength, the new method can reveal a hierarchical structure among the subpopulations that can be useful in exploratory analysis. Simulations using FMR models and real data analysis show the performance of the method is promising.

email: austi260@umn.edu

### Semi-Parametric Models for Clustered Survival Data with Random Cluster Size

**Shuling Liu\***, Emory University

**Amita K. Manatunga**, Emory University

**Limin Peng**, Emory University

We consider semi-parametric regression analysis of the clustered survival data with random cluster size. Our method is motivated by the Mount Sinai Study of Women Office Workers (MSSWOW) in which menstrual cycle lengths are recorded until time-to-pregnancy (TTP) or the end of study for each woman. Since TTP is a discrete random variable, it is natural to view the data as clustered menstrual cycle

lengths with the random cluster size equals to TTP. We consider a semi-parametric framework where survival times (menstrual lengths) are modeled by the Clayton-Oakes model with the dimension indexed by the random cluster size (TTP). We parameterize the covariate effects on clustered survival times using a linear semi-parametric transformation model. A hazard regression model is assumed for the cluster size to incorporate potential risk factors. We propose an estimation procedure that can appropriately accommodate missing and censoring in survival times as well as truncation and censoring in the cluster size. Simulation studies are conducted to illustrate the performance of the proposed method. Finally we apply our method to the MSSWOW study.

email: sliu34@emory.edu

### **Identification of Biologically Relevant Subtypes Via Preweighted Sparse Clustering**

**Sheila Gaynor\***, Harvard University

**Eric Bair**, University of North Carolina, Chapel Hill

Cluster analysis methods are used to identify homogeneous subgroups in a data set. Frequently one applies cluster analysis in order to identify biologically interesting subgroups. In particular, one may wish to identify subgroups that are associated with a particular outcome of interest. Conventional clustering methods often fail to identify such subgroups, particularly when there are a large number of high-variance features in the data set. Conventional methods may identify clusters associated with these high-variance features when one wishes to obtain secondary clusters that are more interesting biologically or more strongly associated with a particular outcome of interest. We describe a modification of the sparse clustering method of Witten and Tibshirani (2010) that can be used to identify such secondary clusters or clusters associated with an outcome of interest. We show that this method can correctly identify such clusters of interest in several simulation scenarios. The method is also applied to a large case-control study of TMD and a leukemia microarray data set.

email: sgaynor@fas.harvard.edu

### **Estimation Methods for Copula Models for Discrete Clustered and Longitudinal Data**

**N. Rao Chaganty\***, Old Dominion University

The multivariate normal distribution is often used in the analysis of clustered and longitudinal continuous data, but no corresponding multivariate distribution analogue has been commonly accepted for discrete data such as binary, count or ordinal. However, copulas facilitate modeling the joint distributions for discrete responses. Specifically, exchangeable copulas can be used to model clustered discrete data, while longitudinal discrete data can be modeled by an appropriate copula with decreasing time-lag dependence. The specification of the multivariate discrete

distribution through the use of copulas provides complete inference, in the sense that maximum likelihood estimation of dependence and marginal parameters is possible. In this talk, I will discuss various methodologies for parameter estimation for the copula models.

email: rchagant@odu.edu

### **Biclustering Via Sparse Clustering**

**Qian Liu\***, University of North Carolina, Chapel Hill

**Guanhua Chen**, University of North Carolina, Chapel Hill

**Michael R. Kosorok**, University of North Carolina, Chapel Hill

**Eric Bair**, University of North Carolina, Chapel Hill

In many situations it is desirable to identify clusters that differ with respect to only a subset of features. Such clusters may represent homogeneous subgroups of patients with a disease, such as cancer or chronic pain. We define a bicluster to be a submatrix  $U$  of a larger data matrix  $X$  such that the features and observations in  $U$  differ from those not contained in  $U$ . For example, the observations in  $U$  could have different means or variances with respect to the features in  $U$ . We propose a general framework for biclustering based on the sparse clustering method of Witten and Tibshirani (2010). We develop a permutation-based method for identifying features that belong to biclusters. This framework can be used to identify biclusters that differ with respect to the means of the features, the variance of the features, or more general differences. We apply these methods to several simulated and real-world data sets and compare the results of our method with several previously published methods. The results of our method compare favorably with existing methods with respect to both predictive accuracy and computing time.

email: qliu@live.unc.edu

### **Composite Likelihood Inference for Multivariate Finite Mixture Models with Application to Flow Cytometry Data**

**Fei Ma\***, University of Rochester

**Ollivier Hyrien**, University of Rochester

Finite mixture models have found numerous applications in various fields, including medicine, biology and genetics, image analysis, where they are routinely used for data clustering, for example. These models can be praised for their flexibility and for offering statistical frameworks that are able to describe latent heterogeneity in the data, yet they face numerous inferential and computational challenges. For example, parameter estimation may be sensitive to departures from modeling assumptions, causing estimators to be biased. This problem appears when dealing with multivariate data if the joint distribution of each component is misspecified. We propose a composite likelihood approach to parameter estimation which requires only the joint distributions to be specified in low dimensions. We develop an EM algorithm for composite likelihood and propose a modification of the algorithm that speeds up calculations. Finite and large sample properties of the method will be presented and investigated using simulations. An application to flow cytometry data will be discussed.

email: fei\_ma@urmc.rochester.edu

### Weighted Quartile Sum Regression for Assessing The Association of Environmental Chemical Mixtures and Oral Health

**Bhanu M. Evani\***, Virginia Commonwealth University  
**Chris Gennings**, Virginia Commonwealth University

Mixture effects of environmental chemicals measured in serum concentrations on the outcome of periodontitis is unknown. Polychlorinated biphenyls (PCBs) occur clustered in the environment due to industrial applications. We anticipate that not all thirty-four PCBs measured by NHANES in 2003-04 are biologically important to periodontitis. Our objective is to determine whether a weighted mixture of PCBs is associated with periodontitis, and if so, identify the bad actors. Ordinary regression methods and other variable selection methods suffer from multicollinearity issues. We used weighted quartile sum (WQS) regression to estimate the body burden due to these clusters of correlated chemicals, adjusting for important covariates. Carrico et al. (2013), show that interpretations of the empirically estimated weights from WQS regression are more robust to the correlation pattern between the chemicals compared to other methods. We use a dendrogram that identifies the clusters of correlated chemicals, and using the estimated weights on the same diagram, we identify important chemicals within the clusters. We validate the method using bootstrapped simulations. Present research focuses on the direct action between an empirically weighted body burden index of PCB congeners and the outcome of periodontitis. Future work will develop methods for identifying mediator(s) for chemical mixtures.

email: evanibm@vcu.edu

## 37. STATISTICAL METHODS FOR LONGITUDINAL DATA

### Sample Size Determination for Longitudinal Binary Response Data Based on Testing the Difference in Rate of Change in Log Odds Ratio Between Groups

**Kush Kapur\***, Boston Children's Hospital and Harvard Medical School

**Dulal K. Bhaumik**, University of Illinois, Chicago

In this talk we will discuss approaches for determining the necessary sample size while comparing the efficacy of an intervention compared to a control for repeated binary response data in order to achieve a pre-specified power using logistic mixed-effects model. The approaches that we will describe will be generalized to accommodate varying attrition rates over time for three level mixed-effects models. Based on simulation study, we have established that the second order Taylor series expansions are not adequate in the context of sample size determination problem for the longitudinal binary data. We will also discuss the relationship of sample size in terms of variance and covariance parameters of the subject-effects. We will finally provide few guidelines for sample size allocation for unbalanced recruitment of subjects in the treatment groups and discuss different links between various types of measurements.

We will end this talk with two illustrations: The first study was conducted for comparison of two doses of depot-medroxyprogesterone acetate drug for contracepting women, and the second study was designed to test the efficacy of hypnotic drug in comparison to placebo for patients suffering from insomnia.

email: kush.kapur@childrens.harvard.edu

### Model Selection of Generalized Estimating Equations with Multiple Imputation and High-Dimensional Covariates For Missing Longitudinal Data

**Ming Wang\***, The Pennsylvania State College of Medicine

Missing longitudinal data with high-dimensional covariates has gained research attentions in biomedical studies. Generalized Estimating Equation (GEE), a marginal statistical method, is commonly used for longitudinal data analysis, and multiple imputation (MI) is popularly employed to handle missingness. Note that how to select working correlation matrix and covariates plays a vital role in improving the efficiency of the parameter estimates and the model goodness-of-fit; however, limit work exists on development of model selection strategies for GEE with MI. In this work, we propose a MI-based weighted Quasi-likelihood approach to account for sampling and imputation uncertainty. Also, we extend this proposal to the cases with high-dimensional covariates using penalized techniques for further evaluation. In addition, several existing alternatives including the quasi-likelihood under the independence model criterion (QIC) and the missing longitudinal information criterion (MLIC) are compared and evaluated to show our proposals outperformance. Finally, the proposed method is illustrated by a real data example from a colorectal study.

email: mwang@phs.psu.edu

### An EM Algorithm for Multilevel Multivariate Mixed Effect Model with Unstructured Error Covariance

**Yun Ling\***, University of Pittsburgh

**Stewart J. Anderson**, University of Pittsburgh

Multivariate longitudinal models that only accommodate single-level random effects with unstructured covariance have been discussed [Shah, 1997; Yucel, 1999]. Software packages, e.g., SAS PROC MIXED and R MLMEM, allow one to estimate parameters. However, often multivariate repeated measurements on subjects are clustered into groups. In ophthalmic data, multiple quantities are repeated measured on each eye, and each patient has two eyes. Here, parameter estimation is straightforward if the error covariance matrix is diagonal. But, for unstructured error covariance, or if some characteristics are not observed at all visits, no existing packages provide parameter estimation. We use the EM algorithm to estimate the parameter for this case. Two difficulties arise: (1) error terms of multiple characteristics are correlated at the same visit; and (2) all characteristics are not observed at all occasions. We apply our method to glaucoma

data where multiple characteristics, e.g. RNFL and GCC thickness are repeatedly measured for each eye. The 2-level bivariate repeated measurements are modeled allowing estimating 2-level random effects and correlation between characteristics.

email: yul27@pitt.edu

### **Regression Methodology for Comparing Longitudinal Rates of Change**

**Matthew W. Bryan\***, University of Pennsylvania  
**Patrick Heagerty**, University of Washington

Longitudinal data can compare groups across an outcome at the mean level and the rate level. Powerful methods for longitudinal data, such as the linear model, have been developed for estimating mean level differences. In addition, the linear model can estimate rate level difference through the inclusion of appropriate interaction terms. However, when differentiating rates of change under the presence of non-linear trend in time, the standard linear models approach is insufficient. Other methods have been developed for modeling rates of change that allow a general time trajectory, but these methods have not addressed comparing rates across covariate defined groups. Thus, we propose regression methods for differentiating longitudinal rates of change. Our proposed method offers a parsimonious comparison of rates of change relative to a generally specified time trajectory. The regression model demonstrates increased power to detect group difference in the rate of change relative to a standard linear models approach. The proposed approach to comparing longitudinal rates of change is illustrated using a study of growth among infants exposed to HIV infection.

email: bryanma@upenn.edu

### **Three-Step Estimation Via Local Polynomial Smoothing for Unevenly Sampled Longitudinal Data**

**Lei Ye\***, University of Pittsburgh  
**Ada O. Youk**, University of Pittsburgh  
**Susan M. Sereika**, University of Pittsburgh  
**Lora E. Burke**, University of Pittsburgh

Parametric and nonparametric mixed models are useful in longitudinal data analysis when the sampling frequency of response and covariate is the same. Three-step estimation via local polynomial smoothing was proposed and demonstrated with data where the dependent variable was more frequently sampled than the independent variable within the same time frame. We first inserted pseudo data points for the dependent variable based on observed measurements to create an even dataset. Then local linear regressions were fitted at each time point to obtain raw estimators of the association between dependent and independent variables. Lastly the local polynomial model was applied to smooth the raw estimators. Rather than using a kernel distribution to assign weights, only analytic weights that indicate the importance of the raw estimators were used. The standard error of the raw estimator and the distance between the pseudo data points and observed measurement points were considered as the measure of

importance of the raw estimators. The results showed that the selection of analytic weights is critical especially when the relationship between the dependent and independent variable is nonlinear.

email: ley9@pitt.edu

### **The Use of Tight Clustering Techniques for Group-Based Trajectory Modeling of Longitudinal Data Accounting for Random Intercepts**

**Ching-Wen Lee\***, University of Pittsburgh  
**Lisa A. Weissfeld**, University of Pittsburgh

Latent group-based trajectory modeling has been widely used to categorize individuals into several homogeneous trajectory groups. In settings where a small number of individuals have unique trajectory patterns that are not similar to those observed in the rest of the population, the latent group-based trajectory modeling techniques may end up either identifying a larger number of latent trajectory groups with several groups containing very few individuals or including these unique trajectory patterns into a latent group where they are a distinct minority. The current techniques make it difficult to identify homogeneous groups within the population and fail to highlight those patterns that may be of greatest interest, namely trajectories that follow distinctly unique patterns. This work applies the idea of the tight clustering method in the human genetics field to group-based trajectory analysis with linear mixed models to classify latent trajectory groups that are more homogeneous, and to identify miscellaneous individuals or outliers whose trajectory patterns are dissimilar to the patterns in the rest of the population. We use the Bayesian information criterion as the criterion for model selection and present simulation studies to examine the properties of the proposed method. An example from a neuroimaging study is provided.

email: erica.gin@gmail.com

### **Monotone Spline-Based Nonparametric Estimation of Longitudinal Data with Mixture Distribution**

**Wenjing Lu\***, University of Iowa  
**Ying Zhang**, University of Iowa

In this paper, we propose a monotone spline-based nonparametric estimation method to analyze longitudinal data with mixture distribution. An iterative algorithm based on k-means clustering technique and non-linear mixed-effects model is implemented to fit the mixture distribution under longitudinal measures. A disparity index based on aggregated areas under the curve (AAUC) is developed to measure the difference between the underlying distributions with longitudinal data. An extensive simulation study is conducted to assess the validity of the proposed method under different values of AAUC. Finally, this method is applied to a multi-site observational study of Huntington's disease (HD), PREDICT-HD, to ascertain sensitive clinical markers in motor and cognitive domains that are possibly capable of distinguishing participants at-risk of HD from healthy controls.

email: luwenjing1028@gmail.com

## 38. RECENT DEVELOPMENTS IN ESTIMATING THE HEALTH EFFECTS OF AIR POLLUTION AND REGULATION

### A Distributed Exposure Time-to-Event Model for Estimating Associations Between Air Pollution and Preterm Birth

**Howard H. Chang\***, Emory University  
**Joshua L. Warren**, University of North Carolina, Chapel Hill  
**Lyndsey A. Darrow**, Emory University  
**Brian J. Reich**, North Carolina State University  
**Lance A. Waller**, Emory University

In reproductive epidemiology, there is a growing interest to examine associations between air pollution exposures during pregnancy and the risk of preterm birth (PTB). One important research objective is to identify critical periods of exposure and estimate the associated effects at different stages of pregnancy. However, population studies have reported inconsistent findings, which may be due to limitations from the standard analytic approach of treating PTB as a binary outcome without considering time-varying exposures together over the course of pregnancy. To address this research gap, we present a Bayesian hierarchical model for conducting a comprehensive examination of gestational air pollution exposure by estimating the joint effects of weekly exposures during different vulnerable windows. We applied the proposed model to a dataset of geocoded birth records in the Atlanta metropolitan area between 1999 - 2005, and examined the risk of PTB associated with gestational exposure to ambient fine particulate matter less than 2.5 micrometers in aerodynamic diameter (PM<sub>2.5</sub>). We found positive associations between PM<sub>2.5</sub> exposure during several long-term exposure windows, and evidence that associations were stronger for preterm births occurring around week 30.

email: howard.chang@emory.edu

### Bayesian Kernel Machine Regression for Estimating the Health Effects of Pollution Mixtures

**Brent A. Coull\***, Harvard School of Public Health  
**Jennifer F. Bobb**, Harvard School of Public Health  
**Gregory A. Wellenius**, Brown University  
**Murray Mittleman**, Beth Israel Deaconess Medical Center

We consider Bayesian kernel machine regression (BKMR) with variable selection as a new approach for studying the health effects of complex mixtures. The method regresses a health outcome on a nonparametric function of a high-dimensional vector of exposure variables (e.g., elemental components of the air pollution mixture) that is specified using a kernel function. This approach flexibly estimates the health effects of exposure to the mixture in a way that accounts for potentially complex nonlinear and interactive effects, while also selecting which components of the mixture are associated with the outcome. We will discuss features of the approach particularly useful in air pollution epidemiology, such as testing for interactions among components under very general assumptions for the exposure-response relationship and the integration of evidence from parallel

toxicological studies. We evaluate the performance of BKMR under realistic conditions through simulation studies, and demonstrate the approach by applying it to investigate the relationship between short-term exposure to elemental pollution components and blood pressure in a Boston-area longitudinal cohort study.

email: bcoull@hsph.harvard.edu

### Estimating the Health Benefit of Reducing Indoor Air Pollution in a Randomized Environmental Intervention

**Roger D. Peng\***, Johns Hopkins University  
**Arlene Butz**, Johns Hopkins University  
**Amber J. Hackstadt**, Johns Hopkins University  
**D'Ann L. Williams**, Johns Hopkins University  
**Gregory B. Diette**, Johns Hopkins University  
**Patrick N. Breyse**, Johns Hopkins University  
**Elizabeth C. Matsui**, Johns Hopkins University

Recent intervention studies targeted at reducing indoor air pollution have demonstrated both the ability to improve respiratory health outcomes and to reduce particulate matter (PM) levels in the home. However, these studies generally do not address whether it is the reduction of PM levels specifically that improves respiratory health. In this paper we apply the method of principal stratification to data from a randomized air cleaner intervention designed to reduce indoor PM in homes of children with asthma. We estimate the health benefit of the intervention amongst study subjects who would experience a substantial reduction in PM in response to the intervention. For those subjects we find an increase in symptom-free days that is almost three times as large as the overall intention-to-treat effect. We also explore the presence of treatment effects amongst those subjects whose PM levels would not respond to the air cleaner. This analysis demonstrates the usefulness of principal stratification for environmental intervention trials and its potential for much broader application in this area.

email: rdpeng@gmail.com

### Influence of Time-Varying Air Pollution Exposure on Rate of Change Estimates for Progression of Cardiovascular Disease

**Lianne Sheppard\***, University of Washington  
**Adel Lee**, University of Washington

An important target parameter in air pollution cohort studies is the effect of the air pollution exposure on the rate of change of a continuous health outcome. For instance, Adar et al (2013) showed that a 1 ug/m<sup>3</sup> increase in subject-specific time-varying average PM<sub>2.5</sub> was associated with a 2.0 (1.0, 3.0) um/yr increase in carotid intima-medial thickness. The pollution exposure varies over time and this estimate does not explicitly highlight the importance of the recent temporal exposure variation. One approach is to separate the time-varying exposure into a subject's baseline

value and its deviation from baseline. Using this partitioning in a model with both terms included, Adar et al showed that the effect of a 1 ug/m<sup>3</sup> increase in PM<sub>2.5</sub> since baseline was 2.8 (1.6, 3.9) um/yr while the effect of the same increase in baseline PM<sub>2.5</sub> was 1.5 (0.5, 2.6) um/yr. We will discuss the implications of different parameterizations of air pollution rate of change in cohort studies when there are time-varying exposures and time-varying health effects.

email: sheppard@uw.edu

## 39. RECENT ADVANCES IN CAUSAL INFERENCE

### **Causal Inference with Social Network Data: Inflated Effective Sample Sizes, Deflated Standard Errors, and Other Perils**

**Elizabeth L. Ogburn\***, Johns Hopkins University

Electronic health records (EHR) provide a wealth of data that have the potential to shed light on treatment effects and effectiveness, personalized medicine, disease etiology, and many other causal questions. However, these data are quite different from traditional types of epidemiological data: the sample often comprises most of the target population; the data are rife with errors; missing data are the rule rather than the exception; and many important confounders are unavailable or available only in the aggregate. I will discuss some of the challenges and opportunities associated with EHR data and provide examples using data from a U.S. health care system that provides care and/or insurance to close to 3 million people.

email: eogburn@jhspsh.edu

### **Causal Inference with Continuous Treatments**

**Yeying Zhu\***, University of Waterloo

**Donna L. Coffman**, The Pennsylvania State University

**Debashis Ghosh**, The Pennsylvania State University

Continuous treatments are very common in practice, such as dosage uses in biomedical studies. In these studies, a main objective is to estimate the dose-response function using inverse probability weighting based on generalized propensity scores. The generalized propensity score is defined as the conditional density of the treatment given covariates. When the covariates are high-dimensional, we propose employing L2 boosting to estimate the mean function first. An important tuning parameter in boosting is the number of trees to be generated. In the case of a binary treatment, it is suggested that the optimal number of trees should be determined by minimizing the average standardized absolute mean difference between the treatment group and the control group. In the case of a continuous treatment, categorizing the treatment into several groups may cause information loss. We propose an algorithm based on distance correlation to determine the optimal number of trees. A data application to Early Dieting in Girls study is conducted to illustrate the proposed methods.

email: y239zhu@uwaterloo.ca

### **Balancing Covariates Via Propensity Score Weighting: A New Perspective**

**Fan Li\***, Duke University

**Alan Zaslavsky**, Harvard Medical School

**Kari Lock Morgan**, Duke University

Balance in the covariate distributions is crucial for an unconfounded descriptive or causal comparison between different groups. However, lack of overlap in the covariates is common in observational studies. This article focuses on weighting strategies for balancing covariates. We propose a general class of weights---the balancing weights---that balance the expectation of the covariates in the treatment and the control groups. The framework is closely related to propensity score and includes several existing weights, such as the inverse-probability weight, as special cases. In particular, we advocate a new type of weight---the overlap weight---that leads to a comparison for the population with the most overlap in the covariates between two groups. We show that the overlap weight minimizes the asymptotic variances of the weighted average treatment effect among the class of balancing weights, and also possess desirable small sample property. Simulated and real examples are presented to illustrate the method and compare with the existing approaches.

email: fli@stat.duke.edu

### **Robust Estimation of Causal Effects of Erythropoiesis-Stimulating Agents (ESAs) on Mortality**

**Roe Gutman\***, Brown University

**David D. Dore**, Brown University

Anemia is a frequent complication of chronic kidney disease (CKD), and if unmanaged can have severe consequences. ESAs are frequently employed in the management of CKD-induced anemia. Recent clinical trials showed that treating patients with high hemoglobin levels with ESAs increased the risk of death. We obtained data from the Renal Management Information System (REMIS) on all end-stage renal disease patients undergoing hemodialysis between 2007-2011. This dataset includes patients' information on specific dialysis encounters. This dataset suffers from limitations common to other observational studies that attempt to estimate causal effects; e.g. missing covariates, definition of estimands, non-collapsibility and confounded assignment mechanism. Through the analysis of the REMIS dataset we introduce a methodology that handles these limitations and enables efficient estimation of the effect of ESAs. This methodology relies on two-staged multiple imputation. The first stage imputes the missing covariates and the second stage imputes the missing potential outcomes. This methodology also allows for estimation of sub-group effects, thus providing interesting scientific insights.

email: roee\_gutman@brown.edu

## 40. SOCIAL NETWORK DATA: CHALLENGES AND OPPORTUNITIES

### What, if Anything, Do We Learn by Fitting an Exponential-Family Random Graph Model?

**Cosma Shalizi\***, Carnegie Mellon University  
**Alessandro Rinaldo**, Carnegie Mellon University

Typically, statistical models of network structure are models for the whole network, while the data is only a sampled sub-graph. Parameters for the whole network, which are the targets of inference, are estimated by applying the model to the sub-network. This assumes that the model is consistent under sampling, or, in terms of the theory of stochastic processes, that it forms a projective family. For the popular class of exponential random graph models (ERGMs), we show that this apparently trivial condition is actually violated by many popular and scientifically appealing models, and that satisfying it drastically limits ERGM's expressive power. (These results are special cases of new general theorems about dependent exponential families.) Using such results, we offer easily checked conditions for the consistency of maximum likelihood estimation in ERGMs, and discuss some possible constructive responses.

email: cshalizi@cmu.edu

### Bayesian Inference for Non-Ignorable Sampling in Social Networks

**Simon Lunagomez\***, Harvard University  
**Edoardo M. Airoldi**, Harvard University

Consider individuals interacting in social network and a response that can be measured on each individual. We are interested in making inferences on a population quantity that is a function of both the response and the social interactions. In this paper, working within Rubin's inferential framework, we propose a Bayesian model that takes into account all relevant sources of uncertainty. Inferences are performed via Bayesian model averaging. Our method provides valid inferences in applications to epidemiology and healthcare in which hard-to-reach populations are sampled using link-tracing designs, including respondent-driven sampling, that carry information about the quantify of interest.

email: simon.lgz@gmail.com

### Targeted Learning of Causal Effects for Networks

**Mark J. van der Laan\***, University of California, Berkeley

Suppose that we observe a population of causally connected units over time where the causal interdependence between the unit-specific longitudinal data structures is defined by a structural equation model. On each unit, we observe a unit-specific longitudinal data structure consisting of baseline and time-dependent covariates, a time-dependent treatment/exposure, and a final outcome of interest. Moreover, at each time-point, and for each unit, this unit-specific longitudinal data structure also includes observing a set of friends of that unit whose observed past data may potentially affect the data for that unit at the next time point. Given this structural equation model, the target quantity

of interest is defined as the mean counterfactual outcome for this group of units if, contrary to the fact, the exposures of the units would be probabilistically assigned according to a known specified mechanism, where the latter is called a stochastic intervention on the unit-specific multiple time-point exposures. This causal model includes causal interference as special case. We develop targeted minimum loss-based estimators and statistical inference, incorporating weak convergence theory for dependent data.

email: laan@berkeley.edu

### Diffusion Matters, But How?

**Kevin A. Bryan\***, Northwestern University

Many phenomena, such as diseases, novel inventions, public policies, or cultural practices are characterized by diffusion through networks. Often, theory suggests that a set of potential networks may simultaneously play a role in diffusion. For example, information about health interventions may spread at differential rates among friends versus colleagues, or adoption of a new public good may spread across cities via both geographic proximity and industrial similarity. If the goal is to probabilistically decompose diffusion into the parts caused by each type of network, I show that regression parameters alone cannot do this, that a set of structural assumptions allows estimation of the correct parameters, and that tools from graph theory make such estimation feasible. I propose to apply this method to the spread of population control measures in 1970s China. These results are relevant to the study of the diffusion of infectious diseases and health behaviors through social networks and to the study of the diffusion of health policies through geopolitical networks.

email: kevincure@gmail.com

## 41. STATISTICS AND COMPUTING FOR HIGH-THROUGHPUT SEQUENCING DATA

### Computational Challenges in Exome and RNA-Seq Analysis

**Steven L. Salzberg\***, Johns Hopkins University

Next-generation sequencing technology allows us to peer inside the cell in exquisite detail, revealing new insights into biology, evolution, and disease that would have been impossible to find just a few years ago. The enormous volumes of data produced by NGS experiments present many computational challenges that we are working to address. In this talk, I will discuss some of the latest solutions to two basic alignment problems: (1) mapping sequences onto the human genome at very high speed, and (2) mapping and assembling transcripts from RNA-seq experiments. I will also discuss some of the problems that can arise during analysis of exome data, in which the gene-containing portions of the genome are sequenced in an effort to identify mutations responsible for disease. My group has developed algorithms to solve each of these problems, including the widely-used Bowtie program for fast alignment and the TopHat and Cufflinks programs for assembly and quantification of genes in transcriptome sequencing (RNA-

seq) experiments. This talk describes joint work with current and former lab members including Ben Langmead, Cole Trapnell, Daehwan Kim, Mihaela Pertea, and Geo Pertea; and with collaborators including Mihai Pop and Lior Pachter.

email: salzberg@jhu.edu

### **Statistical Modeling of Alternative Splicing with RNA-Seq Data**

**Hui Jiang\***, University of Michigan

**Julia Salzman**, Stanford University

**Yang Shi**, University of Michigan

Alternative splicing is a cell mechanism which is of vital importance to gene regulation and expression in higher eukaryotes and can have significant impact on many diseases. Ultra high-throughput sequencing of transcriptomes (RNA-Seq) has enabled the accurate estimation of gene expression at individual isoform level. In this talk I will present some recent works on modeling alternative splicing with RNA-Seq data, which include methods for robust estimation of isoform expression and for detecting differential isoform usage, as well as computational techniques for efficient estimation and testing. These approaches will be demonstrated on both simulated and real data. This talk is based on joint works with Yang Shi and Julia Salzman.

email: jianghui@umich.edu

### **Statistical Analysis of Deep Sequencing Data from Tumor Samples**

**Lin Hou**, Yale School of Public Health

**Mengjie Chen**, Yale University

**Hongyu Zhao\***, Yale School of Public Health

Many cancer sequencing projects have led to many novel discoveries and insights on cancer onset and progression. Although highly informative, the massive data from these sequencing studies present great statistical and computational challenges. In this presentation, we will discuss the unique issues arising from deep sequencing of tumor samples, and statistical methods to address these them.

email: hongyu.zhao@yale.edu

### **Models And Statistics for Detection of Genome Structural Variation**

**Nancy R. Zhang\***, University of Pennsylvania

**David Siegmund**, Stanford University

**Benjamin Yakir**, The Hebrew University

Structural variation, which includes deletion, insertion, and inversion of stretches of DNA, comprise an important class of genome variation in the human population, and are implicated in many diseases. High throughput paired end short read sequencing allows for genome-wide detection of a wide spectrum of structural variation. We develop a general model for this data, based on a Poisson random field, under which signals that are characteristic for each type of structural change can be modeled using a likelihood based framework. Statistics derived from the model integrate information from coverage, insert length, and other aspects

of the data, and thus has improved sensitivity over methods that only utilize any single feature. We also describe how to control the false discovery rate for scan statistics of Poisson random fields, and illustrate our methods on 1000 genomes data. This is joint work with David Siegmund and Benjamin Yakir.

email: nzh@wharton.upenn.edu

## **42. VARIABLE SELECTION AND ANALYSIS OF HIGH DIMENSIONAL DATA**

### **Regularized Semiparametric Functional Linear Regression**

**Helen Zhang\***, University of Arizona

In modern scientific experiments, we often face analysis with functional data, where the observations are sampled from random process, together with a potentially large number of non-functional covariates. The complex nature of functional data makes it difficult to directly apply existing methods to model selection and estimation. We propose and study a new class of semiparametric functional linear models to jointly model functional and non-functional predictors, identify important covariates, and improve efficiency and interpretability of the estimates. Featured with two types of regularization: the shrinkage on effects of scalar covariates and the truncation on principal components of the functional predictor, the new methods treats functional predictor from a nonparametric perspective and focuses on inferring the parametric structure of the scalar covariates. We establish consistency and oracle properties under mild conditions by allowing possibly diverging number of scalar covariates and simultaneously taking the functional predictor into account. An application to fMRI brain image data is shown.

email: hzhang@math.arizona.edu

### **Dimension Reduction for Tensor Regression**

**Peng Zeng\***, Auburn University

**Wenxuan Zhong**, University of Georgia

Tensor is a multiway array. With the rapid development of science and technology in the past decades, large amount of tensor observations are routinely collected, processed and stored in many scientific researches and commercial activities nowadays. To address the statistical challenges that are raised in analyzing the tensor data, in this article, we proposed tensor dimension reduction regression (TDR) models that assume that the response depends on a projection of the tensor predictors through some unknown link function. A novel sequential model estimation approach, called SIDRA, has been proposed under the TDR framework. We also derived the theoretical underpinning of SIDRA approach. The TDR framework has been successfully integrated into the optical electronic nose technique, called colorimetric sensor array (CSA), to improve its sensitivity and specificity. Preliminary studies demonstrate that SIDRA can greatly improve the prediction accuracy and are highly

computationally efficient which provide a high possibility to a prototyping of a portable device for the assessment of personal exposure to pollutions using CSA technique.

email: zengpen@auburn.edu

### **Sparse Group LASSO for Pathway Based GWAS**

**Tatiana V. Apanasovich\***, The George Washington University

Sparse modeling approaches are becoming increasingly popular for the analysis of genome wide datasets. Gene pathway analysis utilizes information about the functional structure of genome and pathways associated with the disease are identified first. Then, significant genes or SNPs are selected within the associated pathways. In our talk we describe a multilevel sparse logistic regression model for the simultaneous identification of pathways, genes and SNPs associated with a risk of complex disorder. Our methodology uses group LASSO to enforce sparsity at the pathway, gene and SNP levels. We discuss how to handle the challenges common to joint modeling of genome-wide data: association between genetic factors and overlapping pathways. Finally, the resampling method that exploits finite sample variability is used to rank pathways, SNPs and genes within pathways. We validate our method through simulations, and use it to perform the pathways-driven SNP selection in a search for pathways, genes and SNPs associated with alcoholism.

email: apanasovich@gwu.edu

## **43. FUNCTIONAL DATA ANALYSIS AND ITS APPLICATIONS IN GENETICS MOST PREDICTIVE INTERVAL SELECTION FOR FUNCTIONAL PREDICTORS, WITH APPLICATION TO CLASSIFYING TUMOR STAGES FROM MASS SPECTRA**

**Andreas Kryger Jensen**, University of Southern Denmark, Odense

**Hans-Georg Müller\***, University of California, Davis

We discuss a problem that arises in clinical proteomics: To classify tumors (benign or malign) from mass spectra and to identify biomarkers. Working within the framework of functional data analysis, we aim to construct a classifier that requires to obtain the mass spectra only on a subinterval, rather than on the entire domain. This reduction in the domain is important, as biomarker identification from mass spectra is a laborious and slow process. The process of obtaining a small set of protein biomarkers to be used in routine screening procedures is laborious and slow but can be facilitated by reducing the entire spectrum to a spectrum on a smaller interval. We study how the location of such a subinterval can be selected, aiming to minimize a target criterion. This approach is implemented by combining functional classifiers that are based on generalized functional linear models with a peaklet expansion for the spectra. The method is illustrated with clinical MALDI-TOF (matrix-assisted laser desorption/ionization time-of-flight) mass spectrometry data.

email: hgmueller@ucdavis.edu

### **Restricted Likelihood Ratio Tests for Functional Effects in the Functional Linear Model**

**Bruce J. Swihart\***, Johns Hopkins

Bloomberg School of Public Health

**Jeff Goldsmith**, Columbia University

**Ciprian M. Crainiceanu**, Johns Hopkins

Bloomberg School of Public Health

The goal of our article is to provide a transparent, robust, and computationally feasible statistical approach for testing in the context of scalar-on-function linear regression models. In particular, we are interested in testing for the necessity of functional effects against standard linear models. Our methods are motivated by and applied to a large longitudinal study involving diffusion tensor imaging of intracranial white matter tracts in a susceptible cohort. In the context of this study, we conduct hypothesis tests that are motivated by anatomical knowledge and which support recent findings regarding the relationship between cognitive impairment and white matter demyelination. R-code and data are provided to reproduce the application.

email: bruce.swihart@gmail.com

### **Gene-Gene Interaction Analysis for Next-Generation Sequencing**

**Momiao Xiong\***, University of Texas

School of Public Health

**Yun Zhu**, Tulane University

**Jinying Zhao**, Tulane University

The critical barrier in interaction analysis for next-generation sequencing (NGS) data is that the traditional pair-wise interaction analysis is difficult to apply to rare variants because of their prohibitive computational time, large number of tests and low power. The great challenges for successful detection of interactions with NGS data are (1) the demands in the paradigm of changes in interaction analysis, (2) severe multiple testing, and (3) heavy computations. To meet these challenges, we shift the paradigm of interaction analysis between two SNPs to interaction analysis between two genomic regions. By intensive simulations, we demonstrate that the functional logistic regression for interaction analysis has the correct type 1 error rates and much higher power to detect interaction than the currently used methods. The proposed method was applied to a coronary artery disease dataset from the Wellcome Trust Case Control Consortium (WTCCC) study and the Framingham Heart Study (FHS) dataset, and the early-onset myocardial infarction (EOMI) exome sequence datasets with European origin from the NHLBI's Exome Sequencing Project. We discovered that six of 27 pairs of significantly interacted genes in the FHS were replicated in the independent WTCCC study and 24 pairs of significantly interacted genes in the EOMI study.

email: momiao.xiong@uth.tmc.edu

### **Functional Regression Models for Association Analysis of Complex Traits**

**Ruzong Fan\***, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**Yifan Wang**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**James L. Mills**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**Alexander F. Wilson**, National Human Genome Research Institute, National Institutes of Health

**Joan E. Bailey-Wilson**, National Human Genome Research Institute, National Institutes of Health

**Momiao Xiong**, University of Texas, Houston

Functional regression models are developed for testing associations between complex traits and multiple genetic variants in a genetic region. Although the observed genetic marker data are discrete, we view them as realizations of continuous genetic variant functions. We believe that the genetic variant functions have intrinsic functional structure. By using modern state-of-the-art functional data analysis technique, the observed high dimension genetic variant data are used to estimate the genetic variant functions based on B-spline or Fourier basis functions or functional principal component decompositions. Then, the estimated genetic variant functions are used in the regression models to connect to phenotype adjusting for covariates. Both fixed and mixed effect functional generalized linear models are built to test the association between dichotomous/quantitative traits and genetic variants adjusting for covariates. After extensive simulation analysis, it is shown that the proposed test methods generate accurate type I errors and have good power. The methods can be used in both gene-based genome-wide/exome-wide association studies and candidate gene analysis.

email: fanr@mail.nih.gov

## **44. EMERGING STATISTICAL CHALLENGES WITH COMPLEX LONGITUDINAL OR FUNCTIONAL DATA**

### **Consistent Estimation of Covariate Effects for Some between-/within-Cluster Covariate Decomposition Methods when Data are Missing at Random**

**John Neuhaus\***, University of California, San Francisco

**Charles McCulloch**, University of California, San Francisco

Conditional maximum likelihood methods and closely related maximum likelihood methods that partition covariates into between- and within-cluster components are useful methods for analyzing longitudinal data. However, both of these methods can produce inconsistent covariate effect estimates when data are missing at random (MAR). Using theory and simulation studies, this talk shows

that decomposition methods using complete covariate information produce consistent estimates. In some practical cases these methods, that ostensibly require complete covariate information, actually only involve the observed covariates. These results offer an easy-to-use approach to simultaneously protect against bias from either cluster-level confounding or MAR missingness.

email: john@biostat.ucsf.edu

### **Handling Missing Data for Multiple Waves of Longitudinal Data**

**Xuan Bi**, University of Illinois, Urbana-Champaign

**Annie Qu\***, University of Illinois, Urbana-Champaign

In many instances of survey data, it is common to collect samples through multiple waves of longitudinal data, which is also called refreshment sampling. A subject is likely to be followed at one time or for a certain period, but not through the entire study period. We will investigate the inference function approach to test whether refreshment samplings cause the biased estimation, and whether the missing data is informative. We will compare several methods based on weighting, imputation and adjusted sampling to correct bias for multiple waves of longitudinal data.

email: anniequ@illinois.edu

### **A Broad Framework for Joint Modeling and Some Tales from the Unexpected**

**Geert Molenberghs\***, I-BioStat, Hasselt Universiteit and Katholieke Universiteit Leuven, Belgium

**Michael G. Kenward**, London School of Hygiene and Tropical Medicine, UK

**Marc Aerts**, Hasselt Universiteit, Belgium

**Geert Verbeke**, Katholieke Universiteit Leuven and Hasselt Universiteit, Belgium

**Anastasios Tsiatis**, North Carolina State University

**Marie Davidian**, North Carolina State University

**Dimitris Rizopoulos**, Erasmus University, The Netherlands

Many statistical properties are derived for a fixed, a priori known sample size. Familiar results then follow, such as the consistency, asymptotic normality, and efficiency of the sample average for the mean parameter when the data follow either a normal or a member of a large class of non-normal distributions. Matters change when sample size itself becomes a random. This can take various forms: the sample size can depend on the data collected or not and, whenever it does, it can be governed by either a deterministic or a probabilistic rule. This include sequential trials. It is insightful to place this into a general joint-modeling-based framework and derive generic results. From there, both parametric and semi-parametric inferences can be drawn. It will be shown that counterintuitive results may follow: the sample average may exhibit small-sample biased in many cases and, even when it is unbiased, like with a completely random sample size, then it is not optimal, without a uniform optimum

existing. This depends critically depend on key attributes, such as (non-)ancillarity of the sample size and the fact that the sample sum combined with the sample size never is a so-called complete sufficient statistic, as long as at least two different sample sizes have a non-zero probability of occurring. Our results have direct implications for estimation after group sequential trials.

email: geert.molenberghs@uhasselt.be

### **Modeling and Estimation Methods for Physical Activity Data**

**Haocheng Li**, Texas A&M University

**Raymond J. Carroll\***, Texas A&M University

**John Staudenmayer**, University of Massachusetts, Amherst

We study physical activity functional data, which is complex because daily records from Monday to Friday are collected within each week, while observations in weeks are nested within subjects. A general framework of functional mixed effects model is proposed for such data including modeling the influence of two factors: weeks and days. We allow for week-specific, day-specific and week-day-interact effects in the mean structures, while subject-specific, week-specific, day-specific and week-day-interact variations are modeled by random structures. The mean and random structures are specified as smooth curves measured at various time-points. The association structure of the three-level data is postulated through the random curves, which are summarized using a few important principal components. We use penalized splines to model the mean curves and the principal component curves, and cast the proposed model into a mixed effects model framework for model fitting, prediction and inference. The method is applied to physical activity data, and is evaluated empirically by a simulation study.

email: carroll@stat.tamu.edu

## **45. GENOME WIDE ASSOCIATION STUDIES**

### **Testing Calibration of Risk Models at Extremes of Disease-Risk**

**Minsun Song\***, National Cancer Institute, National Institutes of Health

**Peter Kraft**, Harvard School of Public Health

**Amit D. Joshi**, Harvard School of Public Health

**Myrto Barrdahl**, German Cancer Research Center

**Nilanjan Chatterjee**, National Cancer Institute, National Institutes of Health

Risk-prediction models need careful calibration to ensure they produce unbiased estimates of risk for subjects in the underlying population given their risk-factor profiles. As subjects with extreme high- or low- risk may be the most affected by knowledge of their risk estimates, checking adequacy of risk models at the extremes of risk is very important for clinical applications. We propose a new approach to test model calibration targeted toward extremes of disease risk distribution where standard goodness-of-fit tests may lack power due to sparseness of data. We construct a test statistic based on model residuals summed over only

those individuals who pass high- and/or low risk-thresholds and then maximize the test-statistic over different risk-thresholds. We derive an asymptotic distribution for the max-test statistic based on analytic derivation of the variance-covariance function of the underlying Gaussian process. The method is applied to a large case-control study of breast cancer to examine joint effects of common SNPs discovered through recent genome-wide association studies. The analysis clearly indicates non-additive effect of the SNPs on the scale of absolute risk, but an excellent fit for the linear-logistic model even at the extremes of risks.

email: songm4@mail.nih.gov

### **An Adaptive Genetic Association Test Using Double Kernel Machines**

**Xiang Zhan\***, The Pennsylvania State University

**Debashis Ghosh**, The Pennsylvania State University

Recently, gene set-based approaches have become very popular in gene expression profiling studies for assessing how genetic variants are related to disease outcome. Since most genes are not differentially expressed, existing pathway tests considering all genes within a genetic pathway suffer from considerable noise and power loss. Moreover, for a differently expressed pathway, it is of interest to select important genes that drive the effect of the pathway. In this article, we propose an adaptive association test using double kernel machines (DKM), which can both select important genes within the pathway as well as test for the overall genetic pathway effect. This DKM procedure first uses the garrote kernel machines (GKM) test for the purposes of subset selection and then the least squares kernel machine (LSKM) test for testing the effect of the subset of genes. An appealing feature of the kernel machine framework is that it can provide a flexible and unified method for multi-dimensional modeling of the genetic pathway effect allowing for both parametric and nonparametric components. This DKM approach is illustrated with application to simulated data as well as data from a neuroimaging genetics study.

email: xyz5074@psu.edu

### **Multi-Marker Tests for Joint Association in Longitudinal Studies Using the Genetic Random Field Model**

**Zihuai He\***, University of Michigan

**Min Zhang**, University of Michigan

**Jennifer Smith**, University of Michigan

**Sharon Kardia**, University of Michigan

**Ana Diez Roux**, University of Michigan

**Seunggeun Lee**, University of Michigan

**Xiuqing Guo**, Columbia University

**Walter Palmas**, Columbia University

**Bhramar Mukherjee**, University of Michigan

Longitudinal studies of common and chronic diseases risk factors provide a valuable opportunity to explore how genetic variants affect traits over time. Statistical power to detect disease susceptibility variants can be improved if we jointly utilize the entire set of longitudinal outcomes. Since

disease risk factors and phenotypes are likely influenced by the joint effect of multiple variants in a gene, a joint analysis of these variants may help to explain additional heritability. In this article, we propose a longitudinal genetic random field model (LGRF), to test the joint association between a set of genetic variants and a phenotype measured repeatedly during the course of an observational study. Several essential methodological improvements are further proposed to enhance the robustness to misspecification of within-subject correlation structure and to improve computational efficiency. The proposed methods were evaluated through simulation studies and illustrated using data from the Multi-Ethnic Study of Atherosclerosis (MESA). Our simulation results indicate substantial gain in power using LGRF when compared to the two commonly used existing alternatives: (i) single marker tests using longitudinal outcome and (ii) existing multi-marker association tests such as the sequence kernel association tests (SKAT) using the average value of repeated measurements as the outcome.

email: zihuai@umich.edu

### **More Powerful Genetic Association Testing Via a New Statistical Framework for Integrative Genomics**

**Sihai D. Zhao\***, University of Pennsylvania

**Tony Cai**, University of Pennsylvania

**Hongzhe Li**, University of Pennsylvania

Integrative genomics offers a promising approach to more powerful genetic association studies. The hope is that combining outcome and genotype data with other types of genomic information can lead to more powerful SNP detection. We present a new statistical model for the relationship between outcome, gene expression data, and genotype data that leads to a more powerful test for genetic association, compared to tests using the outcome and genotype data alone. The model explicitly assumes that genetic variations affect the outcome through perturbing the gene expression levels of a set of genes, and is relatively robust to misspecification of the regulated genes. We use the method on data from patients with advanced heart failure to identify new candidate SNPs that perturb a new candidate gene.

email: dave.zhao@gmail.com

### **Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements?**

**Yiwei Zhang\***, Novartis Pharmaceuticals

**Wei Pan**, University of Minnesota

Genome-wide association studies (GWAS) may suffer from false positives and false negatives due to confounding population structures. Another important issue is unmeasured environmental risk factors. Among many methods for adjusting for population structures, two approaches stand out: principal component regression (PCR) based on principal component analysis (PCA) and linear mixed model (LMM) based method. In this paper, based on the formulation of probabilistic PCA, we show that the PCR approach can be regarded as an approximation to a LMM; such an approximation depends on the number of the top

principal components (PCs) used. Hence, in the presence of population structure, especially for complex structure, the LMM appears to outperform the PCR method. However, due to the different treatments of fixed versus random effects in the two approaches, we show an advantage of PCR over LMM: in the presence of an unknown but spatially confined environmental confounder, the PCs may be able to implicitly and effectively adjust for the confounder while the LMM cannot. Accordingly, to adjust for both population structures and non-genetic confounders, we propose a hybrid method combining the strengths of PCR and LMM. We use a real genotype dataset to confirm the above points, and establish the superior performance of the hybrid method across all scenarios.

email: iid.poisson@gmail.com

### **A Versatile Omnibus Test for Detecting Mean and Variance Heterogeneity for Quantitative Traits**

**Peng Wei\***, University of Texas School of Public Health

**Ying Cao**, University of Texas School of Public Health

**Taylor Maxwell**, University of Texas

School of Public Health

Recent research has revealed genetic loci that display variance heterogeneity through various means such as biological disruption, linkage disequilibrium (LD), gene-by-gene (GxG), or gene-by-environment (GxE) interaction. We propose a versatile likelihood ratio test that allows joint testing for mean and variance heterogeneity or either effect alone in the presence of covariates. Using analytical derivation, simulations and empirical data from a known mean-only functional variant we demonstrate how LD can induce variance-heterogeneity loci (vQTL) in a predictable fashion. We propose that a joint test for mean and variance heterogeneity is more powerful than a variance only test for detecting vQTL. This takes advantage of loci that also have mean effects without sacrificing much power to detect variance only effects.

email: Peng.Wei@uth.tmc.edu

### **Flexible and Robust Methods for Rare-Variant Testing of Quantitative Traits in Pedigrees**

**Yunxuan Jiang\***, Emory University

**Karen N. Conneely**, Emory University

**Michael P. Epstein**, Emory University

Rare-variant sequencing studies are increasingly popular strategies for investigating the missing heritability of complex human traits. Few statistical methods have been developed to analyze rare variants in family-based studies. Family-based designs can overcome potential bias caused by population stratification and can study cosegregation patterns of causal variants. We propose a rare-variant association test for quantitative traits in families that uses a kernel framework. Within a region of interest, the model partitions a family member's genotype at a rare variant into a within-family component and an orthogonal between-family component. Our approach first constructs a kernel test using the between-family component as a screening tool to

identify top hits then follow up these top hits using a kernel test based on the robust and independent within-family component. Our method is flexible in that it still performs well under missingness of parental genotypes. Finally, our method has the practical benefit of efficiently calculating p-values based on asymptotics, which enables practical application to genome-wide data. Using simulated data, we have already shown that our method can avoid inflated false positive caused by population stratification and can improve the power with screening. We will further apply our method to the Sardinia sequencing dataset for illustration.

email: yjian27@emory.edu

## 46. APPLICATIONS OF BAYESIAN METHODS

### **Semi-Parametric Bayesian Clustering of Ophthalmology Data**

**Xin Tong\***, University of South Carolina  
**Hongmei Zhang**, University of Memphis

Zernike aberration polynomials have been commonly used as the standard method of describing the shape of an aberrated wavefront of the human eye. Knowledge on the homogeneity among the Zernike coefficients can potentially help us to improve eye disease diagnosis. Clustering methods are designed to separate heterogeneous data into groups of similar objects such that objects within a group are similar. We propose a new clustering method to classify eye patients. Specifically, subjects are clustered based on the agreement of relationships between variable measures and covariates of interest. A Bayesian method is proposed for this purpose, in which a semi-parametric model is used to evaluate any unknown relationship between variables and covariates of interest, and a Dirichlet process is utilized in the process of subjects clustering. The major novelty of the method exists in its ability to produce homogeneous clusters composed of a certain number of subjects sharing common features on the relationship between response variables and covariates. We test the performance of our procedures using simulations and apply it to Zernike coefficients of pre-LASIK operation patients.

email: cuzntone@gmail.com

### **A Nonparametric Bayesian Latent Factor Model for Body Image Evaluation**

**Kassie Fronczyk\***, Rice University  
**Michele Guindani**, University of Texas  
MD Anderson Cancer Center  
**Marina Vannucci**, Rice University

Body image is recognized as a critical psychosocial issue for individuals with facial cancer, as the disease and its treatment can have devastating consequences involving disfigurement and functional impairment. Healthcare providers would benefit greatly from being able to predict which patients are at greatest risk for experiencing psychosocial impairment arising from body image disturbance. A compilation of measures in the form of questions targeting a patient's body image and distress levels on an ordinal scale are collected

from a number of cancer patients. We develop a multinomial probit factor analysis model to enable data-driven decision making for body image evaluation based on the recorded answers. Under a fully Bayesian setting, we obtain inference for the unknown number of underlying factors affecting body image by employing a nonparametric prior. Our model not only provides information regarding the connection between both questions and patients and the factors of underlying body image and distress, but also presents a possible method for paring down the list of questions by identifying redundant or superfluous measures.

email: kf8@rice.edu

### **Bayes Sensitivity Analysis with Fisher-Rao Metric**

**Sebastian Kurtsek**, The Ohio State University  
**Karthik Bharath\***, The Ohio State University

We propose a geometric framework to assess sensitivity of Bayesian procedures to modeling assumptions based on the nonparametric Fisher-Rao metric on the non-linear manifold of probability densities. While the framework is general in spirit, the focus of this article is restricted to metric-based diagnosis under two settings: assessing local and global robustness in Bayesian procedures to perturbations of the prior; identification of influential observations in a Bayesian regression setting with perturbations to the data. The approach is based on the square-root representation of densities which enables us to compute geodesics and geodesic distances in analytical form, facilitating the definition of naturally calibrated discrepancy measures. The approach proposed in this article ought to be viewed as a first step in the investigation of the exploitation of the geometric structure of the space of probability densities in defining metric-based measures of discrepancies under a Bayesian setting.

email: karthikbharath@gmail.com

### **Bayesian Inference on Multiple Proportions for Misclassified Binomial Data**

**Dewi Rahardja**, U.S. Food and Drug Administration  
**Haiwen Shi\***, U.S. Food and Drug Administration

We consider misclassified binary data composed of an original data set and an independent training data set. The original data set is a multiple-sample binary data set where the binary variable is obtained using a fallible classifier. The training data set has not only the same structure as the original data set but also additional classification of the binary variable using an infallible classifier. The research objectives are to estimate and to test if differences exist among multiple proportion parameters associated with the response variable obtained by the infallible classifier. In this research we achieve the research objectives via Bayesian inference. We developed a Bayesian algorithm that is easy to implement and performs well under various simulation scenarios.

email: rahardja@gmail.com

### **Longitudinal Mediation Analysis**

**Chanmin Kim\***, University of Texas, Austin  
**Michael J. Daniels**, University of Texas, Austin  
**Jason A. Roy**, University of Pennsylvania  
**Beth H. Marcus**, University of California, San Diego

Commit-to-Quit (CTQ) II study was a randomized controlled trial to study the effect of moderate-intensity exercise on smoking cessation in women. During each of the 8 weeks of the intervention period, subjects were supposed to attend either one supervised moderate-intensity exercise session on site and two on their own (the intervention arm) or one lecture on site with supplemental handouts (the control arm). It is of particular interest to understand how much of any potential treatment effect is mediated by weight change (i.e., how large is the indirect effect and to what degree does this mediating effect change over the course of the study?). We propose a Bayesian approach to estimate time-varying natural direct and indirect effects in the above setting of longitudinal mediators and responses. A Bayesian updating model is proposed to model the observed data over time. Non-parametric methods are used to minimize modeling assumption. Several conditional independence assumptions (with sensitivity parameters) are introduced to identify causal effects at each time.

email: chanminkim@utexas.edu

### **Estimation of Contact Network Properties Using Multiple HIV Epidemic Data Sources**

**Ravi Goyal\***, Harvard University  
**Nicole B. Carnegie**, Harvard University

The properties of contact networks can have profound effects on both the spread of diseases and the effectiveness of control programs. Therefore, evaluating the relative merits of different HIV prevention and policy options requires precise information about these properties. We present a method to integrate data from multiple sources, i.e. clinic data, behavioral surveys, and genetic sequences of pathogens, to estimate high-order properties of sexual contact networks that are critical in evaluating both the drivers of the HIV epidemic and efficacy of HIV prevention programs. We jointly model the contact and transmission networks and epidemic process using a Bayesian model, incorporating the network model of Goyal et al. 2013 and an SIIR epidemic model, which captures acute-phase dynamics important to the transmission of HIV, in order to estimate critical network properties, such as clustering, degree distribution, and degree mixing. We will discuss ways to assess the relative contributions of each data source to the precision of the resulting estimates of contact network properties as well as possible extensions to dynamic networks by allowing for the inclusion of information on migration and timing of relationships in order to capture the evolving process of partner change and mobility.

email: rag524@mail.harvard.edu

### **Bayesian Variable Selection for a Regression Model with a Misclassified Binary Covariate**

**Daniel P. Beavers\***, Wake Forest School of Medicine  
**James D. Stamey**, Baylor University

Selecting an optimal set of covariates in the presence of misclassified data is generally beyond the scope of most variable selection procedures. While traditional variable selection methods attempt to identify the set of covariates to optimally predict a single outcome, models containing misclassified data are often exceedingly complex for most selection procedures due to the presence of latent or partially observed true values, the need for optimal prediction of the latent values, and the need for optimal determination of the sensitivity and specificity of the misclassified variable. We propose the use of Gibbs variable selection to identify a parsimonious set of covariates simultaneously spanning the primary model of interest, the measurement of the latent or partially observed gold standard covariate value, the sensitivity of the fallible classifier, and the specificity of the fallible classifier. We justify our approach through simulations, and finally we demonstrate the method using data from a validated questionnaire.

email: dbeavers@wakehealth.edu

## **47. HIGH DIMENSIONAL DATA**

### **Inference for Survival Prediction in the High Dimensional Setting**

**Jennifer A. Sinnott\***, Harvard University  
**Tianxi Cai**, Harvard University

For a risk prediction model to provide clinical utility, it is crucial that it deliver both a prediction for a new patient's risk and an honest assessment of the error inherent in that prediction. When the number of predictors is small, classical methods can capture prediction error; however, when the predictors are high dimensional, estimation of the prediction error can be challenging. In this high dimensional setting, we investigate inference on the survival function estimated using a model, such as the Cox model, under shrinkage. A shrinkage method can be chosen to give nice theoretical properties including asymptotic normality and asymptotically perfect variable selection; nevertheless, in finite samples, the estimated conditional survival distribution can be difficult to approximate using either asymptotic results, which can underestimate the variability, or the standard bootstrap, which may yield overly-conservative standard error estimates. We propose an adaptation of perturbation resampling designed to improve estimation of the error in survival prediction. We demonstrate our method in a study relating a large panel of tissue biomarkers to prostate cancer progression.

email: jsinnott@hsph.harvard.edu

### Testing High-Dimensional Nonparametric Function with Application to Gene Set Analysis

**Tao He\***, Michigan State University  
**Ping-Shou Zhong**, Michigan State University  
**Yuehua Cui**, Michigan State University  
**Vidyadhar Mandrekar**, Michigan State University

This paper proposes a test statistic for testing the high-dimensional nonparametric function in a reproducing kernel Hilbert space generated by a positive definite kernel. We studied the asymptotic distribution of the test statistic under the null hypothesis and a series of local alternative hypotheses in a “large  $p$  small  $n$ ” setup. A simulation study was used to evaluate the finite sample performance of the proposed method. We applied the proposed method to a yeast data set to identify pathways that are associated with BAT2 gene expression.

email: hetao@stt.msu.edu

### Variable Selection and Inference for Ultra-High Dimensional Survival Data with Missing Covariates Under Proportional Hazards Models

**Yang Ning\***, University of Waterloo  
**Grace Yi**, University of Waterloo  
**Baojiang Chen**, University of Nebraska  
**Nancy Reid**, University of Toronto

Proportional hazards models have been perhaps the most popular models used for survival data analysis. Such models, however, break down for settings with high dimensional covariates. Although there is work on model selection for proportional hazards models with high dimensional covariates, most existing methods cannot handle data with ultra-high dimension. Furthermore, in the presence of missing observations, standard inference procedures based on proportional hazards models fail to produce consistent results. In this paper, we address this problem simultaneous inferential procedures that handle both model selection and parameter estimation for ultra-high dimensional survival data with missing covariates. Our methods are developed for a broad class of folded concave penalties, including the LASSO and SCAD penalties to conduct variable selection. Missing data processes are featured by regression models where selection of important covariates can be a concern. To improve efficiency, augmented model selection and parameter estimation algorithms are exploited. The strong oracle property and the asymptotic distributions of our proposed estimators are rigorously established. The performance of the proposed methods are numerically assessed through simulation studies, and the usage of our methods is illustrated by a genetic data set.

email: yning@jhsph.edu

### An EM Test for the Contaminated Chi-Square Model

**Feng Zhou\***, University of Kentucky  
**Hongying Dai**, Children's Mercy Hospital  
**Richard Charnigo**, University of Kentucky

Likelihood-based methods play a central role in parametric testing problems, and among these the likelihood ratio test (LRT) is often preferred. Under standard regularity conditions, the LRT statistic has a simple and elegant asymptotic chi-square distribution under the null hypothesis. But most of the standard asymptotic results for the LRT cannot be applied to mixture-type models. This motivates us to develop an EM test, which is also a likelihood-based method and is implemented with the aid of the Expectation-Maximization (EM) algorithm, for the contaminated chi-square model. The EM test, originally developed for other models by Li, Chen, and Marriott, is appealing in that it enjoys an elegant asymptotic theory compared with the LRT. The EM test also has better power than a moment-based test. Furthermore, the EM test requires no more than two to three iterations of the EM algorithm, which is more time efficient than the modified likelihood ratio test (MLRT).

email: fzh223@gmail.com

### Biostatistical Matrix Time Series Models

**Seyed Yaser Samadi\***, University of Georgia  
**Lynne Billard**, University of Georgia

Many data sets in biology, medicine, and other biostatistical areas deal with multiple sets of multivariate time series. The case of a single univariate time series is very well developed in the literature; and single multivariate series though less well studied have also been developed (under the rubric of vector time series). A class of matrix time series models is introduced for dealing with the situation where there are multiple sets of multivariate time series data. Explicit expressions for a matrix autoregressive model of order one along with its cross-autocorrelation functions are derived. Stationarity conditions are also provided.

email: ysamadi@uga.edu

### Supervised Singular Value Decomposition and its Asymptotic Properties

**Gen Li\***, University of North Carolina, Chapel Hill  
**Haipeng Shen**, University of North Carolina, Chapel Hill  
**Dan Yang**, Rutgers, The State University of New Jersey  
**Andrew Nobel**, University of North Carolina, Chapel Hill

We develop a supervised singular value decomposition (SupSVD) model for supervised dimension reduction. The research is motivated by applications where the low rank structure of the data of interest is potentially driven by additional variables measured on the same set of samples. The SupSVD model can make use of the information in the additional data to accurately extract underlying structures that are more interpretable. The model is very general and includes the principal component analysis model and the reduced rank regression model as two extreme cases. We formulate the model in a hierarchical fashion using

latent variables, and develop a modified expectation-maximization algorithm for parameter estimation, which is computationally efficient. The asymptotic properties for the estimated parameters are derived. We use comprehensive simulations and two real data examples to illustrate the advantages of the SupSVD model.

email: ligen@live.unc.edu

### **brainR: Interactive 3 and 4D Images of High Resolution Neuroimage Data**

**John Muschelli\***, Johns Hopkins  
Bloomberg School of Public Health  
**Elizabeth M. Sweeney**, Johns Hopkins  
Bloomberg School of Public Health  
**Ciprian M. Crainiceanu**, Johns Hopkins  
Bloomberg School of Public Health

We provide software tools for displaying and publishing interactive 3-dimensional (3D) and 4-dimensional (4D) figures, with examples of high-resolution brain imaging. Our framework is based in the R statistical software using the rgl package, a 3D graphics library. We build on this package to allow manipulation of figures including rotations and translation, digital zooming, coloring of brain substructures, adjusting transparency levels, and addition/or removal of brain structures. The need for better visualization tools of ultra high dimensional data is ever present; we are providing a clean, simple, web-based option. We also provide a package (brainR) for users to readily implement these tools.

email: jmuschel@jhspsh.edu

## **48. CLINICAL TRIALS**

### **Outcome-Adaptive Allocation with Natural Lead-in for Three-Group Trials with Binary Outcomes**

**Ghalib A. Bello\***, Virginia Commonwealth University  
**Roy T. Sabo**, Virginia Commonwealth University

Just as Bayes extensions of the frequentist optimal allocation design have been developed for the two-group case, we provide a Bayes extension of optimal allocation in the three-group case. We use the optimal allocations derived for the three-group case (Jeon & Hu, 2010) and estimate success probabilities for each treatment group using a Bayes estimator. We also introduce a natural lead-in design that allows adaptation to begin as early in the trial as possible. Simulation studies show that the Bayesian adaptive designs simultaneously increase the power and expected number of successfully treated patients compared to the balanced design. And relative to the standard adaptive design, the natural lead-in design introduced in this study produces a higher expected number of successfully treated patients whilst preserving power.

email: belloga@vcu.edu

### **Trial Design and Analysis Challenges when Studying Therapies Designed to Control Growth of Brain Metastases in Cancer Patients**

**Sujata M. Patil\***, Memorial Sloan-Kettering Cancer Center

The presence of brain metastases in cancer patients often indicates poor prognosis. Additionally, the presence of brain metastases can directly impact a patient's quality of life. Controlling brain disease is important and has been one current focus of clinical trials and retrospective reviews [Preusser et al, Eur J Cancer 2012; Lin, eCancer 2013]. However, there are challenges in conducting such studies and interpretations of results are not uniform. For instance, patients may progress extracranially before progression in the brain can be assessed, thereby creating a competing risks analytic setting. Assessing true brain recurrence versus radionecrosis and the use of consistent criteria to assess brain recurrence have also been methodological issues. In this work, we describe these statistical challenges and use simulation studies to propose criteria on how best to conduct prospective and retrospective clinical studies in this patient population.

email: patils@mskcc.org

### **Understanding Inconsistencies Between Replicate Trials: Insomnia Case Study**

**Richard Entsuah**, Merck  
**Kenneth Liu\***, Merck  
**Junshui Ma**, Merck  
**Duane Snavely**, Merck  
**Ellen Snyder**, Merck

FDA approval of a new drug often requires replicate confirmatory trials (at least two positive trials). This presentation discusses strategies one can use if one study is positive and the other is negative. One strategy is to identify and down-weight outliers. Another strategy is to calculate the type two error since an effective drug could fail to show efficacy.

email: Kenneth\_Liu@Merck.com

### **Sample Size Determination for a Three-Arm Equivalence Trial of Normally Distributed Responses**

**Yu-Wei Chang\***, Temple University  
**Yi Tsong**, U.S. Food and Drug Administration  
**Xiaoyu Dong**, U.S. Food and Drug Administration  
**Zhigen Zhao**, Temple University

The equivalence assessment is often conducted through a three-arm clinical trial and it usually consists of three tests. The first two tests are to demonstrate the superiority of the test and the reference treatment to placebo. And they are followed by the equivalence assessment between the test and the reference treatment. With continuous response, equivalence is commonly defined in terms of mean difference, mean ratio or ratio of mean difference, i.e. the mean difference of the test and placebo to mean difference of the reference and placebo. The advantage of applying the equivalence test by ratio of mean difference is

that it can test both superiority of the test treatment over placebo and equivalence between the test and the reference simultaneously. In this paper, we derive the test statistics and the power function for the ratio of mean difference hypothesis and solve the required sample size for a three-arm clinical trial. Examples of required sample size are given in this paper, and are compared with the required sample size by the traditional mean difference equivalence test. After a careful examination, we suggest to increase the power of the mean difference ratio approach by appropriately adjusting the lower limit of the equivalence interval.

email: changvick@gmail.com

### **The Utility of Bayesian Predictive Probabilities for Interim Monitoring of Clinical Trials**

**Benjamin R. Saville\***, Vanderbilt University School of Medicine

**Jason Connor**, Berry Consultants

**Gregory Ayers**, Vanderbilt University School of Medicine

**JoAnn Alvarez**, Vanderbilt University School of Medicine

Bayesian predictive probabilities can be used for interim monitoring of clinical trials to estimate the probability of observing a statistically significant treatment effect if the trial were to stop enrolling and follow the current cohort or continue to its predefined maximum sample size. We give an overview of the advantages of Bayesian predictive probabilities for interim monitoring versus alternative strategies such as Bayesian posterior probabilities, p-values, conditional power, and group sequential methods. The benefits of predictive probabilities are discussed in the context of futility and efficacy monitoring, where it is shown that only predictive probabilities properly account for the amount of data remaining to be observed in a clinical trial and have the flexibility to account for information in variables correlated with delayed outcomes.

email: b.saville@vanderbilt.edu

### **Evaluation of Bias for Outcome Response Adaptive Randomization Designs**

**Yaping Wang\***, University of Texas MD Anderson Cancer Center and University of Texas School of Public Health

**Hongjian Zhu**, University of Texas School of Public Health

**J. Jack Lee**, University of Texas MD Anderson Cancer Center

Outcome response adaptive randomization (RAR) designs, which change randomization probabilities sequentially based on observed outcome, assign more patients to the better treatment and can yield higher overall response rates for patients in trials compared to the conventional equal randomization (ER) procedure. However, the resulting maximum likelihood estimator of the response rate tends to be under estimated in RAR trials. Although the bias will converge to zero when sample size increases, it is non-negligible in small trials. In some settings, the bias can

be as large as 10% of the true response rates. To better understand the cause of bias, we derived the large sample approximation for the bias. In addition, we conducted simulation studies to quantify the magnitude of the bias under several settings to assess the accuracy of the asymptotic approximations. We also illustrated the bias and provide intuitive explanation for its cause in simulated trials. Deeper understanding of the bias can help us design better RAR trials and provide more accurate estimates.

email: yaping.wang@mdanderson.org

### **Analysis of the Anticipated Power of a Test: Browne (1995) Revisited**

**Paul W. Stewart\***, University of North Carolina, Chapel Hill

When a preliminary study is conducted to obtain an estimate of the population standard deviation (SD) based on a sample of size  $M$ , Browne (1995) proposed that an upper one-sided confidence limit for SD should be used as the conjectured value of SD for purposes of choosing a sample size,  $N$ , for the future study. We examine the performance of this strategy for cases frequently encountered.

email: paul\_stewart@unc.edu

## **49. PERSONALIZED MEDICINE AND VARIABLE SUBSET SELECTION**

### **Multivariate Markov Models for the Conditional Probability of Toxicity in Phase II Trials**

**Laura L. Fernandes\***, University of Michigan

**Susan Murray**, University of Michigan

**Jeremy MG Taylor**, University of Michigan

In addition to getting an idea of efficacy, phase II trials can also help to determine dose(s) that have an acceptable toxicity profile over repeated cycles. Correct modeling of the dose toxicity relationship in patients receiving multiple cycles of the same dose in oncology trials is crucial. A major challenge lies in taking advantage of the conditional nature of data collection, i.e., each cycle is observed conditional on having no previous toxicities on earlier cycles. We develop a parsimonious model for the conditional probability of toxicity during a cycle  $k$  conditional on not seeing toxicity in any of the  $k-1$  previous cycles using a Markov model. Our model allows the conditional probability of toxicity to depend on randomized dose group, cumulative dose from prior cycles, a measure of how consistently a patient responds to the same dose exposure and individual risk factors. Simulations studying finite sample properties of the model are given and demonstrated in a phase II trial studying two dose levels of ifosfamide plus doxorubicin and granulocyte colony-stimulating factor in soft tissue sarcoma patients over four cycles. Our model provides correct estimates of the probabilities of toxicity in finite sample simulations and correctly models the data from the phase II trial. Further investigation of its application for different trials is required.

email: flaura@umich.edu

## Latent Supervised Learning for Estimating Treatment Effect Heterogeneity

**Susan Wei\***, University of North Carolina, Chapel Hill  
**Michael R. Kosorok**, University of North Carolina, Chapel Hill

It is oft observed in medicine that what works for one patient may not work for another. Determining when and for whom a treatment works and does not work is of great clinical interest. We propose a methodology to estimate treatment effect heterogeneity, i.e. to ascertain for which subpopulations a treatment is effective or harmful. The model studied assumes the relationship between an outcome of interest (e.g. blood pressure, cholesterol, survival) and a set of covariates (e.g. treatment, age, gender) is modified by a linear combination of a set of features (e.g. gene expression). Specifically a threshold on the linear combination divides the population into two subpopulations with different responses to treatment. Techniques from Latent Supervised Learning, a novel machine learning idea, is applied for model estimation. Consistency of the estimator is established. In simulations the proposed methodology demonstrates high classification accuracy in a wide array of settings. Three data analysis examples are presented to illustrate the efficacy and applicability of the proposed methodology.

email: susanwe@live.unc.edu

## Personalized Selection of Radiation Therapy Dose Using Statistical Models for Toxicity and Efficacy with Dose and Biomarkers as Covariates

**Matthew Schipper\***, University of Michigan  
**Jeremy MG Taylor**, University of Michigan  
**Feng-Ming Kong**, Georgia Regents University  
**Randy TenHaken**, University of Michigan  
**Martha Matuzak**, University of Michigan

Dose selection for cancer patients treated with Radiation Therapy (RT) must balance the increased local tumor control (LC) probability with the increased toxicity probability associated with higher dose. Historically, a single dose has been selected for a population of patients. The availability of new, possibly mid-treatment, biologic markers for toxicity allow more personalized dose selection. We consider use of statistical models for toxicity and LC, with dose and biomarkers as covariates, to select an optimal dose that maximizes the probability of LC minus a weighted sum of toxicity probabilities. This function is equal to the expected utility of the bivariate efficacy/toxicity outcome for a particular family of utility values (2x2 matrix). Because metrics such as AUC do not assess the ability of a marker to improve LC at a fixed rate of toxicity, we use a simulation approach to study the effect of marker based dose selection on toxicity and efficacy outcomes. If dose is linearly related to toxicity and LC, then any marker that only acts additively with dose, cannot improve LC without also increasing the toxicity. We illustrate the issues involved with a lung cancer dataset and conclude with implications for early phase RT trial designs.

email: mjschipp@umich.edu

## Simultaneous Inference for Assessing the Effects of a SNP on Treatment Efficacy in Personalized Medicine

**Ying Ding\***, University of Pittsburgh  
**Grace Li**, Eli Lilly and Company  
**Stephen J. Ruberg**, Eli Lilly and Company  
**Jason C. Hsu**, Eli Lilly and Company and The Ohio State University

There has been increasing interest in discovering personalized medicine in current drug development using genetic polymorphisms. Testing for SNPs predictive of treatment efficacy, measured by a clinical outcome, is fundamentally different from association detection for a quantitative trait. In personalized medicine, clinical effect size matters and an important decision to make is which subgroup or union of subgroups of patients should the drug being developed for. Concentrating on the starting point of this practice, making inference within each SNP, we suggest an approach that is informative for the purpose of drug development. For a single SNP, we provide simultaneous confidence intervals for assessing different genetic effects (such as dominant, recessive, and additive) on clinical response. In Type II diabetes, for instance, they would be confidence intervals for mean difference between treatment and control of HbA1c reduction from baseline for all tested effects. This method is more informative for the following reason. A reduction in HbA1c between 0.8 and 1.2 is much more clinically meaningful than a reduction between 0.4 and 0.6. Yet the confidence intervals (0.8, 1.2) and (0.4, 0.6) can have identical p-values. The improvement of the within SNP inference is crucial for a better overall inference across the SNPs in assessing treatment efficacy.

email: yingding@pitt.edu

## Consistent Variable Selection for Quantile Regression with Varying Covariate Effects

**Qi Zheng\***, Emory University  
**Limin Peng**, Emory University

Quantile regression provides a flexible platform for evaluating covariate effects on different segments of the conditional distribution of response. As the effects of covariates may change with quantiles, contemporaneously examining a spectrum of quantiles is expected to produce more robust variable selection than focusing on a single quantile. Under this motivation, we study a new penalization strategy in the quantile regression setting where a continuum of quantile index is considered. We establish the oracle properties of the resulting estimator, which include consistent identification of all covariates that have effects on some or all quantiles of interest, and the same weak convergence behavior to a Gaussian process as if the true set of relevant covariates was known in advance. Furthermore, we investigate a BIC-type uniform tuning parameter selector and show it can ensure consistent model selection, a property which may not be achieved by other methods, such as AIC or generalized cross validation. Our numerical studies confirm the theoretical findings and present an application of the new variable selection procedure.

email: qi.zheng@emory.edu

### Consistent Bi-Level Variable Selection Via Composite Group Bridge Regression

**Indu Seetharaman**, Kansas State University  
**Kun Chen\***, University of Connecticut

We propose a composite group bridge penalized estimation approach for conducting bi-level variable selection in regression models with a diverging number of predictors. The proposed method combines the ideas of bridge regression and group bridge regression, to achieve variable selection consistency in both individual and group levels simultaneously, i.e., the important groups and the important individual variables within each group can both be correctly identified with probability approaching to one as the sample size increases. The method takes full advantage of the prior grouping information, and the established bi-level oracle properties ensure that the method is immune to possible group misspecification. Simulation studies and a real application show that the proposed methods have superior performance in comparison to several existing methods.

email: kun.chen@uconn.edu

### Penalized Regression for Interval-Censored Times of Disease Progression: Selection of HLA Markers in Psoriatic Arthritis

**Ying Wu\***, University of Waterloo  
**Richard Cook**, University of Waterloo

Times of disease progression are interval-censored when progression status is only known at a series of assessment times. This situation arises routinely in clinical trials when events of interest are only detectable upon imaging, based on blood tests, or upon detailed clinical examination. We consider the problem of selecting important biomarkers prognostic for progression from a high dimensional set of covariates when the progression time is interval-censored. We adapt methods for variable selection based on penalized regression (e.g. LASSO, adaptive LASSO and SCAD) to handle interval-censored time of disease progression. An expectation-maximization algorithm is developed which is empirically shown to perform well. Application to data from the motivating study on predicting disease progression in psoriatic arthritis is given and several important human leukocyte antigen (HLA) variables are identified for further investigation.

email: wuyingapril@gmail.com

## 50. ANALYSIS OF CLUSTERED DATA

### Statistical Methods for Assessing Perception in Children with Cochlear Implants

**Michael D. Larsen\***, The George Washington University  
**Cynthia Core**, The George Washington University  
**Janean Wilson**, Children's National Medical Center  
**James Mahshie**, The George Washington University

Research in speech perception typically focuses on group outcomes rather than individual outcomes, yet there is a need for clinical assessment of speech perception. Few tools are available, and those that are available require behaviorally conditioned responses. Thus, there is a need for a perception measure that has a low behavioral response load and that can be used with younger children. Data were collected from multiple perception trials of 9 children ages 3-5 years using a new adaptation of Horn and Houston's Visual Reinforcement Habituation Procedure. In order to incorporate multiple variables describing the data structure and due to the small sample size, data are analyzed using a Bayesian linear model. The statistical method enables comparison of estimates and measures of uncertainty for each child in the small sample situation. This talk presents contrasts to statistical methods, including methods that analyze data from each child separately and random effects models, and the importance of the chosen approach.

email: mlarsen@bsc.gwu.edu

### Identify Common Clusters in Independent Populations with Application to Psychiatry

**Yun Zhang\***, University of Pittsburgh  
**Kehui Chen**, University of Pittsburgh  
**Allan Sampson**, University of Pittsburgh  
**David Volk**, University of Pittsburgh

We consider the problem of identifying whether or not a cluster defined from one population exists in a new independent population. Our work is motivated by schizophrenia and bipolar studies, in which we are interested in assessing whether or not in a population of bipolar individuals there is a cluster characterized by mRNA expression levels previously identified in a population of subjects with schizophrenia. In this presentation, we present a statistical formulation for testing our hypothesis that circumvents some of the non-regularity conditions of more standard approaches. Our procedure recasts the formulation using equivalence testing concepts in a finite normal mixtures framework. The proposed approach is illustrated in a simulation study and applied to our data.

email: yuz52@pitt.edu

### **Generalized Estimating Equation in Analyzing Group-Randomized Trials with Limited Number of Groups**

**Peng Li\***, University of Alabama, Birmingham  
**David T. Redden**, University of Alabama, Birmingham

The generalized estimating equation (GEE) with robust sandwich estimator is increasingly common in analyzing correlated data due to fewer assumptions and the consistent estimation of both the regression parameters and their standard errors. However, the sandwich estimator tends to underestimate the true variance in small samples and results in inflated type I error rates, which limits the application of GEE in group-randomized trials (GRTs) with small number of groups (K). We simulated the correlated binary responses under different GRT scenarios; and compared the small sample properties of GEE Wald statistics using different Bias-corrected variance estimators. Our results suggested that, by using Kauermann and Carroll proposed Bias-corrected estimator and t-distribution approximation with  $K - 2$  degrees of freedom, the Wald test can keep the type I error rate to nominal levels even when the number of groups is as low as 10. Furthermore, the proposed Wald test is robust to the variation of group sizes since the leverages of group were taken into account in the variance estimation. Based on the proposed test, a novel formula of the power calculation for GRTs is proposed. In conclusion, with the control of type I error rates under small sample sizes; we recommend the use of GEE method in GRTs due to fewer assumptions and robustness to the misspecification of the covariance structure.

email: pli@uab.edu

### **Accounting for Covariates in Differential Methylation Analysis with Next-Generation Sequencing**

**Hongyan Xu\***, Georgia Regents University  
**Robert Podolsky**, Wayne State University  
**Duchwan Ryu**, Georgia Regents University  
**Varghese George**, Georgia Regents University

DNA methylation at CpG loci is an important biomedical process involved in many complex diseases including cancers. In recent years, the development of next-generation sequencing (NGS) yields large amount of DNA methylation data. We have developed a statistical approach for detecting differentially methylated CpG sites for NGS data based on clustered data analysis. However, our approach did not allow for covariates. Research has shown that DNA methylation is correlated with age, sex, population and cell type composition. Therefore, it is important to account for the effect of such covariates in differential methylation analysis. In this study, we extend our method by modeling NGS methylation data explicitly as clusters within each individual with logistic regression to allow for covariates. Simulations show that the extended test maintains correct type-I error rate and is robust under several distributions for the measured methylation levels. It improves power over our previous test. Finally, we apply the test to our NGS data on chronic lymphocytic leukemia. The results indicate that it is a promising and practical test for genome-wide methylation analysis.

email: hxu@gru.edu

### **Evaluating Predictors of Individual Dietary Intake Latent Values Under Different Mixed Models**

**Shuli Yu\***, University of Massachusetts, Amherst  
**Edward J. Stanek III**, University of Massachusetts, Amherst

An accurate estimate of subject's latent value is important to establish a diagnosis, determine a patient's treatment, or evaluate a risk factor. We explore a more accurate method to estimate individual latent values of dietary intake when it is repeatedly measured using a 24-hour dietary recall (24HR) and seven day dietary recall (7DDR), accounting for measurement error and bias. We use a finite population mixed model (FPMM) to link identifiable subjects in evaluating the predictor at the individual (cluster) level. The performance of (empirical) predictor of subject's latent value obtained under the FPMM framework is compared with those obtained under the usual mixed model and the measurement error model through a simulation study. We analyzed the predictor of latent value in two cases: for a randomly selected subject and for a specific subject. We illustrated the approach by using dietary intake data from the Seasons Study, and evaluated the performance of the predictors based on the 24HR data and the 24HR and 7DDR combined data. The results reveal the predictor under FPMM is optimal for a randomly selected subject, but not uniformly optimal for a specific subject. The expected MSE of predictor from the combined data is not always smaller.

email: shuli@schoolph.umass.edu

### **A Markov Mixture Model for Longitudinal Course of Youth Bipolar Disorder**

**Jieyu Fan\***, University of Pittsburgh  
**Satish Iyengar**, University of Pittsburgh  
**Boris Birmaher**, University of Pittsburgh  
**Adriana Lopez**, Carnegie Mellon University, Qatar  
**Rasim S. Diler**, University of Pittsburgh  
**David Axelson**, The Ohio State University  
**Benjamin Goldstein**, University of Toronto  
**Tina Goldstein**, University of Pittsburgh  
**Fangzi Liao**, Western Psychiatric Institute and Clinic  
**Mary K. Gill**, Western Psychiatric Institute and Clinic

Bipolar Disorder is characterized by recurrent mood changes ranging from depression, to excessive happiness or irritability. Measuring and making sense of the fluctuations in these moods over time is challenging. To find homogeneous clusters and capture different longitudinal mood change patterns we introduced a Markov mixture model with different transition matrices in an observational study of 412 children and adolescents with bipolar disorder who were followed on average for 344 weeks. The parameters of this model were estimated using the EM algorithm; we determine the number of clusters using several information criteria and cross-validation. Our clusters separate out those who tend to stay in a state from those who jump between states more frequently. We will present further results of these analyses and discuss potential clinical implications.

email: yuyufan05@gmail.com

### Longitudinal Multivariate Outcome Data from Couples: Application To HPV Transmission Couple Studies

**Xiangrong Kong\***, Johns Hopkins University Bloomberg School of Public Health

HPV is a common STI with 14 known oncogenic genotypes causing anogenital carcinoma. While gender-specific infections have been well studied, one remaining uncertainty in HPV epidemiology is HPV transmission within couples. Understanding transmission in couples however is complicated by the multiplicity of genital HPV genotypes and sexual partnership structures that lead to complex multi-faceted correlations in data generated from HPV couple cohorts, including inter-genotype, intra-couple, and temporal correlations. We develop a hybrid modeling approach using Markov transition model and composite pairwise likelihood for analysis of longitudinal HPV couple cohort data to identify risk factors associated with HPV transmission, estimate difference in risk between male-to-female and female-to-male HPV transmission, and compare genotype-specific transmission risks within couples. The method is applied on the motivating HPV couple cohort data collected in the male circumcision trial in Rakai, Uganda to identify modifiable risk factors (including male circumcision) associated with HR-HPV transmission within couples. Knowledge from this analysis will contribute to the public health effort in preventing oncogenic HPV and related cancers in sub-Saharan Africa.

email: xikong@jhsph.edu

## 51. THE ROLE OF STATISTICS IN SHAPING PUBLIC POLICY

### Statisticians: Guardians of Democracy!

**Roderick J. Little\***, University of Michigan

I recently completed an assignment to the U.S. Census Bureau, where my main role was to help set up a new research directorate. I will discuss the important role government statistical agencies play in our democracy, and the role of research at the Census Bureau and government statistical agencies in general. I will also give my views on why Bayesian statistical modeling is particularly important in modern government statistics, why it is not inconsistent with objective analysis of data, and why it should be central to a coherent paradigm for statistical inference.

email: rlittle@umich.edu

### Big Statistics, Major Policies, and... a Little Politics

**Sally C. Morton\***, University of Pittsburgh

Utilizing case studies based on Agency for Healthcare Research and Quality evidence-based medicine projects and Institute of Medicine reports on healthcare reform topics, I will demonstrate how statistics and statisticians can shape health policy. I will give particular attention to the challenges of observational studies in the era of big data and the potential impact of methodological standards for research.

I will illustrate the importance of the science of statistics in policy, as well as underscore the need to engage decision-makers and communicate information from a policy perspective while being ever mindful of the political setting.

email: scmorton@pitt.edu

## 52. PANEL DISCUSSION: HAVING IT ALL: WEIGHTING TO ACHIEVE BALANCE

### Having it all: Weighting to Achieve Balance

**Thomas M. Braun**, University of Michigan

Dr. Braun has been a faculty member with the University of Michigan Department of Biostatistics for fifteen years. He held a research-track position for the first half of his career and then switched to a tenure-track position for the second half. Dr. Braun has one child and is married to a wife who works full-time and can give one male's perspective on how he successfully balances between the demands of being involved in the daily activities of his family and his many roles as a professor, such as collaborating with advisors, publishing manuscripts, lecturing, and supervising graduate students. Dr. Braun has successfully obtained tenure without sacrificing his personal priorities and is thrilled to have a challenging career that still allows him to walk his daughter to school each day.

email: tombrun@umich.edu

### Having it all: Weighting to Achieve Balance

**Mary D. Sammel**, University of Pennsylvania

I am a Professor of Biostatistics at the Perelman School of Medicine, University of Pennsylvania, and spend the majority of my time collaborating with epidemiologists and physicians in the design and analysis of clinical research studies primarily in the areas of reproduction and women's health. For the past 6 years I have taught a 2nd semester course in Biostatistical Methods for Epidemiology, and am active in teaching and mentoring students in both Biostatistics and Epidemiology, as well as medical students and fellows. I have 2 children, ages 17 and 13, and have been married for 22 years. My contribution to the panel will be to share what I learned as a young assistant professor with 2 small children when my spouse was deployed by the US Marine Corps to go to Iraq, not once, but twice. While you may never experience this particular family situation, I hope to share with you some useful survival tips.

email: msammel@upenn.edu

### Having it all: Weighting to Achieve Balance

**Telba Z. Irony**, U.S. Food and Drug Administration

Telba Irony is the Chief of the General and Surgical Devices Branch in the Division of Biostatistics at the Center for Devices and Radiological Health at the Food and Drug Administration. She received her Ph.D. from the University of California at Berkeley and was on the faculty of the School of Engineering at the George Washington University. She joined CDRH to help implement the use of Bayesian methods in Medical Device Clinical Trials. She worked on several NSF grants on Bayesian Statistics and produced more than 50 peer-reviewed articles on Bayesian methods. Telba is a fellow of the American Statistical Association and an elected member of the International Statistical Institute. Telba Irony is currently leading the Decision Analysis initiative at CDRH, which involves Benefit: Risk assessments, Patient Preference Surveys, and Decision Tools for CDRH's Innovation Pathway. In this panel, Telba will discuss the life of a statistician in the government, highlighting the fact that the statistician is a problem solver, who must be interested in science and teaching, and could aspire to leadership positions. She will also present ideas on how to achieve good balance between work and family life when pursuing a government career.

email: telba.irony@fda.hhs.gov

### Having it all: Weighting to Achieve Balance

**Aarti Shah**, Eli Lilly & Company

Brief background on me (Aarti Shah). I am an Indian by origin, married for 24 years and am a mother of two boys (15 and 17 yrs old). I completed my Bachelor and Masters in Statistics and Mathematics from India before venturing to the U.S. to pursue my PhD in Statistics at University of California, Riverside. I joined Eli Lilly and Company in January 1994 as a Sr Statistician. I have recently taken a new role as the Global Brand Development Leader for an autoimmune molecule in development. Previous to this role, I was the Vice President for Biometrics and Advanced Analytics. I have almost 20 years of experience in the pharmaceutical industry and have worked across various phases of drug development in different roles. I also serve on the Board of Directors for the Indianapolis Public Library Foundation. I hope to bring several diverse perspectives to this session based on my personal and professional experiences.

email: aarti@lilly.com

### Having it all: Weighting to Achieve Balance

**Francesca Dominici**, Harvard School of Public Health

I don't have an answer to the question of Why women can't have it all. There is not a single correct answer. This is a topic dominated by personal choices, and strategies highly depend on the situation, the work environment, the cultural background etc. However I do believe you are all here because of a common goal: you are here because you are an ambitious woman that loves her job and wants to be successful at it while at the same time you want to be a good mother and/or caregiver. We will have a frank discussion of how to make that happen.

email: fdominic@hsph.harvard.edu

## 53. BIOSTATISTICAL METHODS FOR INTEGRATIVE GENOMICS

### A Brief Overview of Modelling Approaches in Integrative Genomics, with Special Reference to eQTL Analyses

**Sylvia T. Richardson\***, Cambridge Institute of Public Health

Developing statistical models that relate, combine or integrate different types of genomics data collected on the same set of subjects is a challenging task that has been tackled from a variety of points of view. In the first part of the talk, a brief review of the characteristics of different types of genomic data, statistical methods of integrative analysis, and a few examples will be presented, with emphasis on the design of integrative analysis to address appropriate biological questions. In the second part of the talk, a particularly important example of integrative genomics analysis, the investigation of the genetic regulation of transcription, the so called eQTL studies, will be discussed. In this design, two large data sets are crossed, and the analysis follows a natural structure of parallel regressions between the large set of  $q$  responses, e.g. the expression phenotypes, and an even larger set of  $p$  explanatory variables, the genetic markers or SNPs, where  $p$  is typically much larger than the number of subjects  $n$ . From a statistical point of view, the size and the complex multidimensional structure of such analyses pose a significant challenge. A hierarchical modelling strategy will be presented, its performance reviewed and illustrated.

email: sylvia.richardson@mrc-bsu.cam.ac.uk

### Information Integrative Framework for Sparse K-Means to Combine Multi-Cohort and Multi-Omics Data

**Zhiguang Huo**, University of Pittsburgh  
**Sunghwan Kim**, University of Pittsburgh  
**George C. Tseng\***, University of Pittsburgh

Disease subtype discovery has been an important task in complex diseases. Identifying meaningful disease subtypes with differential disease progression, survival outcome and drug response brings translational impact for precision medicine. We have extended the sparse K-means algorithm to a meta-analysis framework to combine multiple gene expression profiles for more accurate and robust disease subtype discovery. The method identifies a common intrinsic feature set for disease phenotyping via regularization, reports clustering in each study and matches cluster patterns across studies. We also extend the framework to combine multi-omics data, including methylation, miRNA expression, CNV and etc. Simulation and application to breast cancer data have found improved performance than individual study analysis.

email: ctseng@pitt.edu

### EgoNet: Identification of Disease Ego-Network Modules

**Rendong Yang**, Emory University  
**Zhaohui S. Qin**, Emory University  
**Tianwei Yu\***, Emory University

Mining novel gene markers from genome-wide gene expression profiles for accurate disease classification is very challenging due to the reasons of small sample size and high noise in gene expression measurements. Several studies have proposed integrated analyses of microarray data and protein-protein interaction (PPI) networks to find diagnostic subnetwork markers. However, the neighborhood relationship among network member genes has not been fully considered by those methods, leaving many potential gene markers unidentified. Here we present EgoNet, a novel method based on egocentric network-analysis techniques, to exhaustively search and prioritize the disease subnetwork and gene markers from a large-scale biological network for accurate clinical outcome prediction. When applied to a triple-negative breast cancer (TNBC) microarray dataset, the top selected modules contain both known gene markers in TNBC and novel candidates, such as DOK1 and NCOA2, which play a central role in their ego-networks by connecting many differentially expressed genes. Our results suggest ego-network study allows the identification of reliable biomarkers and provides a deeper understanding of their roles in complex diseases.

email: tianwei.yu@emory.edu

### Extensions to Hidden Markov Models and their Application to Integrated Analysis of Multiple Chromatin Immunoprecipitation Data

**Hyung Won Choi**, National University of Singapore  
**Damian Famian**, University of Michigan  
**Alexey Nesvizhskii**, University of Michigan  
**Debashis Ghosh**, The Pennsylvania State University  
**Zhaohui S. Qin\***, Emory University

Multiply correlated datasets have become increasingly common in genome-wide location analysis of regulatory proteins and epigenetic modifications. Their correlation can be directly incorporated into a statistical model to capture underlying biological interactions, but such modelling quickly becomes computationally intractable. In this project, we developed sparsely correlated hidden Markov models (sCHMM), a novel method for performing simultaneous HMM inference for multiple genomic datasets. In sCHMM, a single HMM is assumed for each series, but the transition probability in each series depends not only on its own hidden states, but also the hidden states of other related series. For each series, sCHMM uses penalized regression to select a subset of the other data series and estimate their effects on the odds of each transition in the given series. Following this, hidden states are inferred using a standard forward-backward algorithm with the transition probabilities adjusted by the model at each position, which helps retain the order of computation close to fitting independent HMMs (iHMM). We conducted simulation studies as well as real data analysis to illustrate the advantages of the sCHMM algorithm.

email: zhaohui.qin@emory.edu

## 54. SAFETY SURVEILLANCE MONITORING THROUGH SIGNAL DETECTION

### Methodological Challenges for Sequential Medical Product Safety Surveillance Using Observational Healthcare Data

**Andrea J. Cook\***, Group Health Research Institute  
**Jennifer C. Nelson**, Group Health Research Institute

Post-licensure drug and vaccine safety monitoring activities can range from uncovering new and unexpected adverse events (signal identification) to providing more conclusive evidence about a specific suspected adverse event (signal confirmation). Traditionally, safety signal identification has been accomplished by screening a large number of events (e.g., 100's or 1000's) for potential association with many drug or vaccine exposures without analytic tailoring across different exposure-event pairs. In contrast, signal confirmation generally involves a detailed protocol-based epidemiological study that is specifically customized for a single or just a few exposure-event pairs. More recently, an intermediate type of evaluation has emerged that targets several (e.g., 5-10) pre-specified exposure-event hypotheses and conducts routine sequential monitoring over time, often during the initial uptake period of a new drug or vaccine. In this talk, I will describe these newer surveillance systems, which include the Vaccine Safety Datalink and Mini-Sentinel pilot, and highlight the methodological challenges they face when prospectively monitoring drug and vaccine safety using longitudinal healthcare database information. Further I will introduce new observational sequential methods developed to meet some of the challenges.

email: cook.aj@ghc.org

### Graphical Approaches for Disproportionality Analysis of Spontaneously-Reported Adverse Events in Pharmacovigilance

**Richard C. Zink\***, JMP Life Sciences at SAS Institute, Inc.

While randomized clinical trials are the gold standard for evaluating the efficacy of a new intervention, the available sample size is often insufficient to fully understand its safety profile. The risk a therapy may pose may not be well understood until it has been on the market for many years, taken by individuals who differ from those studied under the inclusion criteria of the clinical development program. Spontaneously-reported adverse events (SRAEs) are collected by regulatory agencies, pharmaceutical companies and device manufacturers to monitor the safety of a product once it reaches market. These data are generally obtained from physicians, patients, or the medical literature. SRAEs present a unique challenge in that there is no convenient measure for the total number of individuals using an intervention. In other words, there is no clear denominator to define an adverse event incidence for a particular drug. In order to identify potential safety-signals, disproportionality analysis methods are used to compare the rate at which a

particular event of interest co-occurs with a given drug with the rate this event occurs without the drug in the event database. We describe how dynamically interactive graphical displays of disproportionality can streamline analyses of SRAEs in post-market surveillance.

email: richard.zink@jmp.com

### **Likelihood Ratio Tests for Active Surveillance**

**Ram C. Tiwari\***, U.S. Food and Drug Administration

In this talk, we will present longitudinal likelihood ratio test (LRT) methods for large databases with exposure information. When the interest is in the evaluation of a signal of an adverse event for a particular drug compared with placebo or a comparator, the special case of the longitudinal LRT referred to as sequential LRT is also presented. The methods are applied to a real drug safety dataset. A small simulation study is also presented to evaluate the performance of the tests using the characteristics such as power and type-I error, over time.

email: ram.tiwari@fda.hhs.gov

### **Discussion of Safety Surveillance Monitoring Through Signal Detection**

**Theodore Lystig\***, Medtronic, Inc.

In addition to summarizing key points and notable contributions from the earlier speakers, the discussion will include a manufacturer's view of these recent developments and opportunities for implementation.

email: theodore.lystig@medtronic.com

## **55. MULTIPLE TESTING AND SIMULTANEOUS INFERENCES IN COMPLEX SETTINGS**

### **False Discovery Rate Control and Group Testing for Complex Omics Data**

**Andrew B. Nobel\***, University of North Carolina, Chapel Hill

**Gen Li**, University of North Carolina, Chapel Hill

**Andrey Shabalin**, Virginia Commonwealth University

**Ivan Rusyn**, University of North Carolina, Chapel Hill

**Fred A. Wright**, North Carolina State University

Although the basic properties and operating characteristics of false discovery control (FDR) are now well understood, effective FDR control for grouped hypotheses remains challenging when the correlation of test statistics is consequential. In this talk, we (i) review the motivation for ranking hypotheses using local FDR, and (ii) describe a group FDR-controlling procedure that is well suited to expression quantitative-trait loci (eQTL) analysis. Our methods are motivated by the ongoing analysis of GTEx data, the focus of which is investigating eQTLs in multiple human tissues. For the GTEx data, we have proposed a hierarchical model for the observed correlations of gene-SNP pairs across the available K tissues in a multi-tissue experiment by fitting a flexible correlation structure for genotype-expression associations. The model is fit using an empirical Bayes

approach, and provides interpretable posterior probabilities for each gene-SNP pair across the range of tissues. We will present general conditions under which local false discovery rates derived from these posteriors control FDR at the gene-SNP level. In addition, we will discuss the need to provide effective FDR control at the gene level and, using GTEx as a case study, demonstrate the behavior of several group FDR-controlling procedures that are well-suited to eQTL data.

email: nobel@email.unc.edu

### **Another Look at Robust PC-Based Stratification Control for Multiple Testing**

**Yi-Hui Zhou\***, North Carolina State University

The issue of robustness in computing principal components has received increased attention in genomics, and is used as a basic tool to properly control sources of spurious correlation in multiple testing. We identify and clarify three sources of error in computing PCs: (i) excessive correlation among sets of genomic features, (ii) individual outlying values, and (iii) high correlation between family members or otherwise related samples. We describe a series of methods to perform PC analysis that is robust to all of these sources of error. We study the effects of outliers for data with high dimension and low sample size and prove consistency under a spiked eigenvalue model. However, the exercise highlights current confusion over the proper role of sample eigenvectors vs. population eigenvectors in controlling spurious stratification. We discuss these issues, and the role they play in extreme multiple testing scenarios with millions of tests performed.

email: yihui2006@gmail.com

### **Simultaneous Inference of Multiple Rare Variants: Design, Power and Interpretation of Findings**

**Andriy Derkach**, University of Toronto

**Jerry F. Lawless**, University of Waterloo

**Lei Sun\***, University of Toronto

The recent focus on rare variants has produced a large number of testing strategies to assess association between a group of rare variants and a trait, with competing claims about the performance of various tests. We show that many of the previous tests fall into either linear or quadratic class, and neither class consistently outperform the other across genetic models. This understanding leads to development of various robust tests that borrow strength from the two complementary classes. However, theoretical and empirical results have shown that it is difficult to achieve high power in the genome-wide setting due to multiple hypothesis testing. To increase power, we investigate various cost-effective response-dependent sampling strategies where samples are preferentially taken from the tails of the response distribution. We also address the selection bias issue, so that the genetic effects of rare variants are not overestimated and replication studies are not underpowered.

email: sun@utstat.toronto.edu

**Extending The Projack to Complex Settings**

**Fred A. Wright\***, North Carolina State University  
**Yi-Hui Zhou**, North Carolina State University

The projack resampling method has been developed to predict the underlying values of a parameter vector, where the hypotheses have been ranked according to observed statistics. Although inherently conditional on the data, the projack has not been applied to more complex settings, for which additional outcome-dependent criteria have been applied. We extend the projack to situations where significance selection has been performed, and study the impact of such selection compared to situations where ranking is performed without significance selection. In addition, we explore the projack as a tool to handle the secondary analysis of case-control data. The results have important implications in a variety of settings, including those of genomewide association scans.

email: fred\_wright@ncsu.edu

## 56. NEW DEVELOPMENTS IN BAYESIAN NONPARAMETRICS

**Scalable Bayesian Nonparametrics**

**David B. Dunson\***, Duke University

Bayesian nonparametric methods provide a useful framework for flexible probabilistic modeling of high-dimensional and complex data in a variety of applications including biomedical studies. Conceptually, there are substantial advantages over alternative approaches relying on kernel methods and penalized optimization, particularly in biomedical studies in which characterizing uncertainty in inference is of critical importance. However, realizing these conceptual advantages in routine applications of big complicated biomedical data is hindered by the lack of scalable and robust algorithms for implementation. Motivated by this problem, we develop new scalable Markov chain Monte Carlo algorithms, which can be implemented very quickly even in huge data problems. Supporting theory is provided, and results are compared with state of the art competitors including variational approximations and optimization approaches.

email: dunson@duke.edu

**Bayesian Models of Structured Sparsity for Discovery of Regulatory Genetic Variants**

**Ryan P. Adams\***, Harvard University  
**Barbara Engelhardt**, Duke University

In genomic science, the amount of data has grown faster than the statistical methodologies necessary to perform them. Moreover, many of the most pressing questions require integrated analysis of high-dimensional, highly structured data. We consider the problem of identifying allelic heterogeneity, or multiple, co-localized genetic

regulators of gene transcription. Sparse regression techniques have been critical to the discovery of allelic heterogeneity because of their computational tractability in large data settings. These traditional methods are hindered by correlation between SNPs introduced by linkage disequilibrium. I will describe a new model for Bayesian structured sparse regression. This model exploits positive definite covariance functions that incorporate LD effects directly into a Gaussian field to yield sparse regression coefficients. This broadly applicable model of Bayesian structured sparsity enables more efficient parameter estimation techniques than models assuming independence would allow. We applied this model to a large study of expression quantitative trait loci, and found that our approach yields highly interpretable, robust solutions for allelic heterogeneity, particularly when the interactions between SNPs are well approximated by an additive model.

email: rpa@seas.harvard.edu

**Bayesian Nonparametric Inference of Population Admixtures**

**Maria De Iorio\***, University College London  
**Stefano Favaro**, Università degli Studi di Torino  
**Yee Whye Teh**, University of Oxford  
**Lloyd Elliott**, University College London

Genetic data obtained on population samples convey information about their evolutionary history. We propose a Bayesian nonparametric model to infer population admixture. Given multilocus genotype data from a sample of individuals, the model allows inferring the demographic origin of an individual, classifying individuals as unadmixed or admixed, inferring the number of subpopulations ancestral to an admixed population and the population of origin of chromosomal regions. We develop methods that allow for correlation between loci due to Linkage Disequilibrium by extending the Hierarchical Dirichlet Process (Teh et al. 2006) to include dependence between loci. We demonstrate the proposed model on simulated and real data and discuss methods to summarise the results of the MCMC output for the analysis of population admixture.

email: m.deiorio@ucl.ac.uk

**Pre-Surgical Assessment of Peritumoral Brain Activation Via a Bayesian Non-Parametric Potts Model**

**Timothy D. Johnson\***, University of Michigan

The Potts model has enjoyed much success as a prior model for image segmentation. Given the individual classes in the model, the data are modeled as Gaussian random variates or as random variates from some other parametric distribution. In this talk, we present a non-parametric Potts model and apply it to a functional magnetic resonance imaging study for the pre-surgical assessment of peritumoral brain activation. We assume the Z-score image can be segmented into activated, deactivated, and null states. Conditional on the state, the Z-scores come from some generic distribution that we model non-parametrically using a mixture of Dirichlet process priors within the Bayesian framework. The posterior distribution of the model parameters is estimated with a Markov chain Monte Carlo algorithm, and Bayesian

decision theory is used for state classification. Our Potts prior model includes two parameters, a spatial regularization parameter and a null prior probability parameter. We assume that these parameters are unknown and estimate them with other model parameters. We show through simulation studies that our model performs on par, in terms of posterior expected loss, with parametric Potts models when the parametric model is correctly specified and outperforms parametric models when the parametric model is misspecified.

email: tdjtdj@umich.edu

## 57. STATISTICAL GENETICS AND GENOMICS

### **Sparse Multivariate Factor Analysis Regression Models and its Applications to Integrative Genomics Analysis**

**Yan Zhou\***, University of Michigan

**Peter Song**, University of Michigan

**Pei Wang**, Fred Hutchinson Cancer Research Center

**Ji Zhu**, University of Michigan

The multivariate regression model is a useful tool to explore complex associations between multiple response variables and multiple predictors. When the multiple responses are correlated, ignoring such dependency will impair statistical power in the data analysis. Motivated by an integrative genomic data analysis, we propose a new methodology -- sparse multivariate factor analysis regression model (smFARM), in which correlations of the response variables are analyzed by a factor analysis model with latent factors. This proposed method not only allows us to address the challenge that the number of regression parameters is larger than the sample size, but also to adjust for unobserved genetic and/or non-genetic factors that potentially conceals the underlying response-predictor associations. The proposed smFARM is implemented efficiently by utilizing the strength of the EM algorithm and the group-wise coordinate descend algorithm. The proposed methodology is evaluated and compared to the existing methods through extensive simulation studies. We apply smFARM in an integrative genomics analysis of a breast cancer dataset on the relationship between DNA copy numbers and gene expression arrays to derive genetic regulatory patterns relevant to breast cancer.

email: zhouyan@umich.edu

### **A General Statistical Framework for Transcript Assemblies**

**Alyssa Frazee\***, Johns Hopkins University

**Geo Pertea**, Johns Hopkins University

**Steven Salzberg**, Johns Hopkins University

**Jeff Leek**, Johns Hopkins University

RNA-sequencing (RNA-seq) is now the most popular technology for quantitatively measuring gene expression. A major advantage of this technology is the ability to assemble - or estimate - the specific transcripts of each gene whose abundances are measured by RNA-seq data. Computational biologists have developed fast deterministic algorithms for

assembling transcripts and estimating their abundances. Here we propose a general statistical framework for analyzing the variation in the transcript assembly and for detecting differential expression at the transcript, exon, and gene level. Our framework is a general-purpose statistical backend for any assembler that both constructs transcripts and estimates abundances for those transcripts. To illustrate the power of our approach, we have built an R package backend called Ballgown for the popular Cufflinks transcript-assembly software. We analyze simulated and public RNA-seq experiments comparing Ballgown to the widely used Cuffdiff statistical software built into Cufflinks. We show (1) Ballgown is significantly more accurate for two-class differential expression, (2) Ballgown permits statistical analysis of a wider variety of experimental designs - such as time course experiments, and (3) Ballgown is much faster computationally. Our R package is freely available from Github.

email: acfrazee@gmail.com

### **Nonparametric Test for Differential Binding Analysis with ChIP-Seq Data**

**Qian Wu\***, University of Pennsylvania

**Kyoung-Jae Won**, University of Pennsylvania

**Hongzhe Li**, University of Pennsylvania

ChIP-seq is a powerful method for detecting genomic binding of DNA-associated proteins. It is important to identify genes with differential binding regions between two conditions, such as different cellular states or different time points. Parametric methods based on Poisson/ Negative Binomial distribution have been proposed to address this problem and most of these methods require biological replications. However, many ChIP-Seq data usually have a few or even no replicates. We propose a nonparametric method to identify differential binding regions, even without replicates. Our method is based on nonparametric hypothesis testing and kernel smoothing in order to capture spatial differences in protein-binding profiles. We demonstrate the method using a ChIP-Seq data on a comparative epigenomic profiling of adipogenesis of murine adipose stromal cells. Our method detects many genes with differential binding for the histone modification mark H3K27ac between two conditions. The test statistics also correlate with the gene expression changes well and are predictive to gene expression changes, indicating that the identified differential binding regions are indeed biologically meaningful.

email: wuqian7@gmail.com

### **A Statistical Framework for Expression QTL Mapping Via Two-Way Mixture Model**

**Ningtao Wang\***, The Pennsylvania State University

**Yaqun Wang**, The Pennsylvania State University

**Bruce Lindsay**, The Pennsylvania State University

**Rongling Wu**, The Pennsylvania State University

Expression quantitative trait loci (eQTL) are genetic regions associated with variation in gene expression among individuals. The single polymorphism associated with variation in the expression level for clusters of genes informs the hotspot with potentially pleiotropic effects. Here we

propose a two way mixture model based approach to detect eQTL hotspots for both gene expression microarray and RNA-seq data. We integrate unsupervised gene expression pattern discovery, interval mapping, and eQTL hotspots detecting into a single framework. A maximum-likelihood approach based on a two way mixture model, implemented with the two layer EM algorithm, is developed to provide the estimates of eQTL positions. More importantly, the two-way mixture model allows for detecting numerous types of eQTLs, including the eQTLs for global expression level, for sub-pattern expression level, and even for cluster interactions. Simulations and real data analysis demonstrate the power of our method.

e-mail: ntwang25@gmail.com

### **SVM With Bootstrap for Soft Clustering of Populations**

**Matey Neykov\***, Harvard University

The need of classifying individuals from a sample into distinct populations often arises in population genetics studies. Clustering based on ancestral population is an approach to study the underlying population structure. However, usually the fractional ancestry of the individuals in the sample and the allele frequencies from the ancestral populations are unknown. One approach is to use the EM (Expectation-Maximization) algorithm for model-based clustering to estimate the fractional ancestry using the haploid genotype of the individuals from the sample. The current paper proposes using a Support Vector Machine (SVM) classifier between homogeneous populations as a statistic, and obtaining the fractional ancestry through a bootstrap procedure as an alternative to the EM algorithm.

e-mail: mneykov@gmail.com

### **Functional Principal Component Analysis for Next Generation Sequencing**

**Lieven Clement\***, Ghent University

The advent of next generation sequencing (seq) technology allows for assessing genome-wide 'omics profiles at an unprecedented resolution. The downstream statistical analysis is commonly based on the number of sequenced reads mapping to the genomic regions of interest. The seq-technology conceptually allows for generating count profiles on a single nucleotide resolution. Many methods, however, aggregate counts based upon existing annotation. This often induces bias due to unreliable annotation, differences in target length and does not allow for discovery of novel regions. Within this context we develop wavelet based functional principal component methods for sequencing applications. Our method considers each 'omics profile as a functional realization across the genomic coordinate and enables an analysis on a single nucleotide resolution. The obtained principal component functions are very informative as they provide a data driven segmentation of the genomic profiles in regions with low and high variability across samples. The principal component functions can be readily adopted for unsupervised analysis, but, they also provide a sparse basis for downstream functional anova and regression approaches.

e-mail: lieven.clement@ugent.be

### **The Generalized Higher Criticism for Testing SNP-Sets in Genetic Association Testing**

**Ian J. Barnett\***, Harvard University

**Xihong Lin**, Harvard University

Genes, gene pathways, and network effects can contribute to the risk of complex genetic diseases. These genetic constructs each contain multiple SNPs, and only a sparse subset of these SNP-sets are generally related to the disease of interest. In this paper we adapt the higher criticism, a test traditionally used in high dimensional signal detection settings, to genetic association testing for SNP-sets. The higher criticism performs well when the signal is sparse, making it a potentially powerful tool for SNP-set association testing. Unlike past treatments of the higher criticism, we propose the generalized higher criticism (GHC) that does not require asymptotics in the number of SNPs in the SNP-set while simultaneously allowing for arbitrary correlation structures among the SNPs in the SNP-set. The power of this method is compared with existing SNP-set tests over simulated regions with varied correlation structures and signal sparsity. The relative performance of these methods is also compared in their analysis of the CGEM breast cancer genome-wide association study.

e-mail: ibarnett@hsph.harvard.edu

## **58. IMAGING**

### **Modeling Covariate Effects in Group Independent Component Analysis with Applications to Functional Magnetic Resonance Imaging**

**Ran Shi\***, Emory University

**Ying Guo**, Emory University

Independent component analysis (ICA) is a powerful computational tool for separating independent source signals from their linear mixtures. ICA has been widely applied in neuroimaging studies to identify and characterize underlying brain functional networks. An important goal in such studies is to assess the effects of subjects' clinical and demographic covariates on these functional networks. Currently, covariate effects can only be evaluated through ad-hoc approaches in ICA which may not be accurate in many cases. In this paper, we propose a hierarchical covariate ICA model that provides a formal statistical framework for estimating and testing covariate effects in ICA decomposition. Estimation in our model is accomplished through a maximum likelihood (ML) approach using the expectation-maximization (EM) algorithms. An approximate EM algorithm is also developed for fast computation when the number of ICs is large. Inferential procedures for testing covariate effects are based on voxel wise approximations to avoid expensive matrix inversions. The performance of the proposed methods is evaluated via simulation studies. The methods are applied to an fMRI study of Zen meditation.

email: rshi3@emory.edu

### **Quantile Mapping for Multi-Modal Imaging Data**

**Huaihou Chen\***, New York University School of Medicine  
**Philip T. Reiss**, New York University School of Medicine  
**Clare Kelly**, New York University School of Medicine  
**Xavier F. Castellanos**, New York University School of Medicine

Quantile map is useful for visualizing the rank of brain regions of interest (ROIs) for a subject. Very often reference centile curves show the distribution of a measurement such as cortical thickness or amplitude of low frequency fluctuations (ALFF) is age-dependent. We utilize Box-Cox transformation to normal for the age-dependent distributions, which are indexed by three curves representing the median, coefficient of variation and skewness. The three curves can be fitted by penalized splines via the GAMLSS method of Rigby and Stasinopoulos. The age-dependent distributions at each ROI are first obtained from some training subjects, and then applied to compute the region-wise quantile map for future subjects according to their ages. The brain quantile map of a subject can be potentially used for clinical diagnosis of a disease and outlier detection in data quality control, since it is based on age-specific individual's image-derived quantities. Moreover, we propose bootstrap based inference to obtain the confidence intervals for the quantile estimates. The proposed methods are applied to the Nathan S. Kline Institute Rockland lifespan sample, in which both structural and functional MRIs are collected for 150 subjects with age ranging from 7 to 85.

email: huaihou.chen@nyumc.org

### **Latent Variable Models for Longitudinal MR Imaging Data with Multiple Outcomes**

**Xiao Wu\***, University of Florida  
**Michael J. Daniels**, University of Texas, Austin

MR imaging analysis involves combining multiple measures of interest over time. In this research, we describe a Bayesian approach to model longitudinal multidimensional continuous outcomes with latent variables. Dependence among different outcomes is induced through latent variables and covariance matrices are estimated simultaneously by using nonparametric priors. A Markov chain Monte Carlo algorithm is proposed for estimating the posterior distributions of the parameters and latent variables. This method is illustrated using data from a Duchenne Muscular Dystrophy study on changes in muscle imaging data to capture disease progression over time.

email: xiaowu@ufl.edu

### **A Novel Brain Connectivity Network Model: Build Bridges Between Network Communities**

**Shuo Chen\***, University of Maryland, College Park

Current scientific research reveals that most brain functions such as emotion, reasoning, and cognition are associated with complex brain networks constructed of clusters of distinct brain regions. Most current brain connectivity network analysis methods assign each brain region to a unique cluster. However, there exist well-connected brain regions bridging multiple brain region clusters. We develop

a novel framework to account for the bridge brain regions that belong to multiple brain region clusters and a likelihood based evaluation method for model comparison. We demonstrate the performance of the new method using simulated data set and a fMRI data set in emotion-cognition study.

email: chenshuochen@gmail.com

### **A Bayesian Model for Brain Activation and Connectivity**

**Zhe Yu\***, University of California, Irvine  
**Hernando Ombao**, University of California, Irvine  
**Raquel Prado**, University of California, Santa Cruz  
**Erin Burke**, University of California, Irvine  
**Steve Cramer**, University of California, Irvine

To study activation and connectivity in the brain, we develop a Bayesian approach for modeling fMRI data. Our approach simultaneously estimates local hemodynamic response function (HRF), local activation, and effective and functional connectivity in a brain network. Existing methods assume HRFs to be identical across brain regions which may lead to erroneous conclusions in activation and connectivity. Our approach addresses this limitation by estimating region-specific HRF. Additionally, our approach will enable neuroscientists to compare effective connectivity between experiment conditions. Furthermore, the use of spike and slab prior makes it straightforward to select effective connectivities. Finally, Bayesian paradigm allows inclusion of knowledge based on neurobiology and empirical results acquired from prior studies. The simulation study demonstrates that, compared to the standard GLM approach, our approach has generally higher power and credible interval coverage, lower type I error and bias, and the ability to capture connectivity difference across conditions; while the standard approach indeed produced erroneous results on activation and connectivity. We also applied our approach to a dataset from a stroke study, and observed interesting findings in HRF, activation and connectivity that were usually not observed from healthy subjects.

email: zhey@uci.edu

### **Pre-Surgical fMRI Data Analysis Using a Spatially Adaptive Conditional Autoregressive Model**

**Zhuqing Liu\***, University of Michigan  
**Veronica J. Berrocal**, University of Michigan  
**Timothy D. Johnson**, University of Michigan

Spatial smoothing is an essential step in the analysis of functional magnetic resonance imaging (fMRI). One standard smoothing method is to convolve the image data with an isotropic Gaussian kernel, which applies a fixed amount of smoothing to the entire image. In pre-surgical brain image analysis where spatial accuracy is paramount, this method, however, is not reasonable as it may cause some regions to be undersmoothed while others oversmoothed, smearing out the boundaries of activation and deactivation regions of the brain. To this end, we propose a novel spatially adaptive intrinsic conditional autoregressive (SACAR) model

with smoothing variances proportional to error variances, allowing the degree of smoothing to vary across the brain. We compare our proposed model with two existing spatially adaptive models and with a Bayesian non-parametric Potts model (Johnson et al., 2011). Simulations studies show our model outperforms these other models; as a real model application, we apply it to pre-surgical fMRI data.

email: zhuqingl@umich.edu

### **Spatial and Temporal Pattern in the Brain Accounting Cognitive Changes after Mild Traumatic Brain Injury**

**Namhee Kim\***, Albert Einstein College of Medicine  
**Craig A. Branch**, Albert Einstein College of Medicine  
**Michael L. Lipton**, Albert Einstein College of Medicine

While most Mild Traumatic Brain Injury (mTBI) patients recover over several months, up to one third have enduring disability with non-specific symptoms and cognitive impairment, e.g. headache, impaired memory and attention. Prediction of mTBI patients with long-term cognitive disability has been an important clinical issue, and we in this study aimed to build a prediction model with longitudinal Fractional Anisotropy (FA) image and cognitive data. A multilinear partial least squares algorithm was adopted to tackle complexity of the data, and to explain the relationship between cognitive changes and underlying longitudinal FA changes in the brain. 16 mTBI patients underwent 6 domains of cognitive tests and MR imaging at each assessment occurred at 1 week, 3 months, and 6 months after injury. As results, a pair of spatial and temporal pattern from longitudinal FA images, which has maximum association with longitudinal cognitive changes, was estimated. Significant areas in the estimated spatial pattern included the brain areas associated with the cognition tested. We also evaluated the prediction power of the proposed model by leave-one-out cross-validation.

email: namhee.kim@einstein.yu.edu

## **59. SEMI-PARAMETRIC AND NON-PARAMETRIC MODELS IN SURVIVAL ANALYSIS**

### **Semiparametric Bayes Estimation of Gap-Time Distribution with Correlated Recurrent Event Data**

**AKM F. Rahman\***, University of South Carolina, Columbia  
**Edsel A. Pena**, University of South Carolina, Columbia

Recurrent event data arise from a wide variety of studies/ fields such as clinical trials, epidemiology, public health, bio-medicine (e.g. repeated heart attack, repeated tumor occurrences of a cancer patient). Semiparametric Bayes inference of the gap-time survivor function with the effect of covariates of a correlated recurrent event in the presence of censoring is considered. A frailty model is considered to allow the association between inter-occurrence gap-times. We assume that for a subject or unit given the unobserved frailty variable  $Z=z$ , the inter-occurrence gap-times are IID with some distribution function  $F(\cdot|z)$ . In our procedure, we assign a Gamma process prior on the baseline cumulative

hazard function and parametric prior distributions on the finite dimensional parameters associated with covariates and the frailty random variable. We derive the conditional posterior distributions from the joint posterior distribution of the unknown parameters of interest. From these conditional posterior distributions we obtain the closed form posterior means which are our Bayes estimators. However, to construct approximate credible intervals, we employ the Gibbs sampler techniques to obtain samples from the joint posterior distribution. Simulation studies demonstrate the effectiveness of the developed method.

email: rahmana@email.sc.edu

### **Quantile Regression Models for Current Status Data**

**Fang-Shu Ou\***, University of North Carolina, Chapel Hill  
**Donglin Zeng**, University of North Carolina, Chapel Hill  
**Jianwen Cai**, University of North Carolina, Chapel Hill

Current status data arise frequently in demography, epidemiology, and econometrics where the exact failure time cannot be determined but is known only to have occurred before or after a random observation time. We propose a quantile regression model to analyze current status data because it relaxes the requirements on the error term and the coefficients are interpretable as direct regression effects on the failure time. Our model assumes that the conditional quantile of failure time is a linear function of covariates. We assume the conditional independence between the failure time and observation time. An M-estimator is developed for parameter estimation and the asymptotic distribution for the estimator is derived. The estimator is computed using the convex-concave procedure, and its confidence intervals are constructed using a subsampling method. The small sample performance of the proposed method is demonstrated via simulation studies. Finally, we apply the proposed method to analyze data from the Voluntary HIV-1 Counseling and Testing Efficacy Study Group.

email: fou@unc.edu

### **Competing Risks Regression Under Random Signs Censoring**

**Jonathan Yabes\***, University of Pittsburgh  
**Joyce Chang**, University of Pittsburgh

Many clinical trials and cohort studies collect failure time data where the participants are at risk of several mutually exclusive events or failure types known as competing risks. The cumulative incidence function (CIF) is commonly reported in such studies. In many situations however, investigators are interested in the marginal survival distribution of latent failure times, rather than the CIF. Because of the identifiability problem in competing risks, we derived an estimator of covariate effects in the Cox proportional hazards model by incorporating the random signs censoring (RSC) principle, which assumes that the main event failure time is independent of the indicator that the main event precedes the competing event. Unlike identifying assumptions that are typically imposed in practice, RSC is verifiable via stochastic ordering in the observed data. We further relaxed the RSC assumption by positing that independence is achieved conditional on some

covariates. We showed that the resulting estimator is not only easy to implement but also has desirable asymptotic properties. We evaluated the estimator's finite sample size performance through simulations. Medical datasets were used to illustrate the proposed methods.

email: jgy2@pitt.edu

### **Regression Analysis of Informatively Interval-Censored Failure Time Data with Cox Model**

**Ling Ma\***, University of Missouri, Columbia

**Tao Hu**, Capital Normal University, China

**Jianguo Sun**, University of Missouri, Columbia

The statistical analysis of interval-censored failure time data has recently attracted a great deal of attention and especially, many procedures have been proposed for their regression analysis under various models. In this paper, we discuss the regression problem in the presence of informative interval censoring, which occurs quite often in practice but for which only limited literature exists. More specifically, motivated by clinical studies, we consider the situation where the presence of the informative censoring is caused by the correlation between the failure time of interest and the interval length. Furthermore, the correlation and the covariate effects can be described by a copula model and the proportional hazards model, respectively. For estimation, we develop a sieve maximum likelihood estimation procedure with the use of monotone l-spline functions and the resulting estimators are shown to be consistent. Furthermore, the estimated regression parameters are asymptotically normal and semiparametrically efficient. The proposed method is examined through a simulation study and illustrated with the data from a well-known breast cancer study.

email: mlbegood@gmail.com

### **Weighted Estimation of the Accelerated Failure Time Model in the Presence of Dependent Censoring**

**Youngjoo Cho\***, The Pennsylvania State University

**Debashis Ghosh**, The Pennsylvania State University

Independent censoring is one of the crucial assumptions in survival analysis. However, this is impractical in many medical studies, where the presence of dependent censoring leads to difficulty in analyzing covariate effects on disease outcomes. The semicompeting risks framework proposed by Lin et al. (1996) and Peng and Fine (2006) offers one approach to handling dependent censoring. These authors proposed estimators based on an artificial censoring technique. However, they did not consider efficiency of their estimators in detail. In this paper, we propose a new weighted estimator for the accelerated failure time (AFT) model under dependent censoring. One of the advantages in our approach is that these weights are optimal among all the linear combinations of the previously mentioned two estimators. To calculate these weights, a novel resampling-based scheme is employed. Attendant asymptotic statistical results for the estimator are established. In addition, simulation studies, as well as an application to real data, show the gains in efficiency for our estimator.

email: yvc5154@psu.edu

### **Model Assisted Cox Regression**

**Shoubhik Mondal\***, New Jersey Institute of Technology  
**Sundarraman Subramanian**, New Jersey Institute of Technology

Semiparametric random censorship (SRC) models (Dikta, 1998), derive their rationale from their ability to gainfully utilize parametric ideas within the random censorship environment. An extension of this approach is developed for Cox regression, producing new estimators of the regression parameter and baseline cumulative hazard function. Under correct parametric specification, the proposed estimator of the regression parameter is shown to be asymptotically more efficient than the standard partial likelihood estimator. Numerical studies are presented to showcase the efficacy of the proposed approach even under significant misspecification. A real example is provided. A further extension to the case of missing censoring indicators is also developed and an illustration with pseudo-real data is provided.

email: sm485@njit.edu

## **60. HIERARCHICAL MODELS**

### **Examining the Spatio-Temporal Trend Between Alcohol Outlets and Violence Using Integrated Nested Laplace Approximations**

**Loni P. Tabb\***, Drexel University

**Tony H. Grubestic**, Oregon State University

The distribution of alcohol outlets has been long linked to violent crime, particularly in urban areas. Privatization removes state control on alcohol sales, and the effect of privatization has been shown to alter this relationship between alcohol outlets and various alcohol-related public health issues. Research on alcohol outlets, though, commonly involves a cross-sectional setting; therefore, limiting the possibility of investigating temporal trends, especially in the presence of policy change. The purpose of this paper is to present the use of integrated nested Laplace approximations to examine the spatio-temporal distribution of alcohol outlets before and after privatization in Seattle, Washington, 2010-2012. Using census block groups, we are able to analyze the patterns of alcohol outlets, as well as characterize the census block group variation via geovisualization methods.

email: lpp22@drexel.edu

### **The Role of Prior Effective Sample Size in the Design of Bayesian Medical Device Studies**

**Gene A. Pennello**, U.S. Food and Drug Administration

**Laura Thompson\***, U.S. Food and Drug Administration

When evaluating the design of a Bayesian medical device study, stakeholders (clinicians, statisticians, regulatory authorities) are often interested in the effective sample size implied by the prior distribution. Several definitions of prior effective sample size have been proposed. We present a definition based on the idea that the posterior and prior variances of a parameter are roughly proportional to sample

size, taking the expectation of the posterior variance over the distribution of the data yet to be observed. We provide an intuitive approximation that is relatively easy to compute and show that this approximation yields reasonable answers for the prior sample size in commonly assumed likelihood-prior combinations. We further show how prior effective sample size can be utilized to design Bayesian studies by simply subtracting it from the usual sample size requirement when no prior information is available. To illustrate, we show how one might design a Bayesian interim analysis plan for a study that borrows strength from multiple prior studies in a hierarchical model. The plans operating characteristics will be presented. Compared with a study without prior information, the plan provides a considerable savings in the sample size necessary to stop the study and draw conclusions.

email: gene.pennello@fda.hhs.gov

### **A Hybrid Bayesian Hierarchical Model Combining Cohort and Case-Control Studies for Meta-Analysis of Diagnostic Tests: Accounting for Disease Prevalence and Partial Verification Bias**

**Xiaoye Ma\***, University of Minnesota

**Yong Chen**, University of Texas

**Stephen Cole**, University of North Carolina, Chapel Hill

**Haitao Chu**, University of Minnesota

Bivariate random effects models have been recommended to jointly model sensitivities and specificities in meta-analysis of diagnostic accuracy studies accounting for between-study heterogeneity. Because the severity and definition of disease may differ across studies due to the design and study populations, the sensitivities and specificities of a diagnostic test may depend on disease prevalence. To account for the potential dependence, trivariate random effects models have been recently proposed. However, the proposed approach can only include cohort studies to estimate study-specific disease prevalence. In addition, some diagnostic accuracy studies only select a subset of samples based on the test results to be verified by the reference test. It is known that ignoring unverified subjects can lead to partial verification bias in the estimation of prevalence and test accuracy in a single study. However, the impact of this bias on the meta-analysis of diagnostic tests has not been investigated. As many diagnostic accuracy studies use case-control designs, we propose a novel hybrid Bayesian hierarchical model combining cohort and case-control studies to account for prevalence and to correct partial verification bias at the same time. We conduct a set of simulation studies, and present a case study on assessing the diagnostic accuracy of MRI in detecting lymph node metastases.

email: maxxx372@umn.edu

### **Group Comparison of Pulsatile Hormone Times Series**

**TingTing Lu\***, University of Michigan

**Timothy D. Johnson**, University of Michigan

Due to its oscillatory and pulsatile nature, analyzing hormone time series data is challenging and many model-based methods have been proposed over the years. Typically, analyses are performed in two stages. First, the number and locations of the episodic events are determined. Second, a model is fit to the data conditional on the number of pulses. However, errors occurring in the first step are carried over to second. In 2007, Johnson proposed the first fully Bayesian deconvolution model that jointly estimates both the number and locations of secretion events and admits a non-constant basal concentration. Thus, both pulsatile and oscillatory components of hormone secretion are simultaneously modeled. Furthermore, the model allows for variation in pulse, shape and size. However, the model cannot handle groups of subject and cannot compare secretion patterns between groups. In this paper we extend Johnson's model in two ways. First, we admit group comparisons of the underlying pulse driving mechanism; second, we model the pulse driving mechanism via a Cox process where the intensity function is not assumed constant as is assumed in Johnson (2007). We take a fully Bayesian hierarchical approach to estimate model parameters, then compare results with a smoothing spline functional analysis approach.

email: ttlu@umich.edu

### **Population Size Estimation with Inactive Lists: Hierarchical Mixture Models and Missing Data with Application to Armed Conflict Data**

**Shira Mitchell\***, Harvard University

**Al Ozonoff**, Harvard University

**Kristian Lum**, Virginia Polytechnic Institute and State University

**Alan M. Zaslavsky**, Harvard University

**Brent A. Coull**, Harvard University

Since 1964, tens of thousands of people have died in Colombia's armed conflict. Underreporting violence obscures the nature of the conflict, precluding development of effective solutions. We develop hierarchical log-linear capture-recapture models to estimate the number of armed conflict killings that occurred in Casanare, Colombia in the years 1998-2007. Lack of data the early years motivates the use of hierarchical models that borrow strength across time. We investigate two methods to handle groups actively collecting data in different but overlapping time-periods. One fills in the inactive periods, treating the counts in those years as missing data. Another does not, instead incorporating the inactivity into the model. We compare these, as well as hierarchical versus unpooled models. A simulation study shows that the Bayesian hierarchical models have shorter confidence interval width, with similar or better coverage than the unpooled models, with robustness to the exchangeability assumption. They enable us to obtain useful intervals for the number of killings in the early years, where there are less data, so we can look at trends across time that guide political analysis of the conflict. We provide guidance for capture-recapture studies with inactive lists and recommend the use of hierarchical modeling in capture-recapture.

email: sam942@mail.harvard.edu

## Bayesian Hierarchical Joint Modeling of Repeatedly Measured Continuous and Ordinal Markers of Disease Severity

**Olive D. Buhule\***, University of Pittsburgh  
**Abdus S. Wahed**, University of Pittsburgh  
**Ada O. Youk**, University of Pittsburgh

In this paper, we propose a joint model to analyze multivariate repeatedly measured outcomes of mixed types, in particular, continuous and ordinal outcomes. The postulated model assumes that the outcomes are from distributions that are in the exponential family and hence modeled as a multivariate generalized linear mixed effects model linked through correlated and/or shared random effects. The MCMC Bayesian approach is used to approximate posterior distribution and draw inference on the parameters. This proposed joint model provides a flexible framework to account for the hierarchical structure of the highly unbalanced data as well as the association between the multiple mixed type outcomes. Moreover, the simulation studies show that estimates obtained from the joint model are consistently less biased and more efficient than those in the separate models.

email: odb3@pitt.edu

## Hierarchical Nearest-Neighbor Gaussian Process Models for Massive Geostatistical Datasets

**Abhirup Datta\***, University of Minnesota  
**Sudipto Banerjee**, University of Minnesota  
**Andrew O. Finley**, Michigan State University

Use of hierarchical spatial process models to analyze geo-referenced datasets has become increasingly popular over the last decade. However, Markov Chain Monte Carlo (MCMC) techniques used for implementing these models inherit heavy calculations. Decompositions of spatial correlation matrices for  $n$  datapoints involve operations of the order  $n$ -cubed thereby slowing the algorithms beyond any pragmatic threshold for large datasets. We propose the hierarchical Nearest Neighbor Gaussian Process (NNGP) model for multivariate spatial responses that uses strong spatial correlation between neighbors to drastically reduce computational complexity. We provide a MCMC algorithm for estimation and prediction avoiding any storage or decompositions of large matrices. We show that the number of matrix operations for this algorithm is linear in  $n$  thereby making it scalable to massive datasets. We also show that this class of models has true likelihoods and hence can be evaluated using any standard model comparison tools. We use simulated datasets to illustrate its huge computational benefits and performance superiority over other candidate models. Finally, we analyze a massive forestry dataset using NNGP model to infer about the spatial distribution of forest biomass in United States.

email: datta013@umn.edu

## 61. METHODS FOR REMOVING SELECTION BIAS AND CONFOUNDING

### Stable Weights that Balance Covariates for Causal Inference and Estimation with Incomplete Data

**Jose Zubizarreta\***, Columbia University

Weighting methods that adjust for observed covariates are widely used for causal inference and estimation with incomplete data. Part of the conceptual appeal of such weighting methods is that one set of weights can be used to estimate a range of treatment effects or to estimate the mean of a variety of outcomes. However, this appeal is eclipsed by the instability of the estimated weights and by the difficulty in some settings of adequately adjusting for observed covariates. This paper presents a new weighting method that overcomes these difficulties. Specifically, by solving a convex optimization problem, this method finds the weights of minimum variance that adjust or balance the empirical distribution of the observed covariates up to levels prespecified by the researcher. Conceptually, this method is based on a well-defined optimization problem that can be solved in polynomial time, permitting to handle relatively large data sets in useful time. In practice this method is easy to use as currently implemented in the new sbw package in R. This paper shows some theoretical properties of the resulting weights and illustrates their use analyzing both a real data set from the 2010 Chilean earthquake and a simulated example.

email: jz2313@columbia.edu

### Matching Using Propensity Score Methods for Time-Varying Treatments

**Pallavi S. Mishra-Kalyani\***, Emory University  
**Brent A. Johnson**, Emory University  
**Qi Long**, Emory University

Analysis of treatment effect among propensity score matched pairs has long been considered a method for removing bias. However, the methodology for utilizing propensity score matching when treatment is administered at various inconsistent times throughout an individual's follow-up period is relatively unexplored. Though some have examined varying time of treatment, these analyses have included only data with even intervals of measurement of time-varying covariates. To account for the varying times of treatment, we propose matching subjects for analysis by using extensions of the standard propensity scores including Generalized Propensity Scores and P-Functions, both incorporating time-varying covariates in an effort to remove bias caused by disease progression over time. Matched pairs will be compared using non-parametric methods, specifically the Wilcoxon Signed-Rank test and rank regression.

email: psmishr@emory.edu

### Estimating Causal Effects in an Observational Study with a Survival Time Endpoint

**Jaeun Choi\***, Harvard Medical School  
**Mary Beth Landrum**, Harvard Medical School  
**A. James O'Malley**, Dartmouth College  
**Bruce Landon**, Harvard Medical School

Estimation of the effect of a treatment in the presence of unmeasured confounding is a common objective in observational studies. The Two Stage Least Squares (2SLS) Instrumental Variables (IV) procedure is frequently used but is not applicable to time-to-event data if some observations are censored. We develop a simultaneous equations model (SEM) to account for unmeasured confounding of the effect of treatment on a survival outcome subject to censoring. Specifically, through joint modeling, we estimate the survival function in the presence of unmeasured confounding while simultaneously accounting for missing survival times. The identification of the treatment effect is assisted by IVs that are related to treatment but conditional on treatment do not directly affect the outcome and the assumed bivariate distribution underlying the data generating process. As the IV and the distributional assumptions cannot be jointly assessed from the observed data, we consider methods for evaluating the sensitivity of the results to these assumptions. The methodology is illustrated on an observational study of time to death following endovascular or open repair of abdominal aortic aneurysm (AAA).

email: [choi@hcp.med.harvard.edu](mailto:choi@hcp.med.harvard.edu)

### Weighting to Strengthen an Instrumental Variable

**Doug Lehmann\***, University of Michigan  
**Yun Li**, University of Michigan  
**Yi Li**, University of Michigan

Instrumental variable (IV) methods have been widely used to control for unmeasured confounding. Groups of patients (e.g., by centers, physicians, or service areas) are often proposed as instruments based on the idea that differential medical practices attributable to these groups are independent of unmeasured confounders for the main outcome of interest after conditioning on medically relevant covariates. These can be weak instruments, however, and as such are sensitive to IV assumptions and can lead to biased results. Near far matching was recently proposed as a matching based IV methodology that strengthens the instrument by removing certain subjects from the analysis. We propose a variation of near far matching that weights pairs based on the instrument rather than remove them entirely. Through simulation we show that this approach effectively strengthens the instrument, leaving it more robust towards violations of the assumption that the instrument is randomly assigned. Additionally, since no subjects are removed, better matches can be obtained. These methods are illustrated using Medicare data to study patient outcomes at kidney dialysis facilities.

email: [lehmannnd@umich.edu](mailto:lehmannnd@umich.edu)

### Propensity Score Bin Bootstrapping Method in Estimation of Cost-Effectiveness

**Zugui Zhang\***, Christiana Care Health System  
**Paul Kolm**, Christiana Care Health System  
**William S. Weintraub**, Christiana Care Health System

The aim of this study was to apply propensity score bin bootstrapping method in estimation of cost-effectiveness of coronary-artery bypass grafting (CABG) versus percutaneous coronary intervention (PCI), using data from the Society of Thoracic Surgeons Database and the American College of Cardiology Foundation (ACCF) National Cardiovascular Data Registry in ASCERT. The STS Database and ACCF NCDR were linked to the Centers for Medicare and Medicaid Services (CMS) claims data from years 2004 to 2008, and patients (86,244 in CABG group and 103,549 in PCI group) at least 65 years old with two or three vessel coronary artery disease were included in the study. Cost-effectiveness is expressed as the incremental cost effectiveness ratio (ICER), the difference in costs of the two forms of coronary revascularization divided by the difference in effectiveness (events prevented, life years gained or QALYs). To reduce the treatment selection bias in this observational study, propensity score bin bootstrapping methods (10,000 replicates) were used to estimate 95% confidence intervals for the mean differences of cost and effectiveness between the two strategies. The results showed that using a common threshold such as \$50,000 or \$100,000 per QALY gained, CABG will have provided moderate to high probability of better clinical benefit, which means that CABG will often be a cost-effective strategy.

email: [zhang@christianacare.org](mailto:zhang@christianacare.org)

### Maximum Likelihood Adjustment for Mis-Measured Exposure Using External Validation Data and Propensity Scores

**Danielle Braun\***, Harvard School of Public Health and Dana-Farber Cancer Institute  
**Malka Gorfine**, Israel Institute of Technology  
**Corwin Zigler**, Harvard School of Public Health  
**Francesca Dominici**, Harvard School of Public Health  
**Giovanni Parmigiani**, Harvard School of Public Health and Dana-Farber Cancer Institute

Propensity score methods are widely used to analyze observational studies in which patient characteristics might not be balanced by treatment group. These methods assume that exposure, or treatment assignment, is error-free, but in reality these variables can be subject to measurement error. This arises in the context of comparative effectiveness research, in which accurate procedural codes are not always available. When using propensity score based methods, this error affects both the exposure variable directly, as well as the propensity score. We propose a two step maximum likelihood approach using validation data to adjust for the measurement error. First, we use a likelihood approach to estimate an adjusted propensity score. Using the adjusted propensity score, we then use a likelihood approach on the outcome model to adjust for measurement error in the exposure variable directly. Simulations show our proposed approach reduces bias in treatment effect estimates, and improves covariate balance across treatment groups.

We illustrate our method on a comparative effectiveness research study assessing the use of surgery to remove a brain tumor in an elderly population diagnosed with glioblastoma. We use SEER-Medicare as our validation data, and apply our method to Medicare Part A data.

email: dbraun@hsph.harvard.edu

### **Examination of Statistical Power in a Propensity Score Analysis Approach**

**Falynn C. Turley\***, University of Alabama, Birmingham  
**David Redden**, University of Alabama, Birmingham

Propensity score analysis is commonly used in applied statistics and medical research. Formally, propensity score analysis begins with calculating a propensity score by logistic regression for each subject with treatment as the outcome variable, resulting in the conditional probability of receiving a treatment given a set of observed variables, which is then used to balance the data through matching, stratification, or covariance adjustments. The consequences of retaining the matching variable in a propensity score analysis have not been fully established. Thus, this research attempts to (i) define a value that serves as an indicator of how well matching was carried out on the propensity scores and (ii) formulate a power calculation approach when using propensity score analysis, specifically when matching is the method used to compare the two groups. Because overall sample size is affected differently in each scenario and a matched sample changes the distribution of interest, the power calculation is dependent on the choice of the matching algorithm. The consequences of ignoring the propensity score matching with regard to Type I error rate and power will be discussed.

email: falynn@uab.edu

## **62. FUNCTIONAL DATA ANALYSIS**

### **Structured Functional Principal Component Analysis**

**Haochang Shou\***, Johns Hopkins  
Bloomberg School of Public Health  
**Vadim Zipunnikov**, Johns Hopkins  
Bloomberg School of Public Health  
**Ciprian M. Crainiceanu**, Johns Hopkins  
Bloomberg School of Public Health  
**Sonja Greven**, Ludwig-Maximilians-Universitat, Germany

Motivated by modern observational studies, we introduce a class of functional models that expands nested and crossed designs. These models account for the natural inheritance of correlation structure from sampling design in studies where the fundamental sampling unit is a function or image. Inference is based on functional quadratics and their relationship with the underlying covariance structure of the latent processes. A computationally fast and scalable estimation procedure is developed for ultra-high dimensional data. Methods are illustrated in two examples: high-frequency accelerometer data for daily activity, and pitch linguistic data for phonetic analysis.

email: haochang.shou@gmail.com

### **A Robust Approach for Functional Linear Regression Model**

**Yihong Zhao\***, New York University Medical Center  
**R. Todd Ogden**, Columbia University Medical Center  
**Huaihou Chen**, New York University Medical Center

One useful approach for fitting linear models with scalar outcomes and functional predictors involves transforming the functional data to the wavelet domain and converting the data fitting problem to a variable selection problem. Applying the LASSO procedure in this situation has been shown to be efficient and powerful. In this study we explore possible directions for improvements to this method. The finite sample performance of the proposed methods will be compared through simulations and real data applications in mental health research. We believe applying these procedures can lead to improved estimation and prediction as well as better stability.

email: zhaoy05@nyumc.org

### **Nonlinear Functional Regression Models with Application to Copy Number Data**

**Adrian Coles\***, North Carolina State University  
**Arnab Maity**, North Carolina State University  
**Ganiraju Manyam**, University of Texas  
MD Anderson Cancer Center  
**Veerabhadran Baladandayuthapani**, University of Texas  
MD Anderson Cancer Center

Genomic abnormalities in the number of copies of DNA are associated with the development and progression of many human disorders such as cancer. One key scientific objective is to detect local genomic regions using the copy number profiles that are associated with clinical outcomes that indicate disease development and progression. To this end, we propose a flexible nonlinear functional regression framework to detect associations between continuous outcomes and a functional covariate, e.g. the DNA copy number profile. In contrast to classical functional regression models where the effect of the functional covariate is modeled using a pre-specified structure, e.g., linear or quadratic functionals, we model the functional effect nonparametrically. We develop estimation and score test procedures for the functional effect. We investigate the finite sample performance of our procedures via simulation studies and illustrate our procedure in an analysis of a multiple myeloma data set. Simulation results show that our procedure outperforms classical models when the functional effect is nonlinear. Analysis of cancer genomics data identifies several significant regions of copy number alterations containing genes with known implications in the etiology of multiple myeloma as well as other cancers.

email: alcoles@ncsu.edu

### Generalized Functional Linear Models for Case-Control Association Studies

**Ruzong Fan**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**Yifan Wang\***, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**James L. Mills**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**Iryna Lobach**, University of California, San Francisco

**Momiao Xiong**, University of Texas, Houston

Generalized functional linear models are developed for testing associations between dichotomous traits and multiple genetic variants in a genetic region. Although the observed multiple genetic marker data are discrete, we view them as realizations of stochastic genetic variant functions. By using functional data analysis technique, the observed high dimension genetic variant data are used to estimate the genetic variant functions based on B-spline or Fourier basis functions or functional principal component decompositions. The estimated genetic variant functions are used in logistic models to connect multiple genetic variants to phenotype adjusting for covariates. After extensive simulation analysis, it is shown that the Rao's efficient score tests of the proposed fixed models are very conservative since they generate low type I errors, and global tests of the mixed models are very robust. The Rao's score test statistics of the proposed fixed models have higher or similar power as sequence kernel association test (SKAT) and its optimal unified version (SKAT-O) for a scenario that the causal variants are both rare and common. When the causal variants are all rare, the Rao's score test statistics or the global score tests have similar or slightly lower power as SKAT and SKAT-O. The methods can be used in both gene-based genome-wide/exome-wide association studies or candidate gene analysis.

email: yifan.wang@nih.gov

### Wavelet-Based Function-On-Function Mixed Models

**Mark J. Meyer\***, Harvard University

**Brent A. Coull**, Harvard University

**Francesco Versace**, University of Texas MD Anderson Cancer Center

**Jeffrey S. Morris**, University of Texas MD Anderson Cancer Center

Medical and public health research increasingly involves the collection of more and more complex and high dimensional data. In particular, functional data---where the unit of observation is a curve or set of curves that are finely sampled over a grid---is frequently obtained. Moreover, researchers often sample multiple curves per person resulting in repeated functional measures. A common question is how to analyze the relationship between two

functional variables. We propose a general function-on-function regression model for repeatedly sampled functional data, presenting both one and two sample settings along with a Bayesian inference procedure. We examine these models via simulation and a data analysis. Motivating data come from a neurological study examining how the brain processes various types of images. Subjects were taken from a pre-screening for a smoking cessation trial. Event-related potentials were then measured after presentation of neutral, positive, negative, and cigarette-related images. Our analysis focuses on the relationship between measurements from pairs of sensors during neutral and cigarette-related image presentation. The resulting analyses suggests that the association between adjacent sensors varies with time, with stronger associations immediately after exposure to the image, but no evidence that this time-varying association differs by image type.

email: mjmeier@fas.harvard.edu

### A Computational Framework for Genetic Mapping of Heterochrony

**Han Hao\***, The Pennsylvania State University

**Ningtao Wang**, The Pennsylvania State University

**Yaqun Wang**, The Pennsylvania State University

**Jianxin Wang**, Beijing Forestry University

**Zhong Wang**, Beijing Forestry University

**Rongling Wu**, The Pennsylvania State University

Heterochrony, the phylogenetic change in the time of developmental events or rate of development, has been thought to play an important role in producing phenotypic novelty during evolution. Increasing evidence suggests that specific genes are implicated in heterochrony, guiding the process of developmental divergence. Here we present a computational framework for genetic mapping by which to characterize and locate quantitative trait loci (QTLs) that govern heterochrony described by three parameters, i.e., when growth starts, growth rate, when growth stops, during organ development. The framework was derived from a dynamic model, functional mapping, aimed to map QTLs for the overall process and pattern of development. By integrating an optimality algorithm, the framework allows the so-called heterochrony QTLs (hQTLs) to be tested and quantified. Specific pipelines are given for testing how hQTLs control the onset and offset of developmental events, the rate of development, and duration of a particular developmental stage. As an example, the framework is used to map hQTLs for plant height growth in rice. Genetic variation in the timing of maximum growth rate and duration of linear growth stage is identified as one mechanism controlling adult height growth of rice and, therefore, as one possible basis for heterochrony to evolve.

email: haohan421@gmail.com

## 63. RECENT ADVANCES IN BAYESIAN METHODS

### **Incorporating Spatial Dependence into Bayesian Multiple Testing of Statistical Parametric Maps in Functional Neuroimaging**

**Andrew Brown\***, Clemson University

**Nicole A. Lazar**, University of Georgia

**Gauri S. Datta**, University of Georgia

**Woncheol Jang**, Seoul National University

**Jennifer E. McDowell**, University of Georgia

The analysis of functional neuroimaging data often involves the simultaneous testing for activation at thousands of voxels, leading to a massive multiple testing problem. This is true whether the data analyzed are time courses observed at each voxel or a collection of summary statistics such as statistical parametric maps (SPMs). It is known that classical multiplicity corrections become strongly conservative in the presence of a massive number of tests. Some more popular approaches for thresholding imaging data, such as the Benjamini-Hochberg step-up procedure for false discovery rate control, tend to lose precision or power when the assumption of independence of the data does not hold. Bayesian approaches to large scale simultaneous inference also often rely on the assumption of independence. We introduce a spatial dependence structure into a Bayesian testing model for the analysis of SPMs. By using SPMs rather than the voxel time courses, much of the computational burden of Bayesian analysis is mitigated. Increased power is demonstrated by using the dependence model to draw inference on a real dataset collected in a fMRI study of cognitive control. The model also is shown to lead to improved identification of neural activation patterns known to be associated with eye movement tasks.

email: ab7@clemson.edu

### **Multivariate Bayesian Censored Models for Predicting Exposure to Multiple Chemical Agents**

**Caroline Groth\***, University of Minnesota

**Sudipto Banerjee**, University of Minnesota

**Tran Huynh**, University of Minnesota

**Gurumurthy Ramachandran**, University of Minnesota

**Richard Kwok**, National Institute of Environmental Health Sciences, National Institutes of Health

**Mark Stenzel**, Exposure Assessment Applications, LLC

**Patricia Stewart**, Stewart Exposure Assessments, LLC

In several public health settings pertaining to exposure modeling, there is a need for analyzing left censored measurements. This presentation formulates and applies a class of multivariate Bayesian censored models for predicting exposure to multiple chemical agents found in crude oil. Our inferential framework aims to better capture correlations between multiple chemical agents, while accounting for various explanatory and design variables. Due to several of the agents having a high percentage of measurements below the measuring instrument's limits of detection (10%-90%), we employ methods for dealing with censored data in multivariate settings. We seek full and exact inference

from the posterior distributions of model parameters in a computationally feasible manner using the BUGS language. We apply our methods to air measurement data from the 2010 Deepwater Horizon Oil Spill.

email: groth203@umn.edu

### **Methods in Functional Data Analysis for Curve Comparison in Spectroscopic Protein Unfolding Data: Applications Using Bayesian Inferential Methods**

**Miranda L. Lynch\***, University of Connecticut Health Center

Many experiments probing protein structure and stability proceed by gathering various types of spectroscopic output, resulting in complicated profiles of protein behavior under different experimental conditions. Spectroscopic profiles can be handled using functional data analytic methods, as the discretely measured curves represent continuous functional output. This work presents methods for curve comparison and feature determination of protein unfolding data from spectroscopic analyses using Bayesian spline-based nonparametric fits. The penalized spline fit for each profile is carried out using MCMC techniques to sample from relevant posterior distributions. The primary method discussed is a Bayesian approach to estimating first- and higher-order derivative functions of the resulting smoothed curves, and subsequent use of derivative estimates to carry out the desired comparisons. Methods for derivative estimation for functional data have not been well examined in a Bayesian inferential framework. Such methods provide novel tools for investigating and comparing curve features that allow for informative prior specification and Bayesian methods for smoothing parameter selection. The proposed methods are exemplified using circular dichroism spectroscopic data from a transcriptional regulatory protein, as well as with simulated data.

email: mlynch@uchc.edu

### **Bayesian Variable Selection for High Dimensional Datasets in the Presence of Error-Prone Time-to-Event Outcomes**

**Xiangdong Gu\***, University of Massachusetts, Amherst

**Raji Balasubramanian**, University of Massachusetts, Amherst

We present a Bayesian Variable Selection algorithm appropriate for high dimensional datasets, where the outcome of interest is a time-to-event random variable that is observed at intermittent time points through error-prone procedures. The proposed methods are motivated by the Women's Health Initiative (WHI) Clinical Trial and Observational Study SHARe, which includes extensive genotypic (> 900K SNPs) and phenotypic data on 12,008 African American and Hispanic women. Due to cost considerations, incident diabetes is ascertained through self-reported questionnaires for all women enrolled in the WHI. Our algorithm incorporates a likelihood-based approach to account for the measurement error due to self-reports of diabetes. This is combined with a stochastic search algorithm (George, E. I. and McCulloch, R.E. (1993)) to identify relevant biomarkers by introducing a latent binary indicator used

to induce a mixture prior on the regression coefficients and explore the space of variable subsets. We extend our approach to incorporate external biological information, such as protein-protein interactions and biological pathways, using a Markov random field prior to improve the ability to detect true biomarkers. We illustrate our proposed algorithm through simulations and by application to GWAS data from the WHI Clinical Trial and Observational Study SHARE.

email: xdgu@schoolph.umass.edu

### **Cortical Thickness Thinning and Cognitive Impairment in Parkinson's Disease without Dementia**

**Lijun Zhang\***, The Pennsylvania State University  
Milton S. Hershey Medical Center

**Ming Wang**, The Pennsylvania State University  
Milton S. Hershey Medical Center

**Nicholas Sterling**, The Pennsylvania State University  
Milton S. Hershey Medical Center

**EunYoung Lee**, The Pennsylvania State University  
Milton S. Hershey Medical Center

**Guangwei Du**, The Pennsylvania State University  
Milton S. Hershey Medical Center

**Mechelle Lewis**, The Pennsylvania State University  
Milton S. Hershey Medical Center

**Xuemei Huang**, The Pennsylvania State University  
Milton S. Hershey Medical Center

Background and Purpose: Parkinsons disease (PD) is a progressive neurodegenerative disorder associated with brain structure changes and cellular remodeling. Cortical thickness may serve as a potential marker of PD progression, particularly because of its correlation with cognitive decline in related neurodegenerative disorders. However, the relationship of demographic and cognitive risk factors with cortical thinning remains unclear. This study investigated cortical thickness in non-demented PD subjects compared to healthy controls and may provide insights to disease pathogenesis. Methods: High-resolution T1-weighted brain MRI and comprehensive cognitive function tests were acquired for 71 non-demented PD subjects and 48 control subjects matched for age, education and gender. Cortical thickness assessment between groups was conducted using a hierarchical Bayesian model. In addition, demographic and cognitive risk factors were compared between PD and controls. Correlation analyses were performed among those brain areas and cognitive domains that showed significant group differences. Results: PD patients demonstrated significant thickness reduction localized predominantly in precunes, precentral, postcentral, insula, and superiorparietal regions. Patients also showed reduced cognitive performance in fine motor speed and executive function compared to controls.

email: lzhang6@hmc.psu.edu

### **Bayesian Modeling of Mixed Outcome Types Using Random Effect**

**Hua Wei\***, Eli Lilly and Company

The problem of analyzing associated outcomes of mixed type arises frequently in clinical trials and other practices. For example, in clinical studies, jointly analyzing associated binary safety outcome and continuous efficacy outcome is a common problem. We develop several Bayesian models for analyzing associated discrete and continuous responses simultaneously using random effects. We also extend these models to overcome the bias in parameter estimation due to ignorance of skewness of the continuous response, a misclassified covariate, and a zero-inflated discrete response. Simulation studies indicate that our models provide good estimates of regression coefficients, response variability, and the correlation between responses. We also show that ignoring the random effects leads to a bias in parameter estimates which is magnified with increasing variability of the random effect. Comparison to corresponding likelihood methods suggests that the Bayesian Poisson-normal (PN) model takes clear advantage of prior information when available, and performs similarly when relatively non-informative priors are used. Finally, we compare the PN model to a model using two separate but correlated random effects. In simulation studies we find there is little advantage to the more complicated model.

email: wei\_hua@lilly.com

### **An Objective Stepwise Bayes Approach to Small area Estimation**

**Yanping Qu\***, U.S. Food and Drug Administration  
**Glen D. Meeden**, University of Minnesota

The term "small area" is commonly used to denote a small geographical area that has a small subpopulation of people within a large area. Small area estimation is an important area in survey sampling because of the growing demand for better statistical inference for small areas in public or private surveys. Some traditional methods for small area problems borrow strength through linear models that provide links to related areas, which may not be appropriate for some survey data. We propose a stepwise Bayes approach which borrows strength through an objective posterior distribution. This approach results in a generalized constrained Dirichlet posterior estimator when auxiliary information is available for small areas. The objective posterior distribution is based only on the assumption of exchangeability across related areas and does not make any explicit model assumptions. The form of our posterior distribution allows us to assign a weight to each member of the sample. These weights can then be used in a straight forward fashion to make inferences about the small area means. Numerically, we demonstrate in simulations that the proposed stepwise Bayes approach can have substantial strengths compared to traditional methods.

email: yanping.qu@fda.hhs.gov

## 64. STATISTICAL LEARNING FOR COMPLEX MULTIVARIATE BIOMEDICAL DATA

### Linear Conditioning for Clustering Functional Data

**Thaddeus Tarpey\***, Wright State University

Cluster analysis is a popular unsupervised learning method for functional data. The ability of standard clustering methods, such as the k-means algorithm, to find interesting sources of heterogeneity in data can be compromised by uninteresting sources of variability. This talk explores various linear transformations of functional data with the goal of improving clustering results. For example, in order to better differentiate specific drug effects from placebo effects in clinical trials, a canonical-discriminant type transformation can be used to steer the clustering algorithms in directions highlighting differences in treatment groups. Also, the use of linear projections will be explored for clustering functional data.

email: thaddeus.tarpey@wright.edu

### Multiple Kernel Statistical Learning to Combine Heterogeneous Data Sources for Prediction

**Tianle Chen**, Columbia University

**Donglin Zeng**, University of North Carolina, Chapel Hill

**Yuanjia Wang\***, Columbia University

Modern high-throughput technologies offer opportunities to collect biomarkers from heterogeneous sources such as genetic data, imaging data and clinical data to study common mental and neurological disorders. Many epidemiological studies on natural history and etiology of mental disorders often span many years and clinical data are collected at multiple visits. It is thus highly valuable to develop statistical methods to efficiently combine heterogeneous sources of clinical, genetic, and imaging data to predict disease risk for diagnosis and screening purposes while accommodating the high dimensionality of genetic and imaging biomarkers. In this work, we propose a multiple kernel statistical learning (MKSL) approach to construct effective time-adaptive decision rules for disease risk prediction. We use different kernels for heterogeneous data sources taking advantage of each modality and then optimally combine data across modality. Furthermore, we account for within-subject correlation of repeated measures by introducing subject-specific random effects modeled through a separate kernel. The use of MKSL is especially appealing in biomedical applications where it is believed there is no clear winner and each data modality contributes partial information to prediction. We will apply developed methods to real data examples.

email: yw2016@columbia.edu

### Margin-Based Learning of Minimum Clinically Important Difference

**Tu Xu**, University of Illinois, Chicago

**Samad Hedayat**, University of Illinois, Chicago

**Junhui Wang\***, City University of Hong Kong

In clinical trials, minimum clinically important difference (MCID) has attracted increasing interest as an important supportive clinical and statistical inference tool. Many estimation methods have been developed based on various intuitions, while little theoretical justification has been established. In this talk, we will propose a new estimation framework of MCID using both diagnostic measurements and patient-reported outcomes (PROs). It first provides a precise definition of population-based MCID so that estimating such a MCID can be formulated as a large margin classification problem. The framework is then extended to personalized MCID to allow individualized thresholding value for patients whose clinical profiles may affect their PRO responses. More importantly, we will show that the proposed estimation framework is asymptotically consistent, and a finite-sample upper bound can be established for its prediction accuracy compared against the ideal MCID. The advantage of our proposed method is also demonstrated in a variety of simulated experiments as well as applications to two benchmark datasets and two phase-3 clinical trials.

email: junhwang@cityu.edu.hk

### Dynamic Directional Model for Effective Brain Connectivity Using Electrocorticographic (ECOG) Time Series

**Tingting Zhang\***, University of Virginia

**Jingwei Wu**, University of Virginia

**Fan Li**, Duke University

**Dana Boatman-Reich**, Johns Hopkins University

**Brian Caffo**, Johns Hopkins University

We introduce a dynamic directional model (DDM) for studying brain effective connectivity based on ElectroCorticographic (ECoG) time series. The DDM consists of two parts: a set of differential equations describing neuronal activity of brain components (state equations), and observation equations linking the underlying neuronal states to observed data. Most existing effective connectivity studies are based on the fMRI or EEG data; the associated DDM usually has complex formulation and thus can only accommodate a small number of brain regions. The highly-localized property and high temporal resolution of the ECoG data result in a much simpler DDM, allowing us to investigate connections within a large brain system with many regions. To identify functionally-segregated sub-networks, a form of biologically economic brain networks, we propose a modified Potts model within the general DDM. We represent the neuronal states of brain components by cubic spline bases and estimate the parameters by minimizing a log-likelihood criterion that combines the state and observation equations. The Potts model is converted to the Potts penalty in the penalized regression approach to achieve sparsity in parameter estimation, for which a fast iterative algorithm is developed. We apply the proposed methods to analyze a ECoG data set measured from patients with intractable epilepsy.

email: tz3b@virginia.edu

## 65. STATISTICAL CHALLENGES IN STUDIES OF ENVIRONMENTAL, REPRODUCTIVE AND PERINATAL HEALTH

### The Current Duration Approach to Estimating Time to Pregnancy

Niels Keiding\*, University of Copenhagen, Denmark

Time to pregnancy (TTP) is the time from a couple decides they want to become pregnant until they succeed. It is considered one of the most direct methods to measure natural fecundity in humans. Statistical tools to measure TTP belong to survival analysis, but several features require special attention. Prospective studies are difficult to carry out, and the most common study design, asking women who have become pregnant how long it took, gives a biased result by excluding sterile couples and couples who gave up trying to become pregnant. In the current duration design a cross-sectional sample of women (or men) in the fertile age are asked whether they are currently trying to become pregnant, and if so, for how long have they tried. The statistical and epidemiological properties of this design will be surveyed and illustrated from experiences in a large French survey (The Observatory of Fecundity in France) and on data from women as well as men in the U.S. National Survey of Family Growth. The current duration design includes sterile and subfertile couples, and follow-up interviews allow combining the retrospective current duration information with a prospective prevalent cohort study, as carried out in the French study.

email: nike@sund.ku.dk

### Prediction of Fecundity Based on Joint Modeling of Multiple Time Scale Longitudinal Intercourse and Menstrual Cycle Characteristics

Rajeshwari Sundaram\*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

We consider the joint analysis and prediction of longitudinal binary process (intercourse) and discrete time-to-event data (Time to Pregnancy), in the presence of multi-scale informative dropout. The data we consider is from a prospective pregnancy study, which collects day-level intercourse data on couples attempting to achieve pregnancy. An aspect of particular interest in such studies is to develop a model for individualized predictions of time to pregnancy (TTP). Obviously, a couple's intercourse behavior plays an integral part in such models. In our motivating data, the intercourse observations comprise of a long series of binary data with a periodic probability profile that is subjected to informative dropout. The number of observations per couple is a function of both the number of observed cycles (i.e., the TTP) and the number of observations per cycle (i.e., the menstrual cycle length) leading to a multiple timescale missingness mechanism of longitudinal process. We propose a semi-parametric shared parameter model for modeling intercourse behavior in

presence of such multiscale informative dropout. Further, we develop dynamic couple-based predictions of intercourse profiles, as well as the risk for subfertility (based on TTP).

email: sundaramr2@mail.nih.gov

### Air Pollution Metric Analysis while Determining Susceptible Periods of Pregnancy for Low Birth Weight and Birth Defects

Montse Fuentes\*, North Carolina State University

Joshua L. Warren, University of North Carolina, Chapel Hill

Amy H. Herring, University of North Carolina, Chapel Hill

Peter H. Langlois, Texas Department of State Health Services, Austin

Multiple ambient air pollution estimate metrics are available for use in environmental health analyses in addition to the standard Air Quality System (AQS) pollution monitoring data. These metrics are crucial in calculating pollution exposures in geographic areas where AQS monitors are not present due to their complete spatial-temporal coverage across a domain. We investigate the impact that two of these metrics, output from a deterministic chemistry model (CMAQ) and from a proposed spatial-temporal statistical downscaler which combines information from AQS and CMAQ, have on risk assessment. Using each metric, we analyze ambient particulate matter and ozone's effect on low birth weight and birth defects utilizing a Bayesian spatial temporal hierarchical multivariate probit regression model. Weekly windows of susceptibility are identified and analyzed jointly for all births in Texas, 2001-2004 and results from the different pollution metrics are compared. Increased exposures during weeks 20-23 of the pregnancy are identified as being significantly harmful in terms of low birth weight development by the statistical downscaler. Use of the CMAQ output results in increased variability of the final risk assessment estimates while the statistical metric results more closely resemble those of the AQS.

email: fuentes@ncsu.edu

### Identifying "Bad Actors" in Mixtures of Prenatal Exposures Associated with Reproductive Health Using Weighted Quantile Sum Regression

Chris Gennings\*, Virginia Commonwealth University

Caroline K. Carrico, Health Diagnostics Laboratory

Polychlorinated biphenyls (PCBs) remain ubiquitous environmental contaminants. Developmental exposures are suspected to impact reproduction. Analysis of mixtures of PCBs may be problematic as components have a complex correlation structure, and along with limited sample sizes, standard regression strategies are problematic. We compared the results of a novel, empirical method (i.e., weighted quantile sum (WQS) regression) to those based on categorization of PCB compounds by (1) hypothesized biological activity previously proposed and widely applied, and (2) degree of ortho-substitution (i.e., chlorination) in a study of the relation of maternal serum PCBs and daughter's time to pregnancy. WQS analyses found some associations previously identified by two classification schemes, but

also identified other bad actors. Simulation studies of environmental health applications demonstrate advantages in accuracy of variable selection of WQS regression compared to other model reduction strategies (e.g., lasso, adaptive lasso, and elastic net). This empirical method can generate hypotheses about mixture effects and mechanisms and overcomes some of the limitations of standard regression techniques.

email: gennings@vcu.edu

## 66. NEW DEVELOPMENTS IN STATISTICAL METHODOLOGIES FOR THE ANALYSIS OF DISEASE DATA

### **Making The Cut: Improved Ranking and Selection in Large-Scale Inference**

**Nicholas Henderson**, University of Wisconsin, Madison

**Michael A. Newton\***, University of Wisconsin, Madison

Identifying leading measurement units from a large collection is a common inference task in various domains of large-scale inference. Testing approaches, which measure evidence against a null hypothesis rather than effect magnitude, tend to overpopulate lists of leading units with those associated with low measurement error. By contrast, local maximum likelihood approaches tend to favor units with high measurement error. Available Bayesian and empirical Bayesian approaches rely on restricted loss functions that result in similar deficiencies. We describe and evaluate a generic empirical Bayesian ranking procedure that populates the list of top units in a way that maximizes the expected overlap between the true and reported top lists for all list sizes. Examples from genome-wide association testing and gene-set enrichment are presented.

email: newton@stat.wisc.edu

### **Generation of Virtual Control Groups for Single Arm Prostate Cancer Adjuvant Trials**

**Zhenyu Jia\***, University of Akron and

Northeast Ohio Medical University

**Michael B. Lilly**, Medical University of South Carolina

**Dan A. Mercola**, University of California, Irvine

It is difficult to construct a control group for trials of adjuvant therapy (Rx) of prostate cancer after radical prostatectomy (RP) due to ethical issues and patient acceptance. We utilized 8 curve-fitting models to estimate the time to 60%, 65%, . . . 95% chance of progression free survival (PFS) based on the data derived from Kattan post-RP nomogram. The 8 models were systematically applied to a training set of 153 post-RP cases without adjuvant Rx to develop 8 subsets of cases (reference case sets) whose observed PFS times were most accurately predicted by each model. To prepare a virtual control group for a single-arm adjuvant Rx trial, we first select the optimal model for the trial cases based on the minimum weighted Euclidean distance between the trial case set and the reference case set in terms of clinical features, and then compare the virtual PFS times calculated by the optimum model with the observed PFSs of the trial cases by the logrank test. The method was validated using

an independent dataset of 155 post-RP patients without adjuvant Rx. We then applied the method to patients on a Phase II trial of adjuvant chemo-hormonal Rx post RP, which indicated that the adjuvant Rx is highly effective in prolonging PFS after RP in patients at high risk for prostate cancer recurrence.

email: zjia@uakron.edu

### **Differential Network Analysis Using Microarray Gene Expression Data**

**Susmita Datta\***, University of Louisville

Complex diseases are caused by the change in cellular responses resulting from exposure to different environmental conditions. It has long been known that the genes do not act alone but rather work in consort during the processes. Naturally, it is important to study their responses in terms of a network and not marginally. We construct a possible way to reverse engineer interaction or association networks under varied experimental conditions. The networks are formed by the strength of association or interaction between pairs of genes, proteins or lipids. We use scores based on partial least squares regression to construct the association networks. Next, we propose formal statistical tests to test i) whether the overall modular structure of the transcription networks are different between two different experimental conditions ii) whether the connectivity has changed for a set of interesting genes between the two conditions and iii) whether the connectivity of a single gene has changed between the two experimental conditions. We show the effectiveness of our analysis in an experimental disease data involving mouse obesity study.

email: susmita.datta@louisville.edu

### **Fused LASSO with the Adaptation of Parameter Ordering (FLAPO) in Meta Analysis of Repeated Measurements**

**Fei Wang**, Ford Motor Credit

**Lu Wang**, University of Michigan

**Peter XK Song\***, University of Michigan

We focus on the marginal model for the regression analysis of repeated measurements measured in several similar studies. When the datasets are sampled from heterogeneous study populations, it is of great importance to examine whether certain common parameters exist across study-specific marginal models so that simpler models, sensible interpretations and meaningful efficiency gain can be obtained. The key challenge in such a meta analysis is that the classical hypothesis testing approach is computationally too intricate to evaluate all possible subsets of common regression parameters due to the forbidding computing burden related to a large number of hypotheses. We develop a new fused lasso method using estimated parameter ordering, which has been often neglected, to only scrutinize adjacent-pair parameter differences so as to reduce the number of needed comparisons dramatically. We show the proposed regularization method enjoys the selection

consistency and the oracle properties as the full fused lasso that concerns all pairwise parameter differences. We also show that the proposed procedure has smaller error bounds and better finite sample performances than the full fused lasso. We illustrate our method through simulation study and real world data analysis.

email: pxsong@umich.edu

## 67. RECENT DEVELOPMENT AND APPLICATION OF BAYESIAN METHODS FOR THE PROBABILITY OF SUCCESS AND DECISION MAKING IN CLINICAL TRIALS

### Using Prior Information to Help Determine Appropriate Metrics for Sound Decision Making in Drug Development

**Christy Chuang-Stein\***, Pfizer Inc.

There are many decision points along the product development continuum. Sound decisions begin with asking the right questions and choosing the appropriate metrics to help set up the decision criteria. In this talk, we will look at metrics that address the unique needs at various stage gates such as proof of concept and late stage development. We will illustrate how prior information should be used to construct these metrics and the decision criteria. The criteria have important implications in designing the trials that form the evidential basis for product approval. We will also illustrate why early observed treatment effect should be discounted when used to project future trial results. The interplay between decision criteria and design features will jointly determine the operating characteristic of the design and the quality of our decisions. We will use examples to illustrate these points.

email: christy.j.chuang-stein@pfizer.com

### Average Power and Average Conditional Power in Clinical Trial Design and Interim Analysis

**Kuang-Kuo G. Lan\***, Janssen R&D, Johnson & Johnson

We consider the use of Bayesian random treatment effects to evaluate the power of a Phase III trial, and the conditional power during interim data monitoring. In general, treatment effect may depend on the choice of primary endpoint and the model chosen for adjustment of covariates. Even when a well-defined primary endpoint is chosen, it is difficult to reach a consensus agreement on the prior distribution of treatment effect among research team members. Nonetheless, we urge the use of average power and related methods for calculating sample size, recognizing that the approaches are likely to be unpopular unless many statisticians and textbooks adopt them.

email: glan@its.jnj.com

### Bayesian Probability of Success for Superiority Trials in the Presence of Historical Data

**Joseph G. Ibrahim\***, University of North Carolina, Chapel Hill

**Ming-Hui Chen**, University of Connecticut

**Mani Y. Lakshminarayanan**, Merck, Inc.

**Guanghan Liu**, Merck, Inc.

**Joseph F. Heyse**, Merck, Inc.

Developing sophisticated statistical methods for GO/NO-GO decisions is crucial for clinical trials, as planning Phase III or Phase IV studies is costly and time consuming. In this talk, we develop a general Bayesian methodology for determining the probability of success of a treatment regimen based on the current data of a given trial. We introduce a new criterion for calculating the probability of success that allows for inclusion of covariates as well as allowing for historical data based on the treatment regimen, and patient characteristics. A new class of prior distributions and covariate distributions are developed to achieve this goal. The methodology is quite general and can be used with univariate or multivariate continuous or discrete data, and it contains the work of Chuang-Stein (2006) as a special case. This methodology will be invaluable for informing the scientist on the likelihood of success of the compound, while including the information on the biomarker, for planning future pre or post-market studies.

email: ibrahim@bios.unc.edu

### Evaluating Regression-to-the-Mean of Treatment Effect from Phase 2 to Phase 3

**Jianliang Zhang\***, MedImmune, LLC

It has been commonly observed that the treatment effect in phase 3 trials is generally smaller than what was observed in the preceding phase 2 trials. Potential causes for this shrinkage in treatment effect have been discussed in the literature. Some of the causes are program-specific while others are commonly seen across development programs. The impact is generally not quantifiable and some conceptual discounting approaches in phase 3 planning have been proposed to account for the shrinkage. In this presentation, I will discuss the impact of the phase 2 portfolio on the shrinkage. It will be shown that the impact is systemic and quantifiable given a prior distribution of the phase 2 portfolio and the size of the phase 2 program. The expected shrinkage in phase 3 is non-negligible and explains the majority of the shrinkage observed in the past given commonly used phase 2 sample sizes.

email: zhangj@medimmune.com

## 68. FUNCTIONAL DATA ANALYSIS: SHOW ME THE DATA

### CSI Statistics: Functional Data Analysis for Dead Bodies

**John Aston\***, University of Cambridge and University of Warwick

**Anjali Mazumder**, University of Warwick

**Anna Zylbersztejn**, University of Warwick and University of Leicester

It is not unusual in cases where a dead body is discovered that it is necessary to determine a time of death or more formally a post mortem interval (PMI). Forensic entomology can be used to estimate this PMI by examining evidence obtained from the body from insect larvae growth. Growth curves however are temperature dependent, and usually direct temperature measurements from the body location are unavailable for the time periods of interest. In this work, we investigate models initially for temperature prediction and then finally for PMI estimation based on functional data analysis.

email: j.a.d.aston@warwick.ac.uk

### Surviving in the ICU: The Case for Uneven Support Functional Data Analysis

**Ciprian Crainiceanu\***, Johns Hopkins University

**Jonathan Gellar**, Johns Hopkins University

Many biosignals are measured on uneven domains; this requires development of statistical methods that can effectively deal with uneven domains with minimal loss of information. In this study we will focus on subjects who were admitted in the Intensive Care Unit with respiratory distress syndrome (SDS). The mortality in this group of patients was roughly 50% in the ICU and each patient stayed in the ICU for a different number of days. We are interested in studying the association between the SOFA score (a daily measure of the severity of disability) and the probability of death both within ICU and after ICU discharge. We develop functional data approaches to deal with these types of uneven-support data and discuss several other applications including sleep EEG data and hand target reaching movements after stroke.

email: ccrainic@jhsph.edu

### Studying the Relationship Between Cerebral Vessel Morphology and Hemodynamic Forces in Arteries Affected by Aneurysms: A Spatial Functional Data Analysis Approach

**Laura M. Sangalli\***, Laboratory for Modeling and Scientific Computing MOX, Italy

**Bree Ettinger**, Emory University

**Simona Perotto**, Laboratory for Modeling and Scientific Computing MOX, Italy

We analyze hemodynamic forces, such as shear stress and pressure, exerted by blood-flow over the wall of a cerebral artery affected by an aneurysm. This study aims at enhancing the still limited knowledge on cerebral aneurysms pathology. The data are obtained within the AneuRisk project (<http://mox.polimi.it/it/progetti/aneurisk/>), via computational

fluid dynamics in real vessel geometries reconstructed from three-dimensional angiographies. To analyze such data, we follow a functional data analysis approach and develop spatial regression models for data occurring over bi-dimensional Riemannian manifolds. These are generalized additive models, able to account for the complex geometry of the non-planar domain, the vessel morphology, and to incorporate space-varying covariates. The models make use of advanced numerical analysis techniques. The estimators have the typical penalized regression form and are linear in the observed data values; classical inferential tools are available.

email: laura.sangalli@polimi.it

### Functional Prediction of Traffic Streams

**Jeng-Min Chiou\***, Academia Sinica

Traffic streams are often characterized by the three basic traffic measurements, flow rate, vehicle speed and occupancy (or density), which are constantly monitored by vehicle loop detectors. Motivated by the need for accurate traffic prediction, we propose functional data methods to analyze traffic patterns and predict future traffic streams. We approach the problem by sampling the triplet traffic trajectories from multivariate random functions. We propose a normalized multivariate functional principal component approach that takes advantage of component dependency through pairwise cross-covariance functions. This method serves as a basic tool in dimension reduction and multivariate functional data analysis. Based on this approach, we derive a multivariate functional regression model for analysis and prediction of multivariate functional responses. These methods are illustrated using an application with interest relating to traffic stream prediction.

email: jmchiou@stat.sinica.edu.tw

## 69. LATENT CLASS MODELS FOR DIAGNOSTIC TESTING WITH APPLICATIONS IN PSYCHIATRY

### Theory and Applications of the Self-Learning Q-Matrix

**Jingchen Liu\***, Columbia University

The Q-matrix, an incidence matrix specifying the item-attribute relationship, is a key element in the specification for many cognitive diagnostic models. It is common practice for the Q-matrices to be specified by experts when items are written, rather than through data-driven calibration. Such a non-empirical approach may lead to misspecification of Q-matrices and substantial lack of model fit, resulting in erroneous interpretation of results. In this talk, we present our recent findings concerning the data-driven construction (estimation) of Q-matrices. Upon writing the model in a regression form, we formulate the Q-matrix estimation to a model selection problem, for which regularized regression estimators are employed. The computation of such estimators is through a combination of the EM algorithm and existing convex optimization methods.

email: jcliu@stat.columbia.edu

### **Making Computerized Adaptive Testing a Diagnostic Tool**

**Hua-Hua Chang\***, University of Illinois, Urbana-Champaign

**Ya-Hui Su**, National Chung Cheng University

Although CAT was originally developed by for high stakes testing, its findings have been beneficial to other domains such as quality of life measurement, patient report outcome, K-12 accountability assessment, survey research, media and information literacy measure, etc. The paper provides a survey of 10 years' progress about Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT). We start with a historical review of the establishment of a large sample foundation for CAT under a framework of martingale theory. Then, we address a number of issues that emerged from large scale implementation in educational testing and show that CAT can be used in the research of patient reported outcome. In addition we show the newest development in the research of multidimensional cognitive diagnostic adaptive assessment. Many issues concerning CAT in quality of life measurement will be discussed.

email: hhchang@illinois.edu

### **Heterogeneous Variance Classification Models for Psychiatric Assessment Survey Data**

**Jonathan Templin\***, University of Kansas

**Lesia Hoffman**, University of Nebraska, Lincoln

**Ryan Walters**, University of Nebraska, Lincoln

**Meghan Sullivan**, University of Nebraska, Lincoln

Unique characteristics of populations of individuals with and without psychiatric disorders do not align with customary psychometric analyses for the development of survey screening instruments. In particular, the often complex high-dimensional features of the disorder will interact with characteristics of the population in that those without the disorder will exhibit reduced variability of responses when compared to those with the disorder. Additionally, the nature of the disorder may be such that those with the disorder display variability of symptom severity whereas those without the disorder will be mostly homogeneous with the absence of such symptoms. In this talk, we present two latent-class based approaches for examining plausibly continuous psychiatric screening survey data. The first follows from the field of diagnostic modeling where symptoms of a disorder are treated as being either present or absent. The second uses a mixture of discrete and continuous variables to evaluate the presence of a disorder, and, if so, the severity of the symptoms. Both approaches use constrained mixtures of multivariate Gaussian distributions where both the mean vector and covariance matrix are modeled as a function of the type of disorder (present/absent or if/how much). Example analyses with such survey data will demonstrate each approach.

email: jontemplin@gmail.com

### **Use of Latent Product Lattice Classification Models for Self-Reported Measures of Depression**

**Curtis Tatsuoka\***, Case Western Reserve University

Initial efforts will be described for validating existing self-report measures from PROMIS batteries within and across the three cognitive sub-groups: persons at risk for Alzheimer's disease (AD), those with mild cognitive impairment (MCI), and those with early AD. This includes depression measures. Item Response Theory (IRT) models are suggested by PROMIS investigators for other populations. We explore possible new methods beyond IRT for psychometrically modeling data such as from PROMIS. These involve the use of latent class models that are structured as products of lattices, which allows for generalization from unidimensionality assumptions. Some theoretical properties of these models in an adaptive testing framework will be presented.

email: curtis.tatsuoka@case.edu

## **70. STATISTICAL METHODS FOR BIOMARKER EVALUATION**

### **Semi-Parametric ROC Analysis Using Accelerated Regression Models**

**Eunhee Kim\***, Brown University

**Donglin Zeng**, University of North Carolina, Chapel Hill

The Receiver Operating Characteristic (ROC) curve is a widely used measure to assess the diagnostic accuracy of biomarkers for diseases. Biomarker tests can be affected by subject characteristics, the experience of testers, or the environment in which tests are carried out, so it is important to understand and determine the conditions for evaluating biomarkers. In this paper, we focus on assessing the effects of covariates on the performance of the ROC curves. In particular, we develop an accelerated ROC model by assuming that the effect of covariates relates to rescaling a baseline ROC curve. The proposed model generalizes the accelerated failure time model in the survival context to ROC analysis. An innovative method is developed to construct estimation and inference for model parameters. The obtained parameter estimators are shown to be asymptotically normal. We demonstrate the proposed method via a number of simulation studies, and apply it to analyze data from a prostate cancer study.

email: ekim@stat.brown.edu

### **Nonparametric ROC Based Evaluation for Survival Outcomes**

**Xiao Song\***, University of Georgia

**Xiao-Hua Zhou**, Puget Sound Health Care System and University of Washington

**Shuangge Ma**, Yale University

For censored survival outcomes, it can be of great interest to evaluate the predictive power of individual markers or their functions. Compared with alternative evaluation approaches, approaches based on the time-dependent receiver operating characteristics (ROC) rely on much weaker assumptions, can be more robust, and hence

are preferred. In this article, we examine evaluation of markers' predictive power using the time-dependent ROC curve and a concordance measure that can be viewed as a weighted area under the time-dependent area under the ROC curve profile. This study significantly advances from existing time-dependent ROC studies by developing nonparametric estimators of the summary indexes and, more importantly, rigorously establishing their asymptotic properties. It reinforces the statistical foundation of the time-dependent ROC-based evaluation approaches for censored survival outcomes. Numerical studies, including simulations and application to an HIV clinical trial, demonstrate the satisfactory finite-sample performance of the proposed approaches.

email: xsong@uga.edu

### **A Generalized C-Index for Survival Data**

**Patrick J. Heagerty\***, University of Washington

A concordance index (C-index) has been proposed as a measure of a model or marker's ability toward properly ordering survival outcomes. Heagerty and Zheng (2005) show that the survival C-index is linked to incident time-dependent accuracy concepts, while Saha-Chaudhuri and Heagerty (2013) outline non-parametric estimation. This presentation will review the survival C-index and discuss how a representation in terms of basic time-dependent accuracy concepts permits generalization to both time-varying covariates and competing risks. We overview non-parametric estimation and demonstrate the utility of the index for comparing models.

email: heagerty@uw.edu

### **Estimating Time-Dependent ROC Curve Using Data Under Outcome-Dependent Sampling**

**Shanshan Li\***, Indiana University Fairbanks

School of Public Health

**Mei-Cheng Wang**, Johns Hopkins Bloomberg

School of Public Health

In this presentation, we consider estimation of time-dependent receiver operating characteristic (ROC) curves when survival data are collected under outcome-dependent sampling. Outcome-dependent sampling can be found in many follow-up studies where data are collected according to a cross-sectional sampling scheme with certain eligibility criterion. Such sampling design tends to over sample individuals with longer survival times. To correct for the sampling bias, we develop both nonparametric and semiparametric estimators for the time-dependent ROC curves and the area under curve. The proposed estimators are consistent and converge to Gaussian processes, while substantial bias may arise if standard estimators for right-censored data are used. To illustrate our method, we analyze data from an Alzheimer's disease study and estimate ROC curves that assess how well the cognitive measurements can distinguish subjects that progress to mild cognitive impairment from subjects that remain normal.

email: sl50@iupui.edu

## **71. SEMI-PARAMETRIC AND NON-PARAMETRIC MODELS**

### **Robust Estimations of Scale, Dependence and Correlation Based on Quick Estimators**

**Lai Wei\***, U.S. Food and Drug Administration

**Alan Hutson**, State University of New York at Buffalo

An empirical quantile-based robust alternative measure of dependence is proposed and coined quick covariance. The statistical properties of quick covariance, analogous to those of the classic moment-based covariance, lead us naturally to develop a robust nonparametric alternative to the correlation coefficient, termed the quick correlation coefficient. In this note we examine the properties of the quick covariance, quick absolute deviation and quick correlation coefficient under bivariate normality and the asymptotic properties under the marginally symmetric assumption. An exact test that the population quick correlation coefficient is zero is developed. Extensive simulation studies are performed to compare the exact test of the quick correlation coefficient with that of the correlation coefficient for various distributions under different sample sizes and distributional assumptions. A data example is provided using the famous Anscombe's quartet data to illustrate how the proposed quick correlation coefficient estimator compares to the classic Pearson correlation coefficient estimator.

e-mail: laiwei@buffalo.edu

### **An RKHS Approach to Estimating High-Dimensional Graphs**

**Kuang-Yao Lee\***, Yale University

**Bing Li**, The Pennsylvania State University

**Hongyu Zhao**, Yale University

We present a new method for estimating graphs in high-dimensional setting. Our method is based on additive conditional independence - a newly proposed statistical relation by Li, Chun, and Zhao (2013). The concept of additive conditional independence aims at relaxing the two assumptions mostly considered in existing methods, a joint (copula) gaussianity among nodes or linear associations between nodes. In the meantime, unlike the fully specified conditional independence, additive conditional independence avoids the loss of efficiency by multivariate smoothing - which makes it especially suitable fitting large scale graphs. We show that at the population level the additive conditional independence can be characterized by identifying nonlinear patterns between variables. We also develop an estimating procedure and demonstrate it using simulations and actual data sets.

email: kuangyao.l@gmail.com

**Quantile Association Regression Models****Ruosha Li\***, University of Pittsburgh**Yu Cheng**, University of Pittsburgh**Jason Fine**, University of North Carolina, Chapel Hill

It is often important to study the association between two continuous variables. In this work, we propose a novel regression framework for assessing conditional associations on quantiles. General methodology is developed which permits covariate effects on both the marginal quantile models for the two variables and their quantile associations. The proposed quantile copula models have straightforward interpretation, facilitating a comprehensive view of association structure which is much richer than that based on standard product moment and rank correlations. The resulting estimators are shown to be uniformly consistent and weakly convergent as a process of the quantile index. Simple variance estimators are presented which perform well in numerical studies. Extensive simulations demonstrate the practical utility of the methodology. I will conclude the talk with an application to a twin dataset.

email: rul12@pitt.edu

**Calibrated Smoothed Bootstrap Confidence Intervals****Santu Ghosh\***, Wayne State University

In statistical analysis, confidence intervals are one of the most important inferential tools. The iterative and smoothed bootstrap methods have important applications in the construction of confidence intervals. It has been shown both methods can reduce the coverage error. However, it is well known that the iterated bootstrap method is computationally expensive. This prevents us to go beyond once iteration in the iterated bootstrap method. In practice, we can not use the iterative bootstrap to produce higher-order accurate confidence intervals. Here, we propose computationally efficient calibrated smoothed bootstrap percentile method confidence intervals which produce coverage error of order  $O(n^{-3/2})$ . This is an improvement upon the iterated bootstrap method. Secondly, we also provide an analytical adjustment to the nominal level to replace the need of double bootstrap for once iteration in the iterated bootstrap method. We further illustrate their finite sample performance through a simulation study.

email: santughosh001@gmail.com

**Semiparametric Group Testing Regression Models****Dewei Wang\***, Clemson University**Christopher S. McMahan**, Clemson University**Colin M. Gallagher**, Clemson University**Kurunarathna B. Kulasekera**, University of Louisville

Group testing, through the use of pooling, has proven to be an efficient method of reducing the time and cost associated with screening for a binary characteristic of interest, such as infection status. The salient feature of group testing that provides for these gains in efficiency

is that testing is performed on pooled specimens, rather than testing specimens one-by-one. Typically, the statistical literature surrounding group testing has investigated the implementation of pooled testing for the purposes of either case identification or estimation. A topic of key interest in the estimation problem involves the development of regression models that relate individual level covariates to testing responses observed from pooled specimens. The research in this area has primarily focused on parametric regression models. In this article, we merge the goals of classification and estimation by proposing a general semiparametric framework which allows for the inclusion of multi-dimensional covariates, decoding information, and imperfect testing. The asymptotic properties of our estimators are presented and guidance on finite sample implementation is provided. We illustrate the performance of our methods through simulation and by applying them to chlamydia and gonorrhea data collected by the Nebraska Public Health Laboratory as a part of the Infertility Prevention Project.

email: dwang@g.clemson.edu

**A Novel Pairwise Conditional Likelihood Ratio Test in a Semiparametric Model For vQTL Mapping****Chuan Hong\***, University of Texas School of Public Health, Houston**Yong Chen**, University of Texas School of Public Health, Houston**Yang Ning**, University of Waterloo**Peng Wei**, University of Texas School of Public Health, Houston

Current tests for single locus association with quantitative traits aim at looking for the mean difference and assume equal variances between genotypes or alleles. However, recent research has revealed functional genetic loci that affect the variance of traits, known as variability-controlling quantitative trait locus (vQTL). In addition, it has been suggested that many genotypes have both mean and variance effects, while some of the mean or variance effects alone would not be strong enough to be detected. A novel pairwise conditional likelihood ratio test is proposed to identify both mean and variance effects. By the conditioning technique, the baseline density function is eliminated in the constructed pairwise likelihood function. Hence the impact of unknown baseline density function is minimized. We show that the proposed test has a simple asymptotic chi-square distribution with four degrees of freedom. Simulation studies show that the proposed test performs well in controlling Type I errors and is powerful and robust to model misspecification. The proposed test is illustrated by an example of identifying both mean and variances effects in blood pressure among a group of Mexican Americans using the Genetic Analysis Workshop (GAW18) data.

email: chuan.hong@uth.tmc.edu

## **Fused Kernel-Spline Smoothing for Repeatedly Measured Outcomes in a Generalized Partially Linear Model with Functional Single Index**

**Fei Jiang\***, Rice University

**Yanyuan Ma**, Texas A&M University

**Yuanjia Wang**, Columbia University

We propose a generalized partially linear functional single index risk score model for repeatedly measured outcomes where the index itself is a function of time. We fuse the nonparametric kernel method and regression spline method, and modify the generalized estimating equation to facilitate estimation and inference. We use local smoothing kernel to estimate the unspecified coefficient functions of time, and use B-splines to estimate the unspecified function of the single-index component. The covariance structure is taken into account via a working model, which provides valid estimation and inference procedure whether or not it captures the true covariance. The estimation method is applicable to both continuous and discrete outcomes. We study the asymptotic properties when the kernel and regression spline methods are combined in a nested fashion. The work is motivated from a Huntington's disease (HD) research where a major goal is to make early prediction of HD diagnosis using cognitive tests preceding traditional clinical diagnosis. The application of the method on the Huntington's disease forms a composite score of four separate cognitive tests that is simple to interpret. The results show strong relationship between the cognitive symptoms and HD diagnosis, which supports the use of the cognitive symptoms to inform HD before the traditional diagnosis.

email: homebovine@gmail.com

## **72. JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA**

### **Joint Latent Class Models with Interval-Censored Survival Data**

**Lan Kong\***, The Pennsylvania State University College of Medicine

**Guodong Liu**, The Pennsylvania State University College of Medicine

In biomedical studies biomarkers and clinical endpoints are often evaluated at regularly scheduled follow-up visits to determine whether biomarker measurements are correlated with the development and progression of a disease. Generally the time to an event of interest is only known to be between two visits, resulting in interval-censored survival data. We consider a joint latent class model to investigate the association between longitudinal biomarker and time-to-event outcome subject to interval-censoring. Comparing to the popular joint models based on the shared random effects, joint latent class models assume a heterogeneous population made up of a finite set of homogeneous subpopulation that share the same trajectory pattern and the same risk of event. We develop an estimating procedure by using a parametric likelihood approach. Our method is evaluated via simulation studies and demonstrated with the neuroimaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

email: lkong@phs.psu.edu

## **A Fast EM Algorithm for Fitting Joint Models of a Binary Response and Multiple Longitudinal Covariates Subject to Detection Limits**

**Paul W. Bernhardt\***, Villanova University

**Daowen Zhang**, North Carolina State University

**Huixia J. Wang**, North Carolina State University

Joint modeling techniques are a popular strategy for studying the association between a response and one or more longitudinal covariates. Motivated by the GenIMS study, where the interest lies in modeling an indicator of survival using censored longitudinal biomarkers, a joint model is proposed for describing the relationship between a binary outcome and multiple longitudinal covariates subject to detection limits. A fast, approximate EM algorithm is proposed that only requires one-dimensional integration in the E-step of the algorithm, regardless of the number of random effects in the joint model. Simulations demonstrate that the proposed approximate EM algorithm leads to satisfactory parameter and variance estimates in situations with and without censoring on the longitudinal covariates.

email: paul.bernhardt@villanova.edu

### **Regression Modeling of Longitudinal Binary Outcomes with Outcome-Dependent Observation Times**

**Kay See Tan\***, University of Pennsylvania

Perelman School of Medicine

**Andrea B. Troxel**, University of Pennsylvania

Perelman School of Medicine

**Stephen E. Kimmel**, University of Pennsylvania

Perelman School of Medicine

**Kevin G. Volpp**, University of Pennsylvania

Perelman School of Medicine

**Benjamin French**, University of Pennsylvania

Perelman School of Medicine

Conventional longitudinal data analysis methods assume that outcomes are independent of the data-collection schedule. However, the independence assumption may be violated, for example, when adverse events trigger additional physician visits in between prescheduled follow-ups. Observation times may therefore be associated with outcome values, and potentially introduce bias when estimating the effect of covariates on outcomes using standard longitudinal regression methods. Existing semi-parametric methods that accommodate outcome-dependent observation times are limited to the analysis of continuous outcomes. We develop new methods for the analysis of binary outcomes, while retaining the flexibility of semi-parametric models. Our methods are based on counting process approaches, and provide marginal, population-level inference. In simulations, we evaluate the statistical properties of our proposed methods. Comparisons are made to "naive" GEE approaches that either do not account for outcome-dependent observation times or incorporate weights based on the observation-time process. We illustrate the utility of our proposed methods in an application to a clinical trial evaluating the effectiveness of

various interventions among patients treated with warfarin. We show that our method performs well in the presence of outcome-dependent observation times, and provide identical inference to “naïve” approaches when observation times are not associated with outcomes.

email: kaystan@mail.med.upenn.edu

### **Joint Model for a Diagnostic Test without a Gold Standard in the Presence of a Dependent Terminal Event**

**Sheng Luo\***, University of Texas, Houston

**Xiao Su**, University of Texas, Houston

**Stacia DeSantis**, University of Texas, Houston

**Xuelin Huang**, University of Texas

MD Anderson Cancer Center

**Min Yi**, University of Texas MD Anderson Cancer Center

**Kelly Hunt**, University of Texas

MD Anderson Cancer Center

Breast cancer patients after breast conservation therapy often develop ipsilateral breast tumor relapse (IBTR), whose classification (true local recurrence versus new ipsilateral primary tumor) is subject to error and there is no available gold standard. Some patients may die due to breast cancer before the IBTR develops. This terminal event is highly correlated with IBTR occurrence. This article presents a joint analysis framework to model the binomial regression with misclassified binary outcome and the correlated time to IBTR, subject to a dependent terminal event and in the absence of a gold standard. Shared random effects are used to link together two survival times. The proposed approach is evaluated by a simulation study and is applied to a breast cancer dataset consisting of 4,477 breast cancer patients. The proposed joint model can be conveniently fit using adaptive Gaussian quadrature tools implemented in SAS procedure NLMIXED.

email: sheng.t.luo@uth.tmc.edu

### **Modeling Short- and Long-Term Characteristics of Follicle Stimulating Hormone as Predictors of Severe Hot Flashes in Penn Ovarian Aging Study**

**Bei Jiang\***, University of Michigan

**Naisyin Wang**, University of Michigan

**Mary D. Sammel**, University of Pennsylvania

**Michael R. Elliott**, University of Michigan

The Penn Ovarian Aging Study tracked a population-based sample of 436 women aged 35-47 years to determine associations between reproductive hormone levels and menopausal symptoms. We develop a joint modeling method that uses the individual-level longitudinal measurements of follicle stimulating hormone (FSH) to predict the occurrence of severe hot flashes in a manner that distinguishes long-term trends of the mean trajectory, cumulative changes captured by the derivative of mean trajectory, and short-term residual variability. Our method allows the potential effects of longitudinal trajectories on the health risks to vary and accumulate over time. We further utilize the proposed methods to narrow down the critical time windows of increased health risks. We find that high

residual variation of FSH is a strong predictor of hot flash risk, and that the high cumulative changes of the FSH mean profiles in the 52.5-55 year age range also provides evidence of increased risk above and beyond that of short-term FSH residual variation by itself.

email: beijiang@umich.edu

### **Regression Analysis of Longitudinal Data with Correlated Censoring and Observation Times**

**Yang Li\***, University of North Carolina, Charlotte

**Haiying Wang**, University of New Hampshire

**Jianguo Sun**, University of Missouri, Columbia

Longitudinal data usually occur in medical follow-up studies when repeated measurements are taken on a series of sampling times. Most existing approaches on longitudinal data analysis assumed that the observation or censoring times are independent of the response process either completely or given some covariates. We present a joint analysis approach in which the possible mutual correlations are characterized by time-varying random effects. Estimating equations are developed for the parameter estimation and a simulation study is conducted to assess the finite sample performance of the approach. The asymptotic properties of the proposed estimates are also given and the method is applied to an illustrative example.

email: Y.Li@unc.edu

### **Mixtures of Gaussian Processes Applied to Medical Monitoring of Lung Function Decline and Pulmonary Exacerbations in Cystic Fibrosis**

**Leo L. Duan\***, Cincinnati Children's Hospital

Medical Center

**John P. Clancy**, Cincinnati Children's Hospital

Medical Center

**Rhonda D. Szczesniak**, Cincinnati Children's Hospital

Medical Center

Patients with cystic fibrosis (CF) lung disease are monitored frequently for the recurrence of an acute respiratory event known as a pulmonary exacerbation (PEX). These events are associated with lung function decline. Joint modeling of longitudinal data and failure time events can provide more accurate estimates than treating these data separately. In this paper, we describe a method to realistically model the linkage between the two variates. Patient-level forecasting or extrapolation is performed via more subtle description of within-subject correlation. Both the intervariate and intrasubject dependencies decay naturally as the time gap increases. In each submodel, the latent process is directly or indirectly defined as a mixture of Gaussian processes, which assimilates the information from the population and individual levels. This design allows us to obtain accurate prediction results by using three factors: joint effects, population effects and patient-level correlation effects. We present a discrete prediction method that achieves both dimension reduction and simplicity of likelihood, which enables us to obtain inferences and forecasts for lung function and PEX risk rapidly under a fully Bayesian framework. We assess the performance of the model through simulation studies and apply it to monitoring data from CF patients.

email: li.duan@cchmc.org

## 73. STATISTICAL METHODS IN EPIDEMIOLOGY

### Comparing Parametric and Semi-Parametric Regression Models for a Skewed, Pooled Outcome

**Emily M. Mitchell\***, National Institute of Child Health and Human Development, National Institutes of Health

**Robert H. Lyles**, Emory University

**Michelle Danaheer**, National Institute of Child Health and Human Development, National Institutes of Health

**Neil J. Perkins**, National Institute of Child Health and Human Development, National Institutes of Health

**Enrique F. Schisterman**, National Institute of Child Health and Human Development, National Institutes of Health

Pooling biological specimens prior to performing expensive laboratory tests can considerably reduce lab assay costs associated with certain epidemiologic studies. While analysis of these pooled measurements is often straightforward in linear regression, many public health studies involve skewed outcome data. In such cases, lognormal or gamma regression models make attractive options, but may require complex analytical techniques in order to calculate maximum likelihood estimates (MLEs) for certain types of pools. As one solution, we exploit an alternate parameterization of the typical gamma regression model that takes advantage of the unique summation properties of this distribution regardless of the pooling strategy. We also demonstrate a more flexible semi-parametric approach that shares key characteristics of the typical gamma regression model, but greatly simplifies the calculation of coefficient estimates. To help choose between models, we extend Akaike's Information Criterion (AIC) for analyzing assay data on pooled specimens. We use simulations to assess the proposed methods and determine the potential consequences of distributional misspecification.

email: emily.mitchell@nih.gov

### Effect Modification and Design Sensitivity in Observational Studies

**Jesse Y. Hsu\***, University of Pennsylvania

**Dylan S. Small**, University of Pennsylvania

**Paul R. Rosenbaum**, University of Pennsylvania

In an observational study of treatment effects, subjects are not randomly assigned to treatment or control, so differing outcomes in treated and control groups may reflect a bias from nonrandom assignment rather than a treatment effect. After adjusting for measured pretreatment covariates, perhaps by matching, a sensitivity analysis determines the magnitude of bias from an unmeasured covariate that would need to be present to alter the conclusions of the naive analysis that presumes adjustments eliminated all bias. Other things being equal, larger effects tend to be less sensitive to bias than smaller effects. Effect modification is an interaction between a treatment and a pretreatment covariate controlled by matching, so that the treatment effect is larger at some values of the covariate than at others. In the presence of effect modification, it is possible that results are less sensitive to bias in subgroups experiencing larger effects. We consider two scenarios (i) an a priori grouping into a few categories based on covariates controlled by

matching and (ii) a grouping discovered empirically in the data at hand, and evaluate the methods using simulation. Sensitivity analysis for a test of the global null hypothesis of no effect is converted to sensitivity analyses for subgroup analyses using closed testing. A study of an intervention to control malaria in Africa is used to illustrate.

email: hsu9@wharton.upenn.edu

### Structural Nested Mean Model for Clustered Outcomes

**Jiwei He\***, University of Pennsylvania

**Marshall Joffe**, University of Pennsylvania

Structural nested models (SNMs) are useful in dealing with confounding by variables affected by treatment. Two types of SNMs are structural nested mean models (SNMMs) and structural nested distribution models (SNDMs). We extend SNMMs to clustered observations. We consider how to formulate models with both cluster- and unit-level treatments, and show how to derive semiparametric estimators of parameters in those models. The properties of parameter estimators are evaluated through simulations. We use a dataset from ophthalmology clinical trial to illustrate our method.

email: jiweihe@mail.med.upenn.edu

### Flexible Models for Comparing Cumulative Effects of Time-Dependent Exposures

**Chenkun Wang\***, Indiana University School of Medicine and Richard M. Fairbank School of Public Health

**Hai Liu**, Indiana University School of Medicine and Richard M. Fairbank School of Public Health

**Sujuan Gao**, Indiana University School of Medicine and Richard M. Fairbank School of Public Health

In pharmacoepidemiologic studies, comparing the risk of adverse events among multiple medication exposures over an extended period of time is often necessary. Statistical methods used for these comparisons need to be able to capture the complex time-dependent exposures and model the varying effects of medications over time. Spline based models have been proposed to model the risk of adverse events from a single exposure. We propose a flexible modeling approach for comparing the cumulative effects of multiple time-dependent exposures using Cox's proportional hazard models. Spline functions are used to model weight functions that summarize past exposures and provide information on medication effects over time. We describe parameter estimation and hypothesis testing procedures that can be conducted using standard statistical software packages. We evaluate the proposed method in a simulation study and also apply the new method to a data set comparing the risk of coronary artery disease between patients taking two different types of antidepressants.

email: wang280@umail.iu.edu

### Instrumental Variables Estimation with Some Invalid Instruments and its Application to Mendelian Randomization

**Hyunseung Kang\***, University of Pennsylvania  
**Anru Zhang**, University of Pennsylvania  
**T. Tony Cai**, University of Pennsylvania  
**Dylan S. Small**, University of Pennsylvania

Instrumental variables have been widely used to estimate the causal effect between exposure and outcome. Unfortunately, using the usual techniques, proper estimation requires completely knowledge about all the instruments' validity; a valid instrument must not have a direct effect on the outcome and not be related to unmeasured confounders. Often, this is impractical as highlighted by Mendelian randomization studies where complete knowledge about instruments' validity is equivalent to complete knowledge about the instruments' gene functions, such as their potential pleiotropic effects on the outcome. In this paper, we present results concerning the estimation of causal effects when this complete knowledge is absent. In particular, we prove that identification and estimation is possible under a weaker requirement that more than 50% of instruments are invalid, without precisely knowing which of the 50%+ instruments are invalid. We show sharp limits on identification in the presence of invalid instruments. In addition, we propose a very fast penalized  $L_1$  estimation method that can estimate the causal effect without knowing which instruments are valid, with theoretical guarantees on its performance. We demonstrate the proposed method on simulated data and a real Mendelian randomization study concerning the effect of body mass index on health-related quality of life index.

email: khyuns@wharton.upenn.edu

### Variable Selection for Case-Cohort Studies with a Diverging Number of Parameters

**Ai Ni\***, University of North Carolina, Chapel Hill  
**Jianwen Cai**, University of North Carolina, Chapel Hill

Case-cohort design is widely used in large cohort studies to reduce the cost associated with covariate measurement. In many of those studies the number of covariates is very large, especially with the increasing availability of massive genetic information. Therefore, an efficient variable selection method is needed for case-cohort design with a diverging number of parameters. In this paper, we study the properties of the Smoothly Clipped Absolute Deviation (SCAD) penalty based variable selection procedure in case-cohort design. We establish the consistency and asymptotic normality of the maximum penalized pseudo-partial likelihood estimator. We also show that the proposed model selection procedure can identify the true model with probability one as sample size goes to infinity, and it estimates the nonzero parameters as efficiently as if the true model is known a priori. Extensive simulation studies are conducted to assess and compare the finite sample performance of the proposed variable selection procedure with AIC- and BIC-based tuning parameter selection methods. We make recommendations for practical use of the variable selection procedures in case-cohort studies. The proposed procedure is applied to the Busselton Health Study.

email: andyni@live.unc.edu

### Weighted Model Selection for Fractional Polynomial Models

**Michael D. Regier\***, West Virginia University  
**Ruoxin Zhang**, West Virginia University  
**John Honaker**, West Virginia University

Fractional polynomials (FP) represent a class of supervised learning prediction models for which model selection and covariate transformations happen in tandem. It has been proposed that FP models may be excellent candidates for understanding complex, non-linear dose-response relationships because the associative effect derived from coefficient estimates can be directly interpreted or graphically visualized to yield dose-response information. A shortcoming of FP models is their inherent instability with respect to the minor changes in the underlying data set. This poses serious challenges for model selection and generalizability. This aspect of the model may be responsible for the limited use of this class of models for dose-response research and more generally epidemiological research. We propose a weighted model selection methodology that accounts for model space cardinality and retains generalizability. We compare this with current methods of FP model selection and apply our comparative approach to both clinical and public health research questions.

email: mregier@hsc.wvu.edu

## 74. ADAPTIVE DESIGNS AND RANDOMIZATION

### Two-Stage Adaptive Optimal Design with Fixed First Stage Sample Size

**Adam Lane\***, Cincinnati Children's Hospital Medical Center  
**Nancy Flournoy**, University of Missouri

Sequential designs increase precision of parameter estimates and are often used when data is costly or difficult to collect, for example in phase I/II clinical, bioassay studies, etc. An adaptive optimal design is a sequential procedure which uses the data from all previous stages to estimate the locally optimal design of the current stage. A challenge present in adaptive optimal design is the derivation of an approximate (or exact) distribution that inference can be based upon. A common solution to this problem is to derive an asymptotic distribution under the assumption that the sample size for each stage go to infinity as the overall sample size goes to infinity. Such an assumption is not appropriate when a small pilot study of fixed size to be followed by a much larger experiment. We study the large sample behavior of such studies. For simplicity, we assume a nonlinear regression model with normal errors. We show that the distribution of the maximum likelihood estimates converges to a scale mixture family of normal random variables. We compare the behavior of these estimates with those obtained from the normal distribution that results when both stage samples are large.

email: adam.lane@cchmc.org

### **Phase II/III Seamless Adaptive Dose Selection Design for Longitudinal Patient Data**

**Caitlyn Ellerbe\***, Medical University of South Carolina  
**Jordan Elm**, Medical University of South Carolina  
**Viswanathan Ramakrishnan**, Medical University of South Carolina  
**Bruce Turnbull**, Cornell University  
**Edward Jauch**, Medical University of South Carolina  
**Stacia DeSantis**, University of Texas Health Sciences  
**Valerie Durkalski**, Medical University of South Carolina

Adaptive designs offer investigators the ability to modify trial parameters to promote safety and trial efficiency. However, in longitudinal phase II/III trials, adaptive designs are limited in the ability to use partial information without inflating the type I error. We propose a two-stage design for a continuous endpoint measured at several visits after enrollment. In stage I several doses of interest are compared to a control, and the optimal dose is selected using all available data. In stage II the efficacy of the selected dose relative to a placebo is tested using data from new subjects as well as the data used for the dose selection in stage I. We propose a correction to the test statistic, which we show, theoretically and through simulations controls the type I error rate, for selecting one of two doses observed over two time points. For more general situations, where there are multiple doses and more than two visits, a bootstrap procedure is proposed to control the type I error. This procedure provides a mechanism to generalize the design to more complex situations that include physician guided dose selection and dose-response models.

email: ellerbcn@muscc.edu

### **The Use of Decreasingly Informative Priors in Adaptive Clinical Trial Designs**

**Roy T. Sabo\***, Virginia Commonwealth University

The natural lead-in can reduce or eliminate several problems resulting from the use of outcome-adaptive allocation, and offers a reasonable and aesthetic alternative to fixed or conditional lead-in methods. However, since natural approaches typically require the adulteration of existing adaptation algorithms designed for some optimality criterion, they may not possess optimality characteristics, though most usually converge to the optimal form as the number of accrued patients approaches the targeted sample size. An alternative approach is presented using the concept of decreasingly informative priors, where prior information is based upon purposeful skepticism weighted by the number of patients not yet accrued into the trial. This approach is presented in both two- and three-sample cases with binary outcomes, and two forms of posterior estimation are discussed. Simulation studies are used to show that the decreasingly informative prior method provides modest adaptation and benefit (with respect to the number of treatment successes), while maintaining desired power and error rates.

email: rsabo@vcu.edu

### **An Adaptive Bayesian Dose Finding Approach for Drug Combinations with Drug-Drug Interaction**

**Yang Yang\***, University of Maryland, Baltimore County  
**Hong-Bin Fang**, Georgetown University  
**Anindya Roy**, University of Maryland, Baltimore County  
**Ming Tan**, Georgetown University

In a single-agent dose finding Phase I trial, the key underlying assumption is that toxicity probability increases monotonically with the dose level. However, in multi-agents trial, this assumption may fail because the drug-drug interaction effect can either decrease or increase the joint toxicity as compared to either one used alone, which may lead to an unforeseen toxicity probability surface. In this article, we develop a novel adaptive dose-finding approach which can be applied to this kind of multi- drug combination trials. With this approach, drug-drug interaction and toxicity probability are modeled jointly through a Bliss independence model. The main goal of our dose finding scheme is to search for Maximum tolerated region (MTR), as opposed to maximum tolerated dose (MTD) in single agent phase I trials. Dose escalation/de-escalation decision rules are determined by the posterior estimates of both drug-drug interaction effect and its corresponding joint toxicity probability, which can be continuously updated by sequentially assigning new patients into the trial while more data being observed. We evaluate the operating characteristics of the proposed method and also compare it with existing methods through extensive simulation studies under various scenarios. The proposed method demonstrates satisfactory performance in general.

email: yang10@umbc.edu

### **Dose Escalation with Over-Dose and Under-Dose Controls for Phase I/II Clinical Trial**

**Zheng Li\***, Emory University  
**Michael Kutner**, Emory University  
**Ying Yuan**, University of Texas  
MD Anderson Cancer Center  
**Zhengjia Chen**, Emory University

To save time and resources in new drug development, Phase I/II clinical trials with toxicity response and drug efficacy as dual primary endpoints have become more and more popular. Escalation with over dose control (EWOC) is a leading Bayesian adaptive Phase I clinical trial design which can estimate maximum tolerated dose (MTD) accurately and control overdosing. To adapt to the Phase I/II clinical trials, EWOC and Gumbel Copula model are used to incorporate additional under-dosing control to guarantee minimum efficacy of drug for patients. Late onset and missing efficacy are common in Phase I/II clinical trials, especially during early stage when patients are treated at low dosage. Therefore, we further employ the Bayesian data augmentation (DA) algorithm to compute values for late onset or missing efficacy data. The new Phase I/II design which can monitor the toxicity and efficacy simultaneously is named as Dose Escalation with Over Dose and Under Dose Control using Data Augmentation (EWOC-DA). The underlying theory

of EWOC-DA is elaborated and extensive simulations are conducted to evaluate its performance and operating characteristics. EWOC-DA has been demonstrated to provide better over-toxic control, optimize utility, and reduce the risk of failure in Phase III clinical trials compared to EWOC.

email: zheng.li@emory.edu

### **An Adaptive Treatment Strategy for the Management of White-Nose Syndrome**

**Nick Meyer\***, North Carolina State University  
**Eric Laber**, North Carolina State University  
**Krishna Pacifici**, North Carolina State University  
**Brian Reich**, North Carolina State University

Bats are a primary consumer of agricultural pests and insect vectors of human disease. Consequently, the emergence of White-Nose Syndrome, a rapidly spreading and fatal fungus afflicting bats, poses a serious threat to U.S. agriculture, ecosystem diversity, and human health. We construct a data-driven adaptive sequential treatment strategy for the management of White-Nose Syndrome which combines systems dynamics models and online updating algorithms. The proposed method uses historical data and ecological theory to estimate a systems dynamics model which is used to derive an initial treatment strategy. The initial treatment strategy is then updated as data accumulates over time using a stochastic approximation algorithm. We show the proposed method is consistent under regularity conditions and derive a null distribution for parameters indexing the estimated optimal treatment strategy. The method is illustrated using simulated experiments.

email: nick.j.meyer@gmail.com

## **75. NEXT GENERATION SEQUENCING**

### **Genotype Calling and Haplotyping in Extended Families**

**Lun-Ching Chang**, University of Pittsburgh  
**Bingshan Li**, Vanderbilt University  
**George C. Tseng**, University of Pittsburgh  
**Wei Chen\***, University of Pittsburgh

The emerging next generation sequencing technologies allow us to examine the variation of human genome in many individuals. However, the methods for analyzing family-based sequence data, particularly from nuclear and multi-generational families, are still lacking. In this talk, we describe a method for genotype calling in settings where sequence data are available for unrelated individuals and/or general families. The method takes both linkage disequilibrium (LD) patterns and the family structure information into consideration while retaining the computational efficiency. We loop all possible trios in each family and iteratively update parents using information from each offspring at each iteration of MCMC step. We summarize all sampled haplotypes for each sample after a pre-defined number of iterations using switch-error minimization. We apply our methods to both simulated and real data sets and show

that our methods can achieve high accuracy of genotype calling and phasing when multiple children are present and can reduce the Mendelian errors greatly in each family. In addition, we extend our method to incorporate external panels (e.g. 1000 Genomes Project) to analyze family-based sequence data with a small sample size. We anticipate that the proposed methods will be useful for many ongoing family-based and population-based sequencing projects.

email: weichen.mich@gmail.com

### **Meta-Analysis of Sequencing Studies Under Random-Effects Models**

**Zheng-Zheng Tang\***, University of North Carolina, Chapel Hill  
**Dan-Yu Lin**, University of North Carolina, Chapel Hill

Recent advances in sequencing technologies have made it possible to explore the influence of rare variants on complex human diseases. Meta-analysis is essential to this exploration because large sample sizes are required to detect rare variants. Several methods are available to conduct meta-analysis for rare variants under fixed-effects models, which assume that the genetic effects are the same across all studies. Such methods will lose power if there is between-study heterogeneity. We propose random-effects models which allow the genetic effects to vary among studies and develop the corresponding meta-analysis methods for gene-level association tests. Our methods take score statistics as input and thus can accommodate any study designs and any phenotypes. We produce the random-effects versions of all commonly used gene-level association tests, including burden, variable threshold and variance-component tests. We demonstrate through extensive simulation studies that our random-effects tests are substantially more powerful than the fixed-effects tests in the presence of moderate and high heterogeneity and achieve similar power to the latter when the heterogeneity is low. In an application to a deep-sequencing project on drug targets, our methods discovered a gene for total cholesterol which was undetected by the existing methods.

email: ztang@live.unc.edu

### **Likelihood Based Complex Trait Association Testing for Arbitrary Depth Sequencing Data**

**Song Yan\***, University of North Carolina, Chapel Hill  
**Yun Li**, University of North Carolina, Chapel Hill

In next generation sequencing (NGS) based genetic association analysis, accurate genotypes calling may not be easily feasible. Moreover, genotype calling may ignore some extent of genotype uncertainty in NGS data and leads to power loss. In the literature, likelihood based methods have been proposed to carry out association testing without genotype calling. Those methods take genotype uncertainty into account by incorporating genotype likelihood function (GLF) of NGS data into analysis directly. However, the existing LRT is computationally inefficient and does not adjust additional covariates. The score test proposed by Skotte et al. (2012) is not applicable when minor allele frequencies (MAFs) are lower than some threshold and also fails to consider the correlation between regression parameters and MAF and thus statistically underpowered. We provide

a LRT with additional covariates adjusted and develop a statistically more powerful score test. A combination strategy is thereby proposed to take advantage of the proposed LRT and score test. Simulations and real data analysis demonstrated that the proposed combination strategy is not only computationally more efficient than the proposed LRT and more applicable than the proposed score test but also statistically more powerful than at least one of the two methods.

email: songyan@unc.edu

### **Design Issue And Power Calculation In RNA-seq Applications**

**Chien-Wei Lin\***, University of Pittsburgh

**George C. Tseng**, University of Pittsburgh

Over the past decade, capabilities of genome-wide expression analysis have increased dramatically and the technology has gradually shifted from traditional microarray to RNA-seq techniques. Although RNA-seq provides more accurate and abundant information, the experimental cost is still inhibitive to allow large sample size in most projects. An important design issue in RNA-seq is the balance between sequencing depth and the number of samples performed under a fixed budget constraint. We will develop analytical and simulation methods to provide a practical guideline of RNA-seq design issues for detecting differentially expressed genes. In the former setting, distributional assumptions are made to allow an analytical answer. In the latter setting, pilot data are provided to allow parameter estimation and simulation for a practical power calculation.

email: masaki396@gmail.com

### **A Simulation-Based Comparative Study Of The Relative Power Of Family-Based Association Tests**

**Jia Jia\***, University of Pittsburgh

**Daniel E. Weeks**, University of Pittsburgh

The statistical genetics community has created a large number of different statistics for testing for association on family data, but it is not necessarily clear which one of these would be best for a particular data set. Accordingly, we compared a number of different association statistics on identical simulated data sets to determine which ones are best under which conditions. By simulating nuclear family data and varying the linkage disequilibrium, family size, completeness of genotyping, and the disease model, we evaluated and compared Type I error, power, robustness and computational speed of different association statistics. When comparing these statistics, it is important to be cognizant of exactly which null hypothesis (e.g., no association and no linkage; no association given complete linkage) is being tested. Our simulation study shows that, in order to test for association or linkage, one should apply the more powerful quasi-likelihood score statistic. In order to test for association given linkage, one should apply the association given linkage test under a recessive penetrance model using the Pseudomarker package. In order to test for association, one should apply Generalized Estimating Equations with independent working correlations using the GDT package.

email: jiajiafbi@aol.com

### **A DNA Variant Caller Adapted to Assess Mitochondrial DNA Variation from Whole-Genome Sequencing Data**

**Jun Ding\***, National Institute on Aging,  
National Institutes of Health

**Carlo Sidore**, Istituto di Ricerca Genetica e Biomedica,  
Consiglio Nazionale delle Ricerche, Monserrato,  
Cagliari, Italy

**Osorio Meirelles**, National Institute on Aging,  
National Institutes of Health

**Mary Kate Wing**, University of Michigan

**Fabio Busonero**, Istituto di Ricerca Genetica e Biomedica,  
Consiglio Nazionale delle Ricerche, Monserrato,  
Cagliari, Italy

**Ramaiah Nagaraja**, National Institute on Aging, National  
Institutes of Health

**Francesco Cucca**, Istituto di Ricerca Genetica e Biomedica,  
Consiglio Nazionale delle Ricerche, Monserrato,  
Cagliari, Italy

**Goncalo R. Abecasis**, University of Michigan

**David Schlessinger**, National Institute on Aging,  
National Institutes of Health

Genome-wide association studies have been successful in identifying variants associated with complex diseases and traits in the past eight years. More recently, the next-generation sequencing technology has become a more powerful tool. Within this framework, we have developed an algorithm specific for identifying variants in mitochondrial DNA (mtDNA) in order to analyze mtDNA variation and its possible effects on aging-related traits. Because each cell has 100-10,000 mtDNA copies that can vary at any site (heteroplasmy), the genotype calling programs for nuclear DNA are not adequate. Our algorithm is adapted to the special features of mtDNA; it incorporates in a likelihood calculation the sequencing error rates in the reads and allows for different allele fractions at a variant site across individuals. The program has been employed to assess homo- and hetero-plasmies in mtDNA sequences of lymphocytes from ~2,000 Sardinia participants. The results provide information about the inheritance of homo- and hetero-plasmies in Sardinia, and about the extent of accumulation of heteroplasmies during aging. The algorithm can be further extended in several ways: for example, to investigate the nuclear DNA variability in cancer cells.

email: jun.ding@nih.gov

### **Analysis of Sequence Data Under Multivariate Trait-Dependent Sampling**

**Ran Tao\***, University of North Carolina, Chapel Hill

**Donglin Zeng**, University of North Carolina, Chapel Hill

**Nora Franceschini**, University of North Carolina,  
Chapel Hill

**Kari E. North**, University of North Carolina, Chapel Hill

**Eric Boerwinkle**, University of Texas Health Science  
Center, Houston

**Dan-Yu Lin**, University of North Carolina, Chapel Hill

High-throughput DNA sequencing is a cutting-edge technology for genetic association studies. Currently, it is prohibitively expensive to sequence all subjects in a large cohort. A cost-effective strategy is to preferentially sequence the subjects with the extreme values of a quantitative trait.

We consider the situation in which the sampling depends on multiple quantitative traits. Under such outcome-dependent sampling, standard linear regression analysis is invalid and inefficient. We construct a semiparametric likelihood that properly reflects the sampling mechanism. In our formulation, quantitative traits are related to genetic variants and covariates through a multivariate linear regression model while the distributions of genetic variants and covariates are arbitrary. We implement a computationally efficient algorithm and establish the theoretical properties of the resulting estimators. We pay special attention to the gene-level association tests for rare variants. Simulation studies demonstrate the superiority of the proposed methods over standard linear regression methods. Data from the Cohorts for Heart and Aging Research in Genomic Epidemiology Targeted Sequencing Study (CHARGE-TSS) are provided.

email: dragontaoran@gmail.com

## 76. STATISTICAL METHODS FOR SURVIVAL ANALYSIS

### Distributional Properties and Peculiarities in HPP-Based Recurrent Event Models

**Piaomu Liu\***, University of South Carolina, Columbia  
**Edsel A. Peña**, University of South Carolina, Columbia

In this pedagogically-oriented paper, distributional properties and some surprising results pertaining to recurrent event models based on homogeneous Poisson process (HPP) and marked Poisson process are obtained. HPPs, both without and with frailties, are observed over a random window. Properties of the number of events seen over the window, the gap-time that covered the termination time, and correlations among gap-times of the observed events are presented. Seemingly peculiar looking results induced by the sum-quota accrual scheme and size-biased sampling are highlighted and discussed. Extensions of results for the HPP-based models are also obtained for a marked Poisson process, which models competing risks with recurrent events. Through the simple HPP-based and marked Poisson process models, it is hoped that a better appreciation of the difficulties when dealing with recurrent events will arise. The results and proofs also highlight the importance of the theorem of total probability, Bayes theorem, the iterated rules of expectation, variance and covariance, the renewal equation, the Cauchy-Schwartz Inequality, and Jensen's Inequality.

email: liu256@email.sc.edu

### Vertical Modeling: Analysis of Multi-State Data with a Cured Fraction

**Mioara Alina Nicolaie\***, Université catholique de Louvain  
**Catherine Legrand**, Université catholique de Louvain

In cancer clinical trials, the assessment of the risk of death due to the disease or of the chance of getting cured plays an important role in treatment selection. Multi-state models with a cured fraction allow one to analyse jointly the development of the disease and the recovery process. In

this presentation, we propose a new approach that extends vertical modeling of Nicolaie et al. (2010) to multi-state models with a cured fraction. The basic idea is to view the multi-state structure as an ensemble of several competing risks components, each of these being modeled by means of semi-parametric vertical modeling technique. This method can accommodate Markov or semi-Markov multi-state models as it provides an alternative to their analysis when either (1) the proportionality assumption on the transition intensities typically used in this context is not fulfilled or (2) there is a cured fraction in the population who will never experience recurrence of the disease. We apply our approach to the analysis of childhood leukemia data from EORTC.

email: mioara.nicolaie@uclouvain.be

### Sequential Stratification for Recurrent Event Outcomes

**Abigail Smith\***, University of Michigan  
**Douglas Shaubel**, University of Michigan

Recurrent events are of increasing interest in observational studies, and in some of these studies the goal is to estimate the effect of a certain treatment on the recurrent event rate. If two or more treatments could potentially occur, but only the first is observed, the ideal comparison is between the treatment of interest and any other potential treatment course. Sequential stratification is a method for estimating the effect of choosing one treatment course relative to waiting and potentially receiving another treatment course. Since the method has only been developed for terminal events such as death, we extend sequential stratification to the recurrent event setting. Asymptotic properties of the proposed estimators are explored. The performance of the method in moderate-sized samples is assessed through simulation. Finally, the proposed methods are applied in a clinical dataset to evaluate the effect of living donor liver transplantation on hospitalization rates.

email: abbysmit@umich.edu

### Random Survival Forests for Interval-Censored Outcomes in the Presence of Imperfect Diagnostic Tests

**Hui Xu\***, University of Massachusetts, Amherst  
**Xiangdong Gu**, University of Massachusetts, Amherst  
**Raji Balasubramanian**, University of Massachusetts, Amherst

In many epidemiologic settings, such as the Women's Health Initiative (WHI), the occurrence of a silent event such as diabetes is ascertained through error-prone procedures such as self-reports. For this setting, we propose a modification to the Random Survival Forests (Ishwaran, H., 2008) (RSF) algorithm by incorporating a likelihood function that accounts for the error-prone nature of the diagnostic tests. To evaluate the performance of our proposed algorithm, we simulated datasets of 100 subjects and 100 features per subject, of which 5 were assumed to be true biomarkers that influence the risk of the event of interest through a proportional hazards model. The parameter settings were selected to mimic diabetes self-reported outcomes in the WHI. Averaging across 100 simulated datasets, the

proportion of times the true biomarkers were ranked among the top 5 features was 0.654 and 0.735, by RSF and our approach, respectively. The proposed algorithm will be applied to GWAS data from the WHI (approximately 12,000 subjects), where the outcome of interest is diabetes that is ascertained through self-report.

email: huix@schoolph.umass.edu

### **Analysis of MD STARnet Data**

**Ke Liu\***, University of Iowa

**Ying Zhang**, University of Iowa

**Paul Romitti**, University of Iowa

**Soman Puzhankara**, University of Iowa

**Kristin Caspers**, University of Iowa

**Elinora Price**, University of Arizona

**Jennifer Andrews**, University of Arizona

**Chris Cunniff**, University of Arizona

Duchenne/Becker Muscular Dystrophy (DBMD) is a recessive X-linked form of muscular dystrophy which results in progressive muscle degeneration and eventual death in young adulthood. Individuals with DBMD are at risk for developing scoliosis; thus, we need to understand the medical practices used to monitor the health status of affected individuals in a well-defined population. The Muscular Dystrophy Surveillance, Tracking, and Research Network (MD STARnet) is a population-based surveillance system for DBMD operating in five states (Arizona, Colorado, Georgia, Iowa, and New York). Information on development and treatment of co-morbidities, including scoliosis, was collected. We analyzed factors associated with the timing (age) of examination for scoliosis and results (Cobb angles) of lumbar radiographs. The age at first radiograph to detect scoliosis was varied by phenotype and race/ethnicity. State differences may be influenced by the race/ethnic composition of their populations. Finds also showed that besides age, the phenotype of DBMD, intensity of steroids usage, were significantly associated with subjects Cobb angle. Furthermore, any steroid use and early-onset of DBMD symptoms strongly impacted the hazard risk for an individual to cease ambulation.

email: ke-liu@uiowa.edu

### **Time-Dependent Tree-Structured Survival Analysis with Unbiased Variable Selection**

**Meredith L. Wallace\***, University of Pittsburgh

Incorporating time-dependent covariates into tree-structured survival analysis (TSSA) may result in more accurate prognostic models than if only baseline values are used. Available time-dependent TSSA methods exhaustively test every binary split on every covariate; however, this approach may result in selection bias towards covariates with more observed values. We propose a novel method that uses unbiased significance levels from permutation tests to select the time-dependent or baseline covariate with the strongest relationship with the survival outcome. The specific splitting value is then identified using only the selected covariate. Simulation results show that the proposed time-dependent TSSA method produces tree models of equal or greater accuracy as compared to baseline TSSA models, even

with high censoring rates and large within-subject variability in the time-dependent covariate. To illustrate, the proposed method is applied to data from a cohort of bipolar youth to identify subgroups at risk for self-injurious behavior.

email: lotzmj@upmc.edu

### **Quantile Regression in Semiparametric Varying-Coefficient Partially Linear Models for Right Censored Length-Biased Data**

**Xuerong Chen**, Georgetown University

**Yeqian Liu\***, University of Missouri, Columbia

**Jianguo Sun**, University of Missouri, Columbia

**Yong Zhou**, Chinese Academy of Sciences, Beijing

This paper discusses regression analysis of right-censored failure time data arising from cross-sectional prevalent cohort studies. It is well known that in these cases, the data could be subject to length-biased sampling and thus it can be difficult to model risk factors on the unbiased failure time for general populations. Most recent studies on this subject adopt the accelerated failure time model or proportional hazards model for the failure time and assume the independence between censoring variable and the covariates of interest. In this paper, we consider a semiparametric partially linear model to assess covariate effects on the population failure times by modeling the length-biased times, and develop a varying coefficient quantile regression approach to obtain the consistent estimators of the regression coefficients. Our approach directly estimates the conditional quantiles of survival times based on a flexible additive-linear model and allows the censoring variable to be informative about the covariates. The large sample properties of the estimators are established and a computationally efficient and stable estimating procedure is also developed via the majorize-minimize algorithm. Numerical results are obtained from a simulation study conducted to assess the performance of the approach and an illustrative example.

email: yldg5@mail.missouri.edu

## **77. PRESIDENTIAL INVITED ADDRESS**

### **A Significance Test for the LASSO**

**Robert J. Tibshirani, PhD**, Stanford University

In this talk, I consider testing the significance of the terms in a fitted regression, fit via the lasso. I propose a novel test statistic for this problem, and show that it has a simple asymptotic null distribution. This work builds on the least angle regression approach for fitting the lasso, and the notion of degrees of freedom for adaptive models (Efron 1986) and for the lasso (Efron et. al 2004, Zou et al 2007). I give examples of this procedure, discuss extensions to generalized linear models and the Cox model, and describe an R language package for its computation. In addition, generalizations to a broad range of adaptive fitting such as graphical models and clustering will be outlined. This work is joint with Richard Lockhart (Simon Fraser University), Jonathan Taylor (Stanford University) and Ryan Tibshirani (Carnegie Mellon University).

e-mail: tibs@stanford.edu

## 78. JABES INVITED SESSION

**Estimating Velocity for Processive Motor Proteins with Random Detachment**

**John Hughes\***, University of Minnesota  
**Shankar Shastry**, The Pennsylvania State University  
**William O. Hancock**, The Pennsylvania State University  
**John Fricks**, The Pennsylvania State University

Processive motor proteins are ATP-powered biological nanomachines that drive many forms of movement in living organisms. The existence of eukaryotic organisms depends on these tiny motors because the passive process of diffusion is not sufficient to transport unwieldy payloads within the cell in a timely fashion. A motor protein overcomes these difficulties by hydrolyzing ATP in order to tow a cargo rapidly and in a directed path along a suitable substrate. Knowledge of these motors could lead to important biomedical applications, e.g., anti-tumor technologies; treatments for neurodegenerative diseases; devices for blood testing and genetic screening; and treatments for diseases caused by motor protein defects. We show that, for a wide range of models, the empirical velocity of processive motor proteins has a limiting Pearson type VII distribution with finite mean but infinite variance. We develop maximum likelihood inference for this Pearson type VII distribution. In two simulation studies, we compare the performance of our MLE with the performance of standard Student's t-based inference. The studies show that incorrectly assuming normality (1) can lead to imprecise inference regarding motor velocity in the one-sample case, and (2) can significantly reduce power in the two-sample case. These results should be of interest to experimentalists who wish to engineer motors possessing specific functional characteristics.

email: [hughesj@umn.edu](mailto:hughesj@umn.edu)

**Bayesian 2-Stage Space-Time Mixture Modeling with Spatial Misalignment**

**Andrew B. Lawson\***, Medical University of South Carolina  
**Jungsoon Choi**, Hanyang University, Korea  
**Bo Cai**, University of South Carolina, Columbia  
**Monir Hossain**, University of Cincinnati  
**Russell Kirby**, University of South Florida  
**Jihong Liu**, University of South Carolina, Columbia

We develop a new Bayesian two-stage space-time mixture model to investigate the effects of air pollution on asthma. The two-stage mixture model proposed allows for the identification of temporal latent structure as well as the estimation of the effects of covariates on health outcomes. In the paper, we also consider spatial misalignment of exposure and health data. A simulation study is conducted to assess the performance of the 2-stage mixture model. We apply our statistical framework to a county-level ambulatory care asthma data set in the US state of Georgia for the years 1999-2008.

email: [lawsonab@musc.edu](mailto:lawsonab@musc.edu)

**Identifying Genes that are Differentially Expressed in Both of Two Independent Experiments**

**Megan Orr\***, North Dakota State University  
**Peng Liu**, Iowa State University  
**Dan Nettleton**, Iowa State University

Identifying genes that are differentially expressed (DE) in two independent experiments generally involves two steps. In the first step, gene expressions from each experiment are analyzed separately to produce a list of genes that are declared DE in each experiment while controlling false discovery rate at some desired level  $\alpha$ . Then, genes common to both lists are declared to be DE in both experiments. We call this approach the "intersection method". Little, if any, research has been done to evaluate how well this method controls the false discovery rate (FDR) or ranks genes based on significance. In addition to exploring these questions, we also propose a new method for estimating FDR. These two methods, as well as another method developed with a different goal in mind, are compared through two simulation studies, one involving independent normal data and one involving real gene expression data. These simulation studies demonstrate the advantages of the proposed method. We conclude the paper by providing an analysis of data from two experiments involving gene expressions in maize leaves.

email: [megan.orr@ndsu.edu](mailto:megan.orr@ndsu.edu)

**A Bayesian Approach to Fitting Gibbs Processes with Temporal Random Effects Generalisations and Challenges**

**Ruth King**, University of St Andrews  
**Janine B. Illian\***, University of St Andrews  
**Stuart E. King**, University of St Andrews  
**Glenna F. Nightingale**, University of St Andrews  
**Ditte K. Hendrichsen**, Norwegian Institute for Nature Research

While spatial point processes are relevant to many dynamic systems, the literature on spatial point processes has focused primarily on purely spatial models. However, space-time modelling is becoming increasingly relevant in this context mainly due to an improving availability of space-time point pattern data sets from many scientific areas. This talk will look at spatio-temporal point process models and discuss different approaches to constructing and fitting these. We consider spatial point pattern data observed repeatedly over time in an inhomogeneous environment. We develop an integrated approach analysing all snapshots within a single analysis and assume that the spatial point patterns can be regarded as independent replicates, given spatial covariates. We demonstrate how different types of models can be considered within a Bayesian framework and different classes of point process models, Gibbs models and log Gaussian Cox processes, are fitted using MCMC and integrated nested Laplace approximation (INLA), respectively. We will further discuss how the methodology may be generalised to models that do not assume independence among the different patterns - in particular those with temporal dependence amongst the points.

email: [janine@mcs.st-and.ac.uk](mailto:janine@mcs.st-and.ac.uk)

## 79. RECENT ADVANCES IN STATISTICAL METHODS FOR MISSING DATA

### Identification and Multiple Imputation of Implausible Gestational Ages for the Study of Preterm Births

**Nathaniel Schenker\***, National Center for Health Statistics, Centers for Disease Control and Prevention

Gestational age is an important variable in the study of infant health. However, information on gestational age compiled from birth records has been known to have inaccuracies, largely due to mis-estimation of gestational ages based on time of last menstrual period. Simply deleting implausible cases from analyses can lead to loss of information as well as bias. This talk will describe research on methods for (a) identifying implausible reported gestational ages using mixture models for the distribution of birth weights conditional on reported gestational ages, and (b) multiply imputing for implausible reported gestational ages using the mixture models together with prediction models for gestational age. The multiple-imputation framework facilitates the reflection of uncertainty in both the assessment of a reported gestational age as incorrect and the prediction of a true gestational age for an incorrectly reported one.

email: nschenker@cdc.gov

### Adjusting for Verification Bias in Estimation of Covariate-Specific Areas Under the ROC Curves

**Xiao-Hua A. Zhou\***, University of Washington  
**Danping Liu**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

When the covariates affect the accuracy of the test, we need to adjust for covariate effects in estimation of test's accuracy. When some covariates affect test's accuracy, we may represent this effect in two ways. One is to calculate covariate-specific ROC curves, representing the accuracy of the test for a specific sub-population of patients. Another way is to calculate the overall ROC curve over the whole population, adjusting for effects of covariates. Furthermore, in many clinical studies, the true disease status may be subject to missingness because it is expensive and/or invasive to ascertain the disease status. Using only complete-case analysis can lead to biased inferences, also known as 'verification bias'. In this talk, we present several new verification bias corrected estimators for estimating the areas under the covariate-specific and the areas under covariate-adjusted ROC curves (AUC<sub>x</sub> and AAUC). We also present the asymptotic distributions for the proposed estimators and report simulation results on the finite sample performance of the proposed estimators. Finally, we apply our method to a data set in Alzheimer's disease research.

email: azhou@uw.edu

### Multiple Imputation Via Flexible, Joint Models

**Jerome Reiter\***, Duke University

Multiple imputation is a common approach to handling missing values. It is usually implemented using either chained equation approaches (as in the software MICE and IVEWARE) or joint models. For the latter, analysts typically use a multivariate normal or log-linear model as imputation engines. While useful, these models impose sets of assumptions that may not be satisfied in many data sets. In this talk, I describe several imputation engines based on nonparametric Bayesian models. I present comparisons against default (main effects only) implementations of chained equations in simulated and actual data.

email: jerry@stat.duke.edu

### On the Use of Box-Cox Transformation for Missing Data Imputation

**Yulei He\***, National Center for Health Statistics, Centers for Disease Control and Prevention  
**Don Malec**, National Center for Health Statistics, Centers for Disease Control and Prevention  
**Nathaniel Schenker**, National Center for Health Statistics, Centers for Disease Control and Prevention

Multiple imputation is a popular approach to handle missing data problems. To impute missing data for variables with strong skewness and/or kurtosis, transformation is often used to make them approximately normal and with homogeneous variance: the typical imputation procedures are conducted on the transformed scale, and the imputed values are transformed back to the original scale. However, there is a lack of thorough research on the performance of this seemingly straightforward strategy. This research aims to provide some insights to several methodological issues such as (1) Do we need to retain the variability of transformation parameters in imputation, which is often ignored in practice? (2) Can we combine the transformation approach with other robust imputation methods, such as predictive mean matching, to make it more robust? (3) Can we incorporate the transformation strategy into the sequential regression imputation framework for multivariate missing variables with non-normal distributions? Using the well-known Box-Cox transformation as the illustrative method, we conduct simulation studies to assess the performance of the proposed methods. Applications include the missing Dual-Energy X-ray Absorptiometry data from the public use files of National Health and Nutrition Examinations Survey.

email: wdq7@cdc.gov

## 80. BIG DATA METHODS IN BIOSTATISTICS

### Some Post-GWAS Strategies for Identifying the Remaining Genetic Determinants

**Zhaoxia Yu\***, University of California, Irvine

Genome-wide association studies have led to the discovery of hundreds of common variants for complex human traits. Despite these successes, the identified variants to date only explain a small fraction of heritability, leaving the majority of genetic determinants yet to be discovered. Emerging evidence suggests that common variants with small effects, rare variants, and gene-gene interactions may explain the missing heritability. However, existing methods with currently used sample sizes are underpowered to detect these variants. Here we present two strategies to address the unprecedented challenges. In the first strategy we first use a data-driven variant selection method to mine potentially disease-related variants in a unit, such as a gene or pathway, and then combine them to further reduce dimensionality. In the second strategy, we improve statistical power by borrowing information across etiologically similar traits. We demonstrate the usefulness of the strategies through both extensive simulation studies and real examples.

email: zhaoxia@ics.uci.edu

### Interactive, Exploratory Visualization and Statistical Analysis of Genome-Scale Data

**Hector Corrada Bravo\***, University of Maryland, College Park

**Florin Chelaru**, University of Maryland, College Park

Data visualization is an integral aspect of the analysis and dissemination of high-throughput functional genomics experimental results, including transcriptomic and epigenomic assays, commonly used to understand basic molecular mechanisms in disease and development. In this talk, I will introduce interactive visualization methods and systems that provide tight-knit integration with algorithmic and statistical data modeling and analysis as a case-study of why integrating interactive visualization with statistical modeling and analysis is an essential aspect of analysis in big data applications.

email: hcorrada@umiacs.umd.edu

### A Semiparametric Bayesian Model for Detecting Multiway Synchrony Among Neurons

**Babak Shahbaba\***, University of California, Irvine

**Bo Zhou**, University of California, Irvine

**Shiwei Lan**, University of California, Irvine

**Hernando Ombao**, University of California, Irvine

**David Moorman**, Medical University of South Carolina

**Sam Behseta**, Cal State Fullerton

We propose a scalable semiparametric Bayesian model to capture dependencies among multiple neurons by detecting their co-firing (possibly with some lag time) patterns over time. After discretizing time so there is at most one spike at each interval, the resulting sequence of 1's (spike) and 0's (silence) for each neuron is modeled

using the logistic function of a continuous latent variable with a Gaussian process prior. For multiple neurons, the corresponding marginal distributions are coupled to their joint probability distribution using a parametric copula model. These are the advantages of our approach: the nonparametric component (i.e., the Gaussian process model) provides a flexible framework for modeling the underlying firing rates; the parametric component (i.e., the copula model) allows us to make inference regarding the relationships among neurons. Using the copula model, we construct multivariate probabilistic models by separating the modeling of univariate marginal distributions from the modeling of dependence structure among variables. Our method is easy to implement using a computationally efficient sampling algorithm, called Spherical Hamiltonian Monte Carlo. As a result, unlike most existing methods, our approach can be easily extended to high dimensional problems.

email: babaks@uci.edu

### Algebraic Properties and Fast Large Covariance Estimation

**Xi Luo\***, Brown University

It becomes popular to study the relationships between thousands or millions of variables, for example in genetics and in neuroimaging. Sparse inverse covariance estimation or Gaussian graphical models is an important tool for characterizing such relationships. Popular and important approaches include optimizing (regularized) likelihood objective functions, especially when the number of variables far exceeds the sample size. In this talk, I will describe alternative approaches based on algebraic properties of covariance matrices. These approaches introduce new and simplified objective functions, and are useful for big data inference. In particular, they enable fast and large-scale computing of vectors, instead of matrices, while in the mean time retaining the desirable theoretical convergence rates. Moreover, they allow inference of each matrix column separately, and adapt to varying sparsity and magnitude. Numerical merits will be illustrated using simulations and real datasets from genetics and fMRI studies.

email: xi.rossi.luo@gmail.com

## 81. STATISTICAL PREDICTION MODELS FOR MEDICAL DECISION MAKING

### Dynamic Prediction of Survival Outcomes and Medical Decision Making

**Xuelin Huang\***, University of Texas

MD Anderson Cancer Center

**Sangbum Choi**, University of Texas

MD Anderson Cancer Center

**Jing Ning**, University of Texas MD Anderson Cancer Center

This proposal is motivated by the need to monitoring the disease progress of chronic myeloid leukemia patients using their BCR-ABL gene expression levels measured on regular time intervals. We provide real-time dynamic prediction for future prognosis using marginal Cox proportional hazards models over time with constraints. Comparing with separate

landmark analyses on different discrete time points after treatment, our approach can achieve more smooth and robust predictions. Comparing with approaches of joint modeling of longitudinal biomarkers and survival, our approach does not need to specify a model for the changes of the monitoring biomarkers, and thus avoids the need of any kind of imputing of the biomarker values on time points they are not available. This helps eliminate the potential bias introduced by mis-specified models for longitudinal biomarkers.

email: xlhuang@mdanderson.org

### **Statistical Prediction Models for Medical Decision Making**

**Michael W. Kattan\***, Cleveland Clinic

The backbone of medical decision making is the prediction of outcomes in individual patients. Statistical prediction models represent an attractive approach towards personalizing medical decision making due to their improved accuracy over many different alternatives. Such alternatives include constructing risk groups, counting risk factors, and using clinical judgment. However, barriers to implementation of statistical prediction models include difficulty with implementation either as stand-alone software or electronic health record integration. Numerous examples of the improvement of statistical prediction models over alternatives, as well as innovative solutions of statistical prediction model integration, will be provided.

email: kattanm@ccf.org

### **ROC Analysis for Multiple Markers with Tree-Based Classification**

**Mei-Cheng Wang\***, Johns Hopkins University  
**Shanshan Li**, Indiana University Fairbanks School of Public Health

This talk considers receiver operating characteristic (ROC) analysis for bivariate or multiple markers. The research interest is to extend ROC framework from univariate marker setting to bivariate or multivariate marker setting for evaluating predictive accuracy of markers. Using a tree-based and-or classifier, an ROC function together with a weighted ROC function (WROC) and their conjugate counterparts are proposed for examining the performance of multiple markers or tests. The proposed functions evaluate the performance of and-or classifier among all possible combinations of marker values. Specific features of ROC and WROC functions and other related statistics are discussed in comparison with those familiar properties for univariate marker ROC analysis. Nonparametric methods are developed for estimating ROC-related functions, (partial) area under curve and concordance probability. The approach is applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data to illustrate the applicability of the proposed procedures. Extension to a different definition of ROC function with optimized TPR will also be discussed.

email: mcwang@jhsph.edu

### **Efficient Evaluation of Risk Markers for Censored Failure Time Outcome: Analyses and Designs**

**Yingye Zheng\***, Fred Hutchinson Cancer Research Center  
**Tianxi Cai**, Harvard School of Public Health

Validating the clinical usefulness of novel biologic markers for stratifying patients in terms of their future outcomes is a critical component for developing effective strategies for disease prevention and treatment management. Such validation is ideally conducted using data from a prospective cohort, which allows the calculation of distributions of risk for subjects with good and bad outcomes and related summary indices that characterize the predictive capacity of novel biomarkers or risk models. Two-phase study methods, in which more detailed or more expensive exposure information is only collected on a sample of individuals with events and a small proportion of other individuals, has become increasingly important in biomarker validation research. In this talk we consider methods for calculating prediction indices when outcomes are censored failure times and novel biomarkers are ascertained with two-phase study designs. In particular this talk will focus on new statistical approaches that aim to improve the statistical efficiency of prediction index estimators under more practical two-phase study designs. Different two-phase design options will be compared in terms of statistical efficiency and practical considerations in the context of risk marker evaluation.

email: yzheng@fhcrc.org

## **82. RECENT DEVELOPMENTS IN STATISTICAL GENETICS, GENOMICS, AND THEIR APPLICATIONS**

### **Joint Analysis of SNP and Gene Expression Data in Genetic Association Studies of Complex Diseases Using Causal Mediation Analysis**

**Yen-Tsung Huang**, Brown University  
**Tyler VanderWeele**, Harvard School of Public Health  
**Xihong Lin\***, Harvard School of Public Health

Genetic association studies have been a popular approach for assessing the association between common SNPs and complex diseases. However, other genomic data involved in the mechanism between SNPs and disease, e.g., gene expressions, are usually neglected in these association studies. In this paper, we propose to exploit gene expression information to more powerfully test the association between SNPs and diseases by jointly modeling the relations among SNPs, gene expressions and diseases. We propose a variance component test for the total effects of SNPs and a gene expression on disease risk. We cast the test within the causal mediation analysis framework with a gene expression as a potential mediator. For eQTL SNPs, we show that the use of gene expression information can enhance power to test for the total effects of a SNP-set, which are the combined direct effect and indirect effects of the SNPs mediated through the gene expression, on disease risk. As the true disease model

is unknown in practice, we further propose an omnibus test to accommodate different underlying disease models. We evaluate the finite sample performance of the proposed methods using simulation studies. We apply our method to study the effect of ORMDL3 gene on the risk of asthma.

email: xlin@hsph.harvard.edu

### The Magic of Score Statistics

**Danyu Lin\***, University of North Carolina, Chapel Hill

Score statistics are mathematically more accurate and numerically more stable than Wald and likelihood ratio statistics, especially for discrete traits and rare variants. Although taking different forms, all popular gene-level association tests for rare variants, including burden test, weighted sum statistic, combined and multivariate collapsing (CMC) test, variable-threshold (VT) test and sequence kernel association test (SKAT), can be constructed from the score vector for testing the global null hypothesis that none of the variants in the gene are related to the trait. In addition, meta-analysis of score statistics is numerically equivalent to joint analysis of individual participant data. In this talk, we will demonstrate the aforementioned attractive features of score statistics, highlighting meta-analysis of gene-level association tests for rare variants under both fixed-effects and random-effects models and providing applications to the Genetic Investigation of Anthropometric Traits (GIANT) and NHLBI Exome Sequencing Projects. We recommend that score statistics be deposited in public databases to facilitate efficient meta-analysis in the future.

email: lin@bios.unc.edu

### Assessing the Sensitivity of Genetic Associations to Unmeasured Confounding Under a Causal Framework

**Nandita Mitra\***, University of Pennsylvania  
**Elizabeth Handorf**, Fox Chase Cancer Center  
**Peter Kanetsky**, University of Pennsylvania  
**Steve Kawut**, University of Pennsylvania

In population-based genetic association studies, estimates of genetic effects may be biased due to unmeasured confounders. Specifically, in candidate gene studies where ancestry informative markers (AIMs) may not be available or may not fully characterize ancestry, it is important to assess the sensitivity of an estimated genetic effect to hidden bias due to residual population stratification. Using a causal framework, we extend the sensitivity approach of Vanderweele [2008] to provide a closed-form relationship between true and observed genetic effects given the hypothesized distribution of an unmeasured confounder. We show how a weighted sum of AIMs, summarizing residual population stratification, is approximated by the normal distribution, and can be readily incorporated into the sensitivity framework. This results in a simple closed-form relationship, which can be used to adjust the estimated genetic effect and corresponding confidence intervals on binary or continuous outcomes. We assess the performance of this approach by simulation under varying magnitudes

of population stratification. We apply the method to a multi-center study aimed to identify genetic risk factors for cardiovascular and renal complications in patients evaluated for liver transplant.

email: nanditam@upenn.edu

### Robust and Powerful Sibpair Test for Rare Variant Association

**Keng-Han Lin**, University of Michigan  
**Sebastian Zoellner\***, University of Michigan

Investigating the impact of rare variants on complex disease requires large samples, increasing the risk of spurious signals caused by population structure. To overcome such stratification, we propose a novel test based on the observation that risk variants are more likely to occur on chromosomes shared between affected relatives even in the absence of a meaningful linkage signal. In sibpairs, this can be tested by comparing the number of risk alleles located on shared chromosomes and the number of risk alleles on non-shared chromosomes. This test is robust to stratification as siblings have matched ancestries. Moreover, this design can correct for genotyping error and batch effects by examining how often a sibpair who share two IBD chromosomes has inconsistent genotype calls. We adjust the test to take into account the genotype error probabilities. We evaluate the power of this approach analytically as well as with computer simulations using a general model for the effect size of rare risk variants and different models of interaction between the locus of interest and the remaining genome. For most models with cumulative risk allele frequency  $<0.05$ , the proposed design shows superior power over the conventional case-control study given the same number of sequenced samples. Especially, if the effect sizes of different risk variants are unequal, the family sharing test is very powerful.

email: szoellne@umich.edu

## 83. IMPROVED STATISTICAL MODELING AND UNDERSTANDING OF GENE EXPRESSION AND TRANSCRIPTION REGULATION USING NEXT GENERATION SEQUENCING AND OTHER HIGH THROUGHPUT TECHNOLOGIES

### Deconvolution of Base Pair Level RNA-Seq Read Counts for Quantification of Transcript Expression Levels

**Han Wu\***, Purdue University  
**Yu Zhu**, Purdue University

RNA-Seq has emerged as a powerful technique for transcriptome study. One of the important goals is to quantify the expression levels of transcripts using short reads generated from a RNA-Seq experiment. However, the observed short sequence reads are mapped to the genome rather than transcriptome, in other words, the label where each read is from is missing. Thus the quantification of

expression levels of transcripts is a difficult task. A number of methods have been proposed to quantify expression levels of transcripts. However they either do not fully utilize the base pair (bp) level information or assume a linear relationship between the unobserved transcripts' expression levels and the observed counts. In this article, we propose to use individual exonic base pairs as observation units and further proposed the convolution of mixture Poisson (CPM) model to model the base level zero as well as non-zero read counts. Both simulation study and real data application have demonstrated the effectiveness of CPM-Seq. CPM-Seq was shown to produce more accurate and consistent quantification results than Cufflinks.

email: wu79@purdue.edu

### **Accounting for Nuisance Covariates when Using RNA-Seq Data to Identify Differentially Expressed Genes**

**Dan Nettleton\***, Iowa State University  
**Yet Nguyen**, Iowa State University

High-throughput DNA sequencing technologies can be used to identify sequences of bases that occur in samples of RNA. This approach (known as RNA sequencing or RNA-seq for short) provides counts that serve as measures of transcript abundance in a biological sample for each of thousands of genes. The amount of messenger RNA (mRNA) produced by a gene is often referred to as a gene's expression level. Thus, RNA-seq provides expression level measurements for thousands of genes. When RNA-seq technology is applied to multiple independent samples of different types, researchers often want to determine which genes are differentially expressed, i.e., which genes have mean levels of expression that differ across sample types. Analyses can often be complicated by the presence of nuisance factors that arise due to experimental design limitations and heterogeneity of experimental units that can be seen in continuous covariates measured for each experimental unit and/or RNA sample. This talk will present examples of such nuisance covariates and describe inference strategies that account for their effects.

email: dnett@iastate.edu

### **Bayesian Models For Integrative Genomics**

**Marina Vannucci\***, Rice University  
**Alberto Cassese**, Rice University  
**Michele Guindani**, University of Texas MD Anderson Cancer Center

Novel methodological questions are now being generated in the biological sciences, requiring the integration of different concepts, methods, tools and data types. Bayesian methods that employ variable selection have been particularly successful for genomic applications, as they allow to handle situations where the amount of measured variables can be much greater than the number of observations. In this talk I will focus on a model that integrates experimental data from different platforms together with prior knowledge. I will look in particular at Bayesian models that relate genotype data or methylation data to mRNAs, for the selection of the markers

that affect the gene expression. Specific sequence/structure information will be incorporated into the prior probability models. All modeling settings employ variable selection techniques and prior constructions that cleverly incorporate biological knowledge about structural dependencies among the variables.

email: marina@rice.edu

### **Understanding Spatial Organizations of Chromosomes Via Statistical Analysis of Hi-C Data**

**Ming Hu\***, New York University  
**Ke Deng**, Tsinghua University  
**Zhaohui Qin**, Emory University  
**Jun S. Liu**, Harvard University

Understanding how chromosomes fold provides insights into the transcription regulation, hence, the functional state of the cell. Using the next generation sequencing technology, the recently developed Hi-C approach enables a global view of spatial chromatin organization in the nucleus, which substantially expands our knowledge about genome organization and function. However, due to multiple layers of biases, noises and uncertainties buried in the protocol of Hi-C experiments, analyzing and interpreting Hi-C data poses great challenges, and requires novel statistical methods to be developed. This work provides an overview of recent Hi-C studies and their impacts on biomedical research, describes major challenges in statistical analysis of Hi-C data, and discusses some perspectives for future research.

email: ming.hu@nyumc.org

## **84. STATISTICAL CHALLENGES IN PUBLIC HEALTH RESEARCH AT THE CDC**

### **Exploring the Optimal Allocation of Sample Sizes in Dual-Frame RDD Telephone Surveys**

**Haci Akcin**, Centers for Disease Control and Prevention  
**Denise Bradford\***, Northrop Grumman

Random-digit dialing (RDD) telephone surveys have long been used to capture data about a target population. To maintain survey coverage and validity, surveys have had to add cellular telephone households to their samples. The Behavioral Risk Factor Surveillance System (BRFSS), for example, one of the largest state-based RDD telephone surveys, began conducting a large pilot study to collect cell phone data in 2008. In 2011, landline and cell phone data were combined and released for public use. Optimal allocation of samples in dual-frame (cell and landline) telephone surveys, however, is still not well defined. In this study, we examined data from the 2011 BRFSS with different characteristics: landline only, combined data with current allocation, and combined data with proposed optimal allocation. The study determines whether there is a cost-effective and optimal sample design feasible for dual-frame RDD telephone surveys. We will examine and compare the 2011 and 2012 BRFSS data.

email: wpm2@cdc.gov

### Area Level Models for County Level Prevalence Estimates Using Publicly Available BRFSS Data

**Betsy L. Cadwell\***, Centers for Disease Control and Prevention

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

**Lawrence E. Barker**, Centers for Disease Control and Prevention

County-level estimates of disease burden and health behaviors may be effective tools for promoting health and well-being and making health policy at local levels. However, publicly available data sources, such as the Behavioral Risk Factor Surveillance System (BRFSS), are often designed for state level analyses. While BRFSS data can be aggregated over several years to produce direct estimates, such estimates are often only practical for large population size counties. Modeling methods have been developed for deriving county level estimates, but these methods require use of restricted-access data. Here, we propose multi-level Bayesian models for creating county level estimates that only use publically available BRFSS and publicly available county level covariate data. The resulting estimates, due to "borrowing of strength" and inclusion of covariates, improve upon aggregated direct county estimates. Further, unlike aggregated estimates, these estimates are available for all counties. We illustrate these methods by producing county level prevalence estimates for self-reported diagnosed diabetes, 2008-2010. We also discuss models considered and model checking and validation approaches. Results from cross validation indicate models with random county and random state effect and either fixed or random county level covariates yield estimates with high correlation to stable direct estimates,  $r=0.77$  and  $0.78$  respectively.

email: bic6@cdc.gov

### Multiple Imputation of Linked National Health Interview Survey and Medicare Data Files

**Guangyu Zhang\***, National Center for Health Statistics, Centers for Disease Control and Prevention

**Jennifer D. Parker**, National Center for Health Statistics, Centers for Disease Control and Prevention

**Nathaniel Schenker**, National Center for Health Statistics, Centers for Disease Control and Prevention

Record linkage is a valuable tool for combining information from different data sources. The National Center for Health Statistics has developed a record linkage program to link the center's population-based surveys with administrative data, including Medicare data. However, not all survey participants provide key information for record linkage. In addition, for Medicare linkages, data are available for the Fee-for-Service program, but less consistently available for the managed care programs, such as Medicare Advantage. In this talk we discuss multiple imputation of missing data in linked National Health Interview Survey (NHIS)-Medicare files. We study mammography status based on Medicare claims for women 65 years and older. In our study, mammography and Medicare Advantage status are missing for NHIS respondents

not linked to Medicare; and mammography status is missing for some linked respondents who have Medicare Advantage coverage. To address this scenario, we explore three imputation methods: imputing screening status first and then imputing the FFS/MA plan type, or imputing FFS/MA plan type first and then imputing the screening status, or imputing the two longitudinal processes simultaneously. We conduct simulation studies and apply these methods to the linked NHIS-Medicare files.

email: VHA1@cdc.gov

### Using Longitudinal Data Analysis to Link Policy and Legislation to Public Health Impacts

**Simone Gray\***, Centers for Disease Control and Prevention

**Patricia Sweeney**, Centers for Disease Control and Prevention

**Joseph Prejean**, Centers for Disease Control and Prevention

**David W. Purcell**, Centers for Disease Control and Prevention

**Aruna Surendera Babu**, Centers for Disease Control and Prevention

**Brett Williams**, Centers for Disease Control and Prevention

**Jenny Sewell**, Centers for Disease Control and Prevention

**Jonathan Mermin**, Centers for Disease Control and Prevention

Longitudinal data analysis methods are well-suited to evaluate the impact of policy changes. However, a number of statistical challenges exist when trying to link the effects of policies with public health outcomes. When conducting a nationwide ecological analysis, it can be difficult to match policy and public health data because states can fall into 3 categories of implementation; 1) never implemented the policy, 2) implemented the policy prior to the study period, and 3) implemented the policy at various times during the study period. In addition to the data matching challenges, it can be difficult to distinguish the temporal trends of the outcome from the actual effects of the policy. The heterogeneity of the trends and policy effects may also vary by state. As with any analysis that uses repeated measurements over time, it is also important to account for the induced correlation. Using generalized estimating equations, we highlight and address many of these challenges with a data example that focuses on comparing changes in laws criminalizing HIV exposure with HIV and AIDS diagnosis data in all 50 states over 10 years. Findings from this type of analysis can provide information about the utility of laws in supporting public health.

email: simonegray@cdc.gov

## 85. INNOVATIVE BAYESIAN NONPARAMETRICS IN BIOSTATISTICS

### Longitudinal Data Analysis Using a Random Partition Model with Regression on Covariates

**Gary L. Rosner\***, Johns Hopkins University

**Peter Mueller**, University of Texas, Austin

**Fernando Quintana**, Pontificia Universidad Catolica de Chile

**Michael Maitland**, University of Chicago

In this talk I discuss applying Bayesian nonparametric methods to complex problems in biostatistics. A particular example considers inference for longitudinal data based on mixed-effects models with a nonparametric Bayesian prior on the treatment effect. The proposed nonparametric Bayesian prior is a random partition model with regression on patient-specific covariates. The main feature and motivation for the proposed model is the use of covariates with a mix of different data formats and possible high-order interactions in the regression. The regression is not explicitly parametric but is implied by the random clustering of subjects. The motivating application is a study of the hypertensive side effect of an anticancer drug. The study involves blood pressure measurements taken periodically over several 24 hour periods for 54 patients.

email: grosner1@jhmi.edu

### A Bayesian Feature Allocation Model for Tumor Heterogeneity

**Peter Mueller\***, University of Texas, Austin

**Juhee Lee**, University of California, Santa Cruz

**Yuan Ji**, NorthShore University Health System

We develop a feature allocation model for inference on genetic tumor variation. We analyze data on  $S$  SNV's (single nucleotide variants) in  $n$  available samples. We characterize tumor variability by  $C$  hypothetical latent cell types that are defined by the presence of some subset of the recorded SNV's. Assuming that each sample is composed of some sample-specific proportions of these cell types we can then fit the observed proportions of SNV's for each sample. Taking a Bayesian perspective, we proceed with a prior probability model for all relevant unknown quantities, including in particular a prior probability model on the binary indicators that characterize the latent cell types by selecting (or not) the recorded SNV's. Such prior models are known as feature allocation models. We define a simplified version of the Indian buffet process, one of the most traditional feature allocation models.

email: pmueller@math.utexas.edu

### A Bayesian Nonparametric Approach to Monotone Missing Data in Longitudinal Studies with Informative Missingness

**Antonio Linero**, University of Florida

**Michael Daniels\***, University of Texas, Austin

We develop a Bayesian nonparametric model for inference for a longitudinal response in the presence of nonignorable missing data. Our general approach is to first specify a  $\{\em working model\}$  that flexibly models the missingness and full outcome processes jointly. We specify a Dirichlet process mixture of independent models as a prior on the joint distribution of the working model. This aspect of the model governs the fit of the observed data by modeling the observed data distribution as the marginalization over the missing data in the working model. We then separately specify the conditional distribution of the missing data given the observed data and dropout. This approach allows us to identify the distribution of the missing data using identifying restrictions as a starting point. We propose a framework for introducing sensitivity parameters, allowing us to vary the untestable assumptions about the missing data mechanism smoothly through a space. Informative priors on the space of missing data assumptions can be specified to combine inferences under many different assumptions into a final inference. We demonstrate this by applying the method to simulated data (and comparing with standard methods) and to data from two clinical trials.

email: mjdaniels@austin.utexas.edu

### Bayesian Quantile Regression for Censored Data

**Brian J. Reich\***, North Carolina State University

**Luke B. Smith**, North Carolina State University

We propose a semiparametric quantile regression model for censored survival data. Quantile regression permits covariates to affect survival differently at different stages in the follow-up period, thus providing a comprehensive study of the survival distribution. We take a semiparametric approach, representing the quantile process as a linear combination of basis functions. The basis functions are chosen so that the prior for the quantile process is centered on a simple location-scale model, but flexible enough to accommodate a wide range of quantile processes. We show in a simulation study that this approach is competitive with existing methods. The method is illustrated using data from a drug treatment study, where we find that the Bayesian model often gives smaller measures of uncertainty than its competitors, and thus identifies more significant effects.

email: brian\_reich@ncsu.edu

## 86. NEW DEVELOPMENTS IN SURVIVAL ANALYSIS

### A Local Agreement Index Based on Hazard Functions for Survival Outcomes

**Tian Dai\***, Emory University  
**Ying Guo**, Emory University

The need to assess agreement often arises in biomedical and clinical research when measurements are taken by different raters or methods on the same subject. When measuring agreement between survival outcomes, standard agreement measures cannot be directly applied due to the unique features of time-to-event data. In this paper, we propose a chance-corrected local agreement index based on bivariate hazard functions to characterize the local agreement pattern between two survival times within a finite region. We show the proposed index fully captures the dependence structure between bivariate survival times. We develop a nonparametric estimation method for the proposed agreement index based on kernel estimates. Our estimator is shown to be strongly consistent and asymptotically normal. We then evaluate the performance of the proposed estimator through simulation studies and illustrate the method using a prostate cancer data example.

email: tian.dai88@gmail.com

### A Frailty Model for Bivariate Interval-Censored Data Allowing Weak Dependence and Independence

**Naichen Wang\***, University of South Carolina  
**Lianming Wang**, University of South Carolina

Interval-censored data commonly arise in real-life epidemiologic, social, and medical studies, in which participants undergo multiple examinations at different times. The failure time of interest is never observed exactly but is known to fall within some examination times. For multivariate interval-censored failure times, the gamma frailty PH model is widely used but is known to produce a large estimation bias when the events are weakly correlated or independent. In this paper, we propose a mixture of frailty model to solve this issue. Our approach allows one to test independence among the events of interest. A Gibbs sampler is proposed based on a data augmentation and is straightforward to implement. Our method is evaluated by simulation studies and illustrated by a real-life medical data application.

email: wangn@email.sc.edu

### Survival Analysis with Correlated Frailties and the Bootstrap

**J. C. Loredo-Osti\***, Memorial University

Consider the problem of mapping major genes affecting the survival time in a replicable population where two levels of clustering should be considered. Gene mapping tests the association of phenotype with marker data. In modern mapping problems, these hypotheses are not independent nor are their associated test-statistics and their number exceeds the sample size, i.e., the same data is used to test a large number of non-independent and non-nested hypotheses which makes the whole inference problem anomalous. Procedures carried out through methods that rely on the standard assumptions of independence and regularity are affected by these issues. One way to address the problem is resampling. There is an extensive literature regarding the shared frailty model and its associated methods of estimation and inference, although all these methods and procedures assume that the frailties are independent (which in a genetics context, it implies ignoring the polygenic background and kinship between individuals); thus, in order to be usable in mapping studies, these methods have to be modified or extended. Here, I discuss some issues with the current resampling procedure and propose the implementation of a bootstrap procedure to address the problem and provide an example using recombinant congenic strains.

email: jcloredoosti@mun.ca

### Semiparametric Methods to Contrast Restricted Mean Gap Times

**Xu Shu\***, University of Michigan  
**Douglas E. Schaubel**, University of Michigan

Times between successive events (i.e., gap times) are of great importance in survival analysis. Very few existing methods allow for comparisons between gap times. Motivated by the comparison of primary and repeat transplantation, our interest is specifically in contrasting the gap time survival functions. Two major challenges in gap time analysis are non-identifiability of the marginal distributions and the existence of dependent censoring (for all but the first gap time). We use Cox regression to estimate the (conditional) survival distributions of each gap time (given the previous gap times). Combining fitted survival functions based on those models, along with multiple imputation applied to censored gap times, we then contrast the first and second gap times with respect to survival and restricted mean lifetime. Large-sample properties are derived, with simulation studies carried out to evaluate finite-sample properties. We apply the proposed methods to kidney transplant data obtained from a national registry.

email: shuxu@umich.edu

### **Extending the Peters-Belson Approach for Assessing Disparities to Right Censored Time-to-Event Outcomes**

**Lynn E. Eberly\***, University of Minnesota  
**James S. Hodges**, University of Minnesota  
**Kay Savik**, University of Minnesota  
**Olga Gurvich**, University of Minnesota  
**Donna Z. Bliss**, University of Minnesota

The Peters-Belson method was developed for quantifying and testing disparities between groups, where the disparities are computed from group-specific observed and expected outcomes. This methodology has been developed for linear regression and logistic regression, for both random samples and survey weighted outcomes. For an NIH-funded project assessing racial/ethnic disparities in nursing home care (the REDSKIN Study), we extended the Peters-Belson approach to proportional hazards survival analysis. Our extension includes a test, and a graphical summary of the disparities, using the theory and methods of expected survival based on Cox regression for a reference population. We also discuss a bootstrap approach to account for clustering in the data. We describe the extension, show how we applied it in the REDSKIN Study with clustering by nursing home, and discuss issues in interpretation and implementation.

email: lynn@biostat.umn.edu

### **Consistency on Change-Point Estimators on Hazard Regression Models with Long-Term Survivors and Right Censoring**

**Wei Zhang\***, Florida Atlantic University  
**Lianfen Qian**, Florida Atlantic University

In this paper, we propose a change-point detection algorithm in single change-point hazard regression model for fitting failure times that allows the existence of both susceptibles and long-term survivors, we also show the consistency of the proposed estimators. The proposed method is used to run two real data analysis.

email: wzhang6@fau.edu

### **Nonparametric Estimation of Quantile Residual Life for Length-Biased Survival Data**

**Samia H. Lopa\***, University of Pittsburgh  
**Jong-Hyeon Jeong**, University of Pittsburgh

Length biased data occurs when a prevalent sampling is used to recruit subject into a study that investigates the time from an initial event to a terminal event. In this paper we propose two ways to estimate the quantiles of the residual life time at fixed time points accounting for the length biased and censored nature of the data. We provide the asymptotic properties of these estimators and investigate them through simulation studies considering that the variance of these estimators require density estimation, we suggest an alternate approach taken by Jeong et al. (2008, Biometrics 64, 157-163) to obtain the confidence intervals for available residual function. We apply these methods to a breast cancer dataset from National Surgical Adjuvant Breast and Bowel Project (NSABP).

email: lopa@nsabp.pitt.edu

## **87. CAUSAL INFERENCE**

### **Estimation of the Optimal Regime in Treatment of Prostate Cancer Recurrence from Observational Data Using Flexible Weighting Models**

**Jincheng Shen\***, University of Michigan  
**Lu Wang**, University of Michigan  
**Jeremy M.G. Taylor**, University of Michigan

For many diseases, patients need multiple stages of treatment, then the goal of identifying the optimal dynamic treatment regime is very appealing. The challenge is to find the best regime amongst a set of defined regimes from observational data, in which the regime being followed by each subject is not well characterized. Inverse probability weighting (IPW) based estimators are used in the estimation of causal parameters in defined regimes as an efficient way to utilize information from an observational study. In this paper, we focused on the case where 1) the outcome is time-to-event, and 2) some of the covariates are time-varying, and possibly follow a complicated pattern. We consider a class of dynamic treatment regimes that are fully determined by the longitudinal covariates. A novel Random Forest based inverse probability weighting scheme is proposed to adjust for the complexity in the mechanism of adherence in the observational data while still allowing for some patients to remain treatment-free as defined by the regime. The optimal regime is then identified as the one with the largest restricted mean survival time. The performance of the proposed method is assessed through simulation studies, which are designed to mimic the situation of salvage therapy to reduce the risk of cancer recurrence in prostate cancer. We also apply the method to an observational prostate cancer study.

email: jcshen@umich.edu

### **A Simulation Study of a Multiply-Robust Approach for Causal Inference with Missing Covariates**

**Jia Zhan\***, Indiana University School of Medicine  
**Changyu Shen**, Indiana University School of Medicine  
**Lingling Li**, Harvard Medical School  
**Xiaochun Li**, Indiana University School of Medicine

Confounding bias and missing data are two major barriers to valid comparative effectiveness studies using observational data. Each respective problem has been extensively studied, however, a principled approach to causal inference on data with missing covariates is still lacking. We have developed a unified multiply-robust (MR) methodology to simultaneously handle both issues. Our MR method builds upon the well-established doubly-robust theory and is 4-fold robust in that it is consistent and asymptotically normal if at least one of four sets of modeling assumptions holds. In this simulation study, we assess the finite sample performance of MR under various realistic scenarios. For comparison we also include results from the full data likelihood and the complete case approaches. Our simulation results show that the MR approach has reasonable finite-sample performance and is 4-fold robust in most considered settings. It is much more

robust to model misspecification than the complete-case approach and the likelihood based approach. The coverage probability based on an asymptotic approximation is around the nominal level with realistic sample sizes.

email: jiazhan@uemail.iu.edu

### Estimating Causal Treatment Effect for Complex Intervention Study Designs

**Pan Wu\***, Christiana Care Health System

Complex intervention study designs, such as multi-layered, multi-level intervention, adaptive or semi-randomized trials, are emerging and increasingly popular in behavioral, psychosocial, and policy research. Estimating causal treatment effect for such kind of studies is becoming a central issue for investigators to identify the causal pathways and mechanisms of interventions by controlling for measurable confounding variables. Some existing approaches of causal inference for selection bias, treatment non-compliance, or informative missing data in simple intervention studies are unsuitable for such complex treatment trials, due to inappropriate model assumptions and low efficiency of model estimation. In this talk, I will discuss a new class of Structural Functional Response Models (SFRM) to address causal treatment effects in one of such complex study designs, which has multiple intervention layers and involve pre- or/and post-treatment confounders from different intervention layers. The new approach not only provides an integrated framework to address multiple sources of confounders in estimating causal treatment effect, but also applies to a wider class of data distributions with high computational efficiency. I will also demonstrate the performance of the proposed method in a real community-based, multi-layered randomized intervention trial.

email: pwu@christianacare.org

### Regression Analysis of Sequentially Randomized Trials Through Artificial Randomization

**Semhar B. Ogbagaber\***, University of Pittsburgh  
**Abdus S. Wahed**, University of Pittsburgh

Adaptive treatment strategies (ATs) are decision rules that take in inputs such as patient characteristics, covariate history, and previous treatments, and output a treatment option. ATs are often compared via sequentially multiple assignment randomized trials (SMARTs). Hypothesis testing to compare adaptive treatment strategies are usually based on inverse weighting and g-estimation. However, regression methods that allow for comparison of treatment strategies that flexibly adjust for baseline covariates are not as straightforward due to the fact that one patient can belong to multiple strategies. For instance, in a two-stage SMART design, one may be tempted to compare the four strategies: A1B1, A1B2, A2B1, A2B2 using a regression model where strategy A<sub>j</sub>B<sub>k</sub> is defined as "if responds to A<sub>j</sub> continue the same initial treatment, otherwise switch to B<sub>k</sub>". Note that a patient responding to A<sub>1</sub> is consistent with both strategies A1B1 and A1B2 which poses a challenge for data analysts as it violates basic assumptions of regression modeling of unique group membership. In this paper, we propose an artificial randomization technique to make the data appear that each subject belongs to a specific ATs. This enables

treatment strategy indicators to be inserted as covariates in a regression model. The properties of this method are investigated analytically and through simulation.

email: sbo8@pitt.edu

### Why Do Treatments Work Differently for Some People? Understanding Treatment-Effect Mechanisms in Stratified Medicine

**Sabine Landau\***, King's College London  
**Richard Emsley**, University of Manchester, United Kingdom  
**Hanhua Liu**, University of Manchester, United Kingdom  
**Graham Dunn**, University of Manchester, United Kingdom

The development of stratified medicine is intrinsically based on a theory regarding treatment-effect mechanisms (effects on therapeutic targets that mediate the effect of the treatment on clinical outcomes). Yet the evaluation of these mechanisms is often absent from the design and analysis of stratified medicine studies, and even if present, is subject to unmeasured confounding between mechanism and outcome. For experimental settings, we apply methods from the causal mediation literature to evaluate mechanisms in the presence of hidden confounding. We illustrate the potential of the predictive biomarker-stratified trial design, together with baseline measurement of all known prognostic markers, to enable the evaluation of both, the utility of the predictive biomarker in such a stratification, and the estimation of the portion of the treatment effect explained by changes in the putative mediator. We call this a biomarker-stratified efficacy and mechanisms evaluation (BS-EME) trial design. Using Monte Carlo simulation we show that the recommended instrumental variables analysis approach provides consistent estimates, with adjustments for all known prognostic markers increasing precision. We also investigate the impact of misclassification of the predictive marker on bias and efficiency. We conclude that valid treatment-effect mediation assessment adds credibility to stratification evaluation with the BS-EME design.

email: sabine.landau@kcl.ac.uk

### Inference for Surrogate Endpoint Validation in the Binary Case

**Ionut Bebu\***, Uniformed Services University of the Health Sciences  
**Thomas Mathew**, University of Maryland Baltimore County  
**Brian K. Agan**, Uniformed Services University of the Health Sciences

This article investigates surrogate endpoint validation in the case of a binary surrogate endpoint and a binary true endpoint, using the criteria of proportion explained (PE) and the relative effect (RE). Logistic regression models suggested in the literature are used for this purpose, and the PE and RE are functions of the parameters appearing in the model. The concepts of generalized confidence intervals and fiducial intervals are used for computing confidence intervals for PE and RE, and their performance is numerically investigated. The numerical results indicate that the proposed confidence intervals are quite satisfactory in terms of maintaining the coverage probability, whereas the intervals based on Fieller's

theorem and the delta method fall short in this regard. This indicates that the conclusion regarding the validation of a surrogate can change depending on the method of inference used, underscoring the importance of using an accurate method of inference for surrogate endpoint validation. It is further noted that our methodology can be applied to interval estimation problems in a causal inference based approach to mediation.

email: ibebu@idcrp.org

### **Longitudinal Analyses of the Causal Path Between Multiple Sclerosis and Depression Using Structural Equation Modeling**

**Douglas Gunzler\***, Case Western Reserve University

While multiple sclerosis (MS) patients commonly experience depressive symptoms, clinicians cannot reliably distinguish the indirect pathways through which different trajectories of MS leads to depression over time. In this talk, a longitudinal structural equation modeling (SEM)-based approach, latent growth modeling (LGM) in the multilevel modeling framework, will be discussed for examining the hypothesized indirect pathways through which MS, as defined by type and baseline time since diagnosis, leads to depression using the Knowledge Program (KP) at the Cleveland Clinic's Neurological Institute data base. SEM is a very general technique combining complex path models with latent (unobserved) variables. LGM is a practical application of SEM for longitudinal data to estimate growth trajectory, analogous to mixed effects modeling in more traditional analyses. The KP links patient-reported depression (via the PHQ-9) responses to the EPIC EHR and provides a powerful opportunity to study and improve patient care and clinical research. SEM is a very appropriate approach to handling the patient reported outcomes, latent variables, causality questions and irregular follow-up times in the KP data base.

email: dgunzler@metrohealth.org

## **88. NON-PARAMETRIC ANALYSIS OF BIOMEDICAL DATA**

### **A Spatio-Temporal Nonparametric Bayesian Variable Selection Model of fMRI Data for Clustering Correlated Time Courses**

**Linlin Zhang\***, Rice University

**Michele Guindani**, University of Texas  
MD Anderson Cancer Center

**Marina Vannucci**, Rice University

We present a novel Bayesian nonparametric model for the analysis of functional magnetic resonance imaging (fMRI) data. Our goal is to provide a joint analytical framework that allows to detect regions of the brain which exhibit neuronal activity in response to a stimulus and, simultaneously, infer the association of spatially remote voxels that exhibit fMRI time series with similar characteristics. We account for the complicated temporal and spatial correlation structure of the experiments by employing appropriate prior distributions which embody our knowledge about the structure of the

brain. We use Markov Chain Monte Carlo (MCMC) techniques for posterior inference, and explore the performance of the proposed model on simulated data and real fMRI data.

email: lz17@rice.edu

### **Inferences About the Mean Area Under the Curve in Pre-Clinical Destructive Sampling Designs**

**Yi Shi\***, State University of New York at Buffalo

**Rameela Chandrasekhar**, Vanderbilt University

**Alan Hutson**, State University of New York at Buffalo

**Gregory Wilding**, State University of New York at Buffalo

Destructive sampling designs are commonly used in pre-clinical pharmacokinetic and toxicology studies. Although the individual area under the curve (AUC) is not accessible under the destructive sampling design, statistical inference regarding the mean AUC is still desirable. Traditionally, parametric methods are applied to perform inferences on the mean AUC based on assumed normality. Although such methods perform well with larger sample sizes, the pre-clinical experiments of interest are often associated with small sample size and oftentimes highly skewed data. We investigated several methods to make inference on the mean AUC, which can be expressed in the form of the linear combination of log-normal distributed random variables. Inferential properties of our approach are compared with those from standard methods of analysis using Monte Carlo simulation studies. The robustness of the methods are also evaluated when distribution assumptions are violated.

email: iamshiyi@gmail.com

### **Investigating a Method for Testing a Hypothesis About the Ratio of Two Medians Using Conover's Rank Transformation Method**

**Donald J. Schuirmann\***, U.S. Food and Drug Administration

Suppose we have two populations with medians,  $M_1$  and  $M_2$  respectively, that we wish to compare. If we are unsure about a normality assumption, we might wish to use Conover's Rank Transformation method (e.g. Conover and Iman 1981) to do the analysis. Suppose we wish to test  $H_0: M_1/M_2 \leq R$  against  $H_1: M_1/M_2 > R$ , for specified  $R$ . One possible approach would be to multiply the observations from population 2 by  $R$ . Let  $M_1^*$  and  $M_2^*$  be the medians of the modified populations (note that  $M_1^* = M_1$  and  $M_2^* = R M_2$ .) At the boundary of our original null hypothesis,  $M_1/M_2 = R$ , we would have  $M_1^* = M_2^*$ , so we could test our original  $H_0$  by testing the hypothesis  $H_0^*: M_1^* = M_2^*$  using the Rank transformation method on the modified dataset. A possible concern with this approach is that multiplying the observations from population 2 by  $R$  will modify their scale. Here, we investigate the impact this may have on the level of significance of the approach, for the specific case of  $R=0.80$  or  $R=1.25$ . A number of underlying probability distributions are considered.

email: donald.schuirmann@fda.hhs.gov

### Restricted Confidence Intervals for Ordered Binary and Survival Data

**Yongseok Park\***, University of Pittsburgh  
**Jeremy M G Taylor**, University of Michigan

In biomedical research, we often encounter situations where there is prior knowledge about the ordering of underlying parameters. In this situation, by appropriately utilizing this ordering information in the estimation process, we can potentially obtain more precise estimators with less variability, particularly when the parameters are close to each other in small sample cases. We consider the problem of constructing restricted confidence intervals when there are groups of observations with an associated parameter for each group, and the parameter values are known to be monotonically ordered. Specifically, we consider the probability parameters for binomial data and the survival rates for censored time-to-event data. Following the procedure introduced by Park, Kalbfleisch and Taylor (2014) for constructing confidence intervals for order restricted normal means, we introduce intermediate random variables to be used to construct restricted confidence intervals. Simulation study shows the proposed method yields narrower intervals than the unrestricted ones and have good coverage rates even when the sample sizes are relatively small.

e-mail: yongpark@umich.edu

### Novel Algorithm for Stratifying Patients into Survival Risk Groups Using Mutation Data at Selected Genes

**Irina Ostrovnyaya\***, Memorial Sloan-Kettering Cancer Center

**Sean Devlin**, Memorial Sloan-Kettering Cancer Center  
**Mithat Gonen**, Memorial Sloan-Kettering Cancer Center

Many cancer studies routinely sequence commonly mutated genes in hopes that patterns of somatic mutations in the tumor may affect patient prognosis. These mutations might co-occur or be mutually exclusive, and they might affect the survival through both main effects and interactions. We consider the problem of building a prognostic model that will define a patient's survival risk based on the somatic mutations that are observed in the patient's tumor. Various decision tree methods are available for such problem. We developed an alternative algorithm that clusters each mutational combination into a small number of risk groups based on its survival distribution. We will show the operating characteristics of this method under various conditions and identify the situations in which it outperforms standard recursive algorithms.

e-mail: ostrovni@mskcc.org

### Two-Sample Parameter Estimation Using Empirical Characteristic Functions

**Cornelis J. Potgieter\***, Southern Methodist University  
**Fred Lombard**, North-West University, Potchefstroom, South Africa

Two random variables  $X$  and  $Y$  belong to the same location-scale family if there are constants  $\lambda$  and  $\tilde{A}$  such that  $Y$  and  $\lambda + \tilde{A}X$  have the same distribution. Potgieter and Lombard (2012) considered non-parametric estimation of these parameters using asymptotic likelihood considerations under minimal assumptions regarding the form of the distribution functions of  $X$  and  $Y$ . We consider here an approach to the estimation problem that is based on minimizing a distance function between empirical characteristic functions of  $X$  and  $Y$ . Asymptotic properties of the estimators are considered and the estimators are also shown to often be near optimal when compared to fully parametric methods. Small-sample performance of the estimators is also considered in a series of Monte Carlo simulations.

e-mail: cpotgieter@smu.edu

### Multiple Imputation Methods for Nonparametric Inference on Cumulative Incidence with Missing Cause of Failure

**Minjung Lee\***, Seoul National University

**James J. Dignam**, University of Chicago  
**Junhee Han**, University of Arkansas, Fayetteville

We propose nonparametric inferences for cumulative incidence estimation when causes of failure are unknown or missing for some subjects. Under the missing at random assumption, we estimate the cumulative incidence function using multiple imputation methods. We develop asymptotic theory for the cumulative incidence estimates obtained from multiple imputation methods. We also discuss how to construct confidence intervals for the cumulative incidence function and perform a test for comparing the cumulative incidence functions in two samples. Through simulation studies, we show that the proposed methods perform well. The methods are illustrated with data from a randomized clinical trial in early stage breast cancer.

e-mail: minjung.lee09@gmail.com

## 89. HIGH DIMENSIONAL IMAGING DATA

### A Parallel Group Independent Component Analysis Algorithm

**Shaojie Chen\***, Johns Hopkins University  
**Lei Huang**, Johns Hopkins University  
**Huitong Qiu**, Johns Hopkins University  
**Ani Eloyan**, Johns Hopkins University  
**Brian Caffo**, Johns Hopkins University

Independent component analysis (ICA) is widely used for blind source separation, especially in the field of functional neuroimaging. Existing ICA algorithms fail when used to analyze massive data with a large number of subjects as the nature of the algorithms does not immediately scale. This is problematic, as new data sets are being compiled with thousands of resting state subjects. Parallel computing has been widely adopted in scientific computing community since the 1950s. Different levels of parallel models, including thread level, process level, cluster level and most recently GPGPU level paralleling, can greatly improve scalability and speed. In this manuscript, a likelihood-based scalable two-stage iterative true group ICA methodology is reviewed and built upon, which naturally fits the mechanism of the MPI model for distributed memory clusters. With this algorithm, brain networks were estimated by analyzing a resting state fMRI dataset of over 500 subjects within a reasonable time. Algorithmic performance is compared to other algorithms.

e-mail: schen89@johnshopkins.edu

### Ultra-High Dimensional Test Via Sparse Projections

**Qiang Sun\***, University of North Carolina, Chapel Hill  
**Hongtu Zhu**, University of North Carolina, Chapel Hill  
**Joseph G. Ibrahim**, University of North Carolina, Chapel Hill

The aim of this paper is to develop a sparse projection regression modeling (SPReM) framework to perform multivariate regression modelling with a large number of responses and a multivariate covariate of interest. We propose two novel heritability ratios to simultaneously perform dimension reduction, response selection, estimation, and testing, while explicitly accounting for correlations among multivariate responses. Our SPReM is devised to specifically address the low statistical power issue of many standard statistical approaches, such as the Hotelling's  $T^2$  test statistic or a mass univariate analysis, for high dimensional data. We formulate the estimation problem of SPReM as a novel sparse unit rank projection (SURP) problem and propose a fast optimization algorithm for SURP. Furthermore, we extend SURP to the sparse multi-rank projection (SMURP) by adopting a sequential SURP approximation. Theoretically, we have systematically investigated the convergence properties of SURP and the convergence rate of SURP estimates. Our simulation results and real data analysis have shown that SPReM outperforms other state-of-the-art methods.

e-mail: qsun@live.unc.edu

### Statistical Approaches for Exploring Brain Connectivity with Multi-Modal Neuroimaging Data

**Phebe B. Kemmer\***, Emory University  
**Ying Guo**, Emory University  
**F. DuBois Bowman**, Columbia University

By combining various types of neuroimaging data, multimodal imaging analyses enable us to study the relationship between brain structure and function, and investigate the connectivity disruption pathways that characterize certain brain diseases. We develop a novel measure, sSC, to quantify the strength of structural connectivity (SC) underlying functional networks identified using data-driven methods such as independent component analysis (ICA). The sSC statistic can be defined on both the voxel- or region-level using diffusion tensor tractography. We provide a framework to conduct statistical inference for sSC, which overcomes many computational challenges due to spatial correlations within the data and the estimation of a large variance-covariance matrix. We will discuss our estimation methods and illustrate the application using an fMRI dataset.

e-mail: brennep@gmail.com

### Spatially Regularizing High Angular Resolution Diffusion Imaging

**Shangbang Rao\***, University of North Carolina, Chapel Hill  
**Hongtu Zhu**, University of North Carolina, Chapel Hill  
**Jian Cheng**, University of North Carolina, Chapel Hill  
**Pew-Thian Yap**, University of North Carolina, Chapel Hill  
**Joseph Ibrahim**, University of North Carolina, Chapel Hill

High angular resolution diffusion imaging (HARDI) has recently been of great interest in mapping the orientation of intra-voxel crossing fibers, and such orientation information allows one to infer the connectivity patterns prevalent among different brain regions and possible changes in such connectivity over time for various neurodegenerative and neuropsychiatric diseases. The aim of this paper is to propose a penalized multi-scale adaptive regression model (PMARM) framework to spatially and adaptively infer the orientation distribution function (ODF) of water diffusion in regions with complex fiber configurations. In PMARM, we reformulate the HARDI imaging reconstruction as a weighted regularized least-squares regression (WRLSR) problem. Similarity and distance weights are introduced to account for spatial smoothness of HARDI, while preserving the unknown discontinuities (e.g., edges between white matter and grey matter) of HARDI. The L1 penalty function is introduced to ensure the sparse solutions of ODFs, while a scaled L1 weighted estimator is calculated to correct the bias introduced by the  $L_1$  penalty at each voxel. In PMARM, we integrate WRLSR with the propagation-separation method (polzehl 2000) to adaptively estimate ODFs across voxels. Experimental results indicate that PMARM can reduce the angle detection errors on fiber crossing area and provides more accurate reconstructions than standard voxel-wise methods.

e-mail: srao@live.unc.edu

### Parametrization of White Matter Manifold-Like Structures Using Principal Surfaces

**Chen Yue\***, Johns Hopkins University  
**Vadim Zipunnikov**, Johns Hopkins University  
**Pierre-Louis Bazin**, Max Planck Institute  
**Dzung Pham**, National Institute of Neurological Disorders and Stroke, National Institutes of Health  
**Daniel S. Reich**, National Institute of Neurological Disorders and Stroke, National Institutes of Health  
**Ciprian Crainiceanu**, Johns Hopkins University  
**Brian Caffo**, Johns Hopkins University

In this manuscript, we are concerned with data generated from a diffusion tensor imaging (DTI) experiment. The goal is to parameterize manifold-like white matter tracts, such as the corpus callosum, using principal surfaces. The problem is approached by finding a geometrically motivated surface-based representation of the corpus callosum and visualized fractional anisotropy (FA) values projected onto the surface; the method applies to any other diffusion summary as well as to other white matter tracts. An algorithm is provided that 1) constructs the principal surface of a corpus callosum; 2) flattens the surface into a parametric 2D map; 3) projects associated FA values on the map. The algorithm is applied to a longitudinal study containing 466 diffusion tensor images of 176 multiple sclerosis (MS) patients observed at multiple visits. For each subject and visit the study contains a registered DTI scan of the corpus callosum at roughly 20,000 voxels. Extensive simulation studies demonstrate fast convergence and robust performance of the algorithm under a variety of challenging scenarios.

e-mail: cyue@jhsp.edu

### Predicting Enhancement in Magnetic Resonance Images Using Scan Stratified Case Control Sampling

**Gina-Maria Pomann\***, North Carolina State University  
**Elizabeth M. Sweeney**, Johns Hopkins University  
**Russel (Taki) Shinohara**, University of Pennsylvania  
**Ana-Maria Staicu**, North Carolina State University  
**Daniel S. Reich**, National Institute of Neurological Disorders and Stroke, National Institutes of Health

Multiple sclerosis (MS) is a disease associated with inflammation of the brain, morbidity, and disability. Irregular blood flow in white matter of the brain is associated with clinical relapses in patients with MS. This abnormality in active MS lesions is commonly evaluated through identification of enhancing voxels on Magnetic Resonance (MR) images. Current clinical practice includes intravenous injection of contrast agents that are occasionally toxic to the patient and can increase the cost of imaging by over 40%. Local image regression methodology has recently been developed to predict enhancing MS lesions without using such contrast agents. We extend this model to account for the rarity of enhancement as well as to incorporate historical information about lesions. Incorporation of a historical covariate which characterizes newly enhancing lesion

behavior is found to substantially improve prediction. We consider 77 brain MR imaging studies on 15 patients which include historical imaging covariates along with T1-w, T2-w images. Additionally, we present a novel scan-stratified case-control sampling technique that accounts for the rarity of enhancement and reduces the computational cost.

e-mail: gina.pomann@gmail.com

### Persistence Landscape of Functional Signal and its Application to Epileptic Electroencephalogram Data

**Yuan Wang\***, University of Wisconsin, Madison  
**Hernando Ombao**, University of California, Irvine  
**Moo K. Chung**, University of Wisconsin, Madison

Persistent homology is a recently popular multi-scale topological data analysis framework that has many potential scientific applications, particularly in neuroscience. The method can be effectively applied to yield patterns in nonlinear imaging data that are otherwise undetected by existing mono-scale techniques. Among several persistent homological features, recently proposed persistence landscape is used as a new signal detection method in one-dimensional functional data. For this purpose, weighted Fourier series expansion is used for estimating the functional shape of the data before the persistent landscape is obtained. We utilize the proposed method to study topological differences between electroencephalogram (EEG) data during pre-seizure and seizure periods in a patient diagnosed with left temporal epilepsy.

e-mail: yuanw@stat.wisc.edu

## 90. NEW METHODS IN GENOMICS

### Inference of Epigenetic Modulation of Gene Expression with Meta-Pathway Analysis

**Elana J. Fertig\***, Johns Hopkins University  
**Ana Markovic**, University of California, San Francisco  
**Ludmila V. Danilova**, Johns Hopkins University  
**Daria A. Gaykalova**, Johns Hopkins University  
**Leslie Cope**, Johns Hopkins University  
**Christine H. Chung**, Johns Hopkins University  
**Joseph A. Califano**, Johns Hopkins University  
**Michael F. Ochs**, The College of New Jersey

Global DNA methylation changes occur heterogeneously in distinct cancer subtypes. However, these changes have not been linked to expression or functional changes specific to these subtypes. We create a new matrix factorization algorithm to integrate global gene expression and DNA methylation measurements. This algorithm incorporated novel models to transform methylation and expression measurements to a common scale, and thus inferred meta-pathways that link gene expression changes to DNA methylation. We applied this algorithm to head and neck squamous cell carcinoma (HNSCC), and where it differentiated known clinical subtypes. For example, this algorithm uniquely linked DNA methylation changes and reactivation of genes in the Hedgehog pathway in the worse-prognosis HPV-negative HNSCC. We confirmed

that GLI1, the primary Hedgehog target, showed higher expression in tumors compared to normal samples with HPV-tumors having the highest GLI1 expression, suggesting that increased expression of GLI1 is a potential driver in HPV-HNSCC. Our algorithm for integration of DNA methylation and gene expression can infer biologically significant molecular pathways that may be exploited as therapeutics targets. Similar integrative analysis of high-throughput coupled DNA methylation and expression datasets may yield novel insights into epigenetically regulated pathways in other diseases and developmental processes.

e-mail: ejfertig@jhmi.edu

### **Integrative Modeling of Multiplatform Genomic Data**

**Yen-Tsung Huang\***, Brown University

Given the availability of genomic data, there have been emerging interests in integrating multi-platform data. Here we propose to model multiplatform genomic data as a biological process to delineate phenotypic traits under the framework of causal mediation modeling. We propose a regression model for the joint effect of multiple genomic data and their non-linear interactions on the outcome, and study three path-specific effects. We characterize correspondences between the three path-specific effects and coefficients in the regression model, which are influenced by causal relations across genomic data. A score test for variance components of regression coefficients is developed to assess path-specific effects. The test statistic under the null follows a mixture of chi-square distributions, which can be approximated using a characteristic function inversion method or a perturbation procedure. We construct tests for different candidate models determined by various combinations of multiple genomic data and their interactions, and further propose an omnibus test to accommodate different models. The utility of the method will be illustrated in numerical simulation studies and a glioblastoma data from The Cancer Genome Atlas (TCGA).

e-mail: Yen-Tsung\_Huang@brown.edu

### **The Most Informative Spacing Test as an Outlier and Subgroup Identification Method**

**Iwona Pawlikowska\***, St. Jude Children's Research Hospital

**Gang Wu**, St. Jude Children's Research Hospital

**Michael Edmonson**, St. Jude Children's Research Hospital

**Tanja Gruber**, St. Jude Children's Research Hospital

**Jinghui Zhang**, St. Jude Children's Research Hospital

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

Several outlier and subgroup identification statistics (OASIS) have been proposed to discover transcriptomic features with outliers or multiple modes in expression that are indicative of distinct biological processes. Here, we borrow ideas from the OASIS methods in the bioinformatics and statistics literatures to develop the most informative spacing test (MIST) for unsupervised detection of such transcriptomic features. For each individual expression variable, MIST computes the differences between consecutive order statistics (spacings) and multiplies each spacing by the geometric mean of the sizes of the two groups it defines. The spacing with the

largest value of this statistic is considered to be the most informative spacing and its significance is determined by simulation. The performance of MIST in simulation studies and example applications is similar to or superior to that of other OASIS methods in the literature. In a pediatric leukemia study, MIST more effectively identified features that divide patients according to gender or the presence of a prognostic fusion-gene in both RNA-seq and microarray expression data than any other OASIS method. MIST may be generalized to identify transcriptomic features that divide subjects into more than two groups and to select features for class discovery analysis.

e-mail: iwona.pawlikowska@stjude.org

### **Cross-Platform Gene Expression Profile Classification Using Top-Scoring Pairs**

**Prasad Patil\***, Johns Hopkins School of Public Health

**Benjamin Haibe-Kains**, Institut de Recherches

Clinques de Montreal

**Jeffrey T. Leek**, Johns Hopkins School of Public Health

Top-Scoring Pairs (TSPs) are lightweight and invariant features for model-building using gene expression data. A TSP is a pair of genes whose relative ranking flips between disease subtypes. Since pairs are rank-based, they are robust to monotone transformations of the gene expression data. It has been suggested that this property makes classifiers based on TSPs more robust to changes in the underlying measurement technology. Here we develop a procedure for performing multi-class classification with TSPs including when some pairs are not observed on a particular platform. Using 26 curated breast cancer microarray datasets spanning 11 different platforms, we show that models based on up to five TSPs are as accurate as the existing standard for predicting breast cancer statuses and subtypes (of which many are predicated upon 50 genes). Moreover, in many cases these models retain their accuracy and do not require retraining across platforms. These results suggest that an inexpensive and interpretable model could be universally instituted as a platform-independent and stable standard for phenotype prediction with gene expression data.

e-mail: prpatil@jhsphe.edu

### **A Survival Copula Mixture Model for Comparing Two Genomic Rank List**

**Yingying Wei\***, Johns Hopkins University

**Hongkai Ji**, Johns Hopkins University

Analyzing high-throughput data often leads to rank lists. How to characterize the commonalities and differences between two genomic rank lists is a common problem with a variety of applications. One example is measuring the reproducibility between two rank lists derived from experimental replicates. Another example is detecting co-binding regions for two types of transcription factors from ChIP-seq peak lists. A simple Venn diagram shows the overlap between two lists but fails to account for the rank ordering. On the other hand, the irreproducibility discovery rate (IDR) recently proposed by Li et al (2011) measures the consistency of ranks only among the overlapping set in the

two lists, which could mistakenly claim high reproducibility for two lists with very little overlap but high concordance of ranks in the common set. Moreover, when one of the two replicates is of poor quality, IDR would reject many good features being signals. In this work, we propose a new solution to comparing two genomic rank lists by translating the problem into a bivariate survival problem. We generalize the idea of IDR and employ a survival copula mixture model. The effectiveness of our approach is illustrated by simulations and real datasets from the ENCODE project.

e-mail: ywei@jhspsh.edu

### **An Integrated Method for Detecting MicroRNA Target Proteins Through Reverse-Phase Protein Lysate Arrays**

**Jiawen Zhu\***, Stony Brook University

**Song Wu**, Stony Brook University

**Jie Yang**, Stony Brook University

Understanding functions of microRNAs (or miRNAs), particularly their effects on protein degradation, is biologically important. Emerging technologies, including the reverse-phase protein lysate array (RPPA) for quantifying protein concentration and RNAseq for quantifying miRNA expression, provide a unique opportunity to study miRNA-protein regulatory mechanisms. A naïve and commonly used way to analyze such data is to directly examine the correlation between the raw miRNA measurements and protein concentrations estimated from RPPA through simple linear regressions. However, the uncertainty associated with protein concentration estimates is ignored, which may lead to less accurate results and significant power loss. We propose an integrated nonlinear hierarchical model for detecting miRNA targets through original RPPA intensity data. This model is fitted within a maximum likelihood framework and the correlation between miRNA and protein is assessed using Wald tests. Our simulation studies demonstrated that the integrated method performed consistently better than the simple method, especially with limited sample sizes. The model was also illustrated through a TCGA real dataset.

e-mail: jie.yang@stonybrook.edu

### **Testing in Metagenomic Profiling Studies with the Microbiota Regression-Based Kernel Association Test (MiRKAT)**

**Ni Zhao\***, Fred Hutchinson Cancer Research Center

**Michael C. Wu**, Fred Hutchinson Cancer Research Center

The massively parallel sequencing technology has enabled high-throughput profiling of microbiota sampled directly from the human body. Evaluating the association between microbiota composition and phenotypic outcome is of considerable interest. Current strategies for association testing include the distance based approach, which suffers from two major challenges. First, adjusting for additional covariates is not straightforward under its permutation framework. Secondly, given the existence of many distance measures, it is unclear how to choose the best distance

metric to obtain optimal power. Therefore, we propose the microbiome regression-based kernel test (MiRKAT). MiRKAT uses the existing kernel machine framework to compare pairwise distance/similarity in the outcome to pairwise distance/similarity in the microbiome profiles while adjusting for covariates. However, modifications to the standard kernel machine framework are required to evaluate significance. Additionally, we developed an optimal test which simultaneously examines multiple distance metrics, selects the best distance metric to compute a p-value, and adjusts for having taken the optimal distance metric. Simulations and real studies show that MiRKAT can provide higher power with improved control of type I error compared to existing distance based and kernel approaches.

e-mail: nzhao@fhcrc.org

## **91. IMS MEDALLION LECTURE**

### **Statistical Genetics and Genomics in the Big Data Era: Opportunities and Challenges in Research and Training**

**Xihong Lin, Ph.D.**, Harvard School of Public Health

The human genome project in conjunction with the rapid advance of high throughput technology has transformed the landscape of health science research. The genetic and genomic era provides an unprecedented promise of understanding genetic underpinnings of complex diseases or traits, studying gene-environment interactions, predicting disease risk, and improving prevention and intervention, and advancing personalized medicine. A large number of genome-wide association studies conducted in the last ten years have identified over 1,000 common genetic variants that are associated with many complex diseases and traits. Massive next generation sequencing data as well as different types of omics data have become rapidly available in the last few years. These big genetic and genomic data present statisticians with many exciting opportunities as well as challenges in data analysis and in interpretation of results. They also call for more interdisciplinary knowledge and research, e.g., in statistics, machine learning, data curation, molecular biology, genetic epidemiology and clinical science. In this talk, I will discuss some of these challenges, such as low-level pre-processing, analysis of rare variants in next generation sequencing association studies; integrative genomics, which integrates different types of omics data; and study of gene-environment and genetreatment interactions. I will also discuss strategies of training next generation quantitative genomic scientists at the interface of statistical genetics and genomics, computational biology and genetic epidemiology, to meet these challenges.

e-mail: xlin@hsph.harvard.edu

## 92. PARAMETRIC OR NONPARAMETRIC; WHICH IS THE ANSWER?

### Super Learning to Hedge Against Incorrect Inference from Arbitrary Parametric Assumptions in Marginal Structural Modeling

Romain Neugebauer\*, Kaiser Permanente

We present analyses of electronic health records data to evaluate the effect of various pharmacotherapy strategies for patients with type 2 diabetes based on Marginal Structural Modeling. The sensitivity of results to modeling assumptions motivates the application of Super Learning in Comparative Effectiveness Research. In particular, results contradict the common intuition that more data-adaptive estimators of the treatment and censoring mechanisms to implement inverse probability weighting estimation leads to unstable weights and thus large increase in estimation variability.

e-mail: romain.s.neugebauer@kp.org

### Fitting ICU Data Complexity: Need for Innovative Prediction Tools Mortality Prediction by Superlearner

Romain Pirracchio\*, Hôpital Saint Louis, Paris, France  
Maya Petersen, University of California, Berkeley  
Sylvie Chevret, Hôpital Saint Louis, Paris, France  
Mark van der Laan, University of California, Berkeley

BACKGROUND: Predicting the outcome of patients hospitalized in Intensive Care Units (ICU) is crucial. The Super Learner (SL) is a machine learning method that can incorporate a large customized library of different data-fitting algorithms. OBJECTIVES: To assess performances of SL based prediction of ICU mortality as compared to the predictions obtained with the SAPS2 and the first SOFA. METHODS: We used the MIMIC-II database that includes all patients admitted to an ICU at Boston's Beth Israel Deaconess Medical Center from 2001 to 2012. The prediction of hospital mortality based on the SAPS2, the SOFA score and the SL were compared. RESULTS: 24,508 patients were included: age: 65[51-77], SAPS2: 38[27-51], SOFA: 5[2-8], medical: 2453(10%), trauma: 9006(37%). 3002(12.2%) patients died in hospital. The SL predictor outperformed each single candidate algorithm included in its library. As compared to the prediction performances obtained with the SAPS2 (0.71[0.71-0.72]) or with the SOFA (0.78[0.77-0.78]), the one obtained with the SL were far better with an AUROC of 0.89 [0.88-0.89] ( $p < 0.0001$  for AUROC comparisons) (Figure 1). CONCLUSIONS: As compared to usual severity scores, the Super Learner seems to offer a great benefit in predicting hospital mortality from ICU patients.

e-mail: romainpirracchio@yahoo.fr

### Sensitivity Analysis for Causal Inference Under Unmeasured Confounding and Measurement Error Problems

Iván Díaz\*, Johns Hopkins Bloomberg School of Public Health

Mark van der Laan, University of California, Berkeley

We present a sensitivity analysis for drawing inferences about parameters that are not estimable from observed data without additional assumptions. We present the methodology using two different examples: a causal parameter that is not identifiable due to violations of the randomization assumption, and a parameter that is not estimable in the nonparametric model due to measurement error. Existing methods for tackling these problems assume a parametric model for the type of violation to the identifiability assumption, and require the development of new estimators and inference for every new model. The method we present can be used in conjunction with any existing asymptotically linear estimator of an observed data parameter that approximates the unidentifiable full data parameter, and does not require the study of additional models.

e-mail: idiaz@jhu.edu

### From Causal Roadmaps to Hedging Your Bets in the Adventures of Comparative Effectiveness Research: An Illustration Using an Effect Modification Analysis of STAR\*D

Wenjing Zheng, University of California, Berkeley

Zhehui Luo\*, Michigan State University

Mark van der Laan, University of California, Berkeley

The causal nature of the research questions in Comparative Effectiveness Research underscores the need to separate limitations of research design from analytic constraints. The complexity of the data structure underscores the need to avoid strong parametric modeling constraints in estimation and inference, as well as sensitivity analysis. We illustrate these two aspects in the context of the Sequential Trial of Alternatives to Relieve Depression (STAR\*D), a multi-level longitudinal clinical study of treatment strategies for major depression. We estimate the causal effect of switching medication vs augmenting medication on symptom remission in the current level, by variables (effect modifiers) collected prior to the start of the level. Under a nonparametric causal framework (Pearl 2009), we establish the statistical estimands of interest that account for confounding of the treatment effect and potential selection bias introduced by dropout and missing effect modifier. We then illustrate the use of super learning and targeted maximum likelihood estimator in estimating and inference, and the application of a nonparametric sensitivity analysis.

e-mail: wzheng@stat.berkeley.edu

## 93. CAUSAL INFERENCE IN HIGH DIMENSIONAL SETTINGS

### Calibrated Observational Studies

**David Madigan\***, Columbia University

Observational healthcare data, such as administrative claims and electronic health records, play an increasingly prominent role in healthcare. Pharmacoepidemiologic studies in particular routinely estimate temporal associations between medical product exposure and subsequent health outcomes of interest, and such studies influence prescribing patterns and healthcare policy more generally. Some authors have questioned the reliability and accuracy of such studies, but few previous efforts have attempted to measure their performance. The Observational Medical Outcomes Partnership (OMOP, <http://omop.org> [omop.fnih.org]) has conducted a series of experiments to empirically measure the performance of various observational study designs with regard to predictive accuracy for discriminating between true drug effects and negative controls. In this talk I will describe a procedure we have developed for “calibrating” statistical outputs of observational studies so that they have the desired nominal properties.

e-mail: madigan@yahoo.com

### Connectivity and Causality in Brain Imaging

**Martin A. Lindquist\***, Johns Hopkins  
Bloomberg School of Public Health

To date human brain mapping has primarily been used to construct maps indicating regions of the brain that are activated by certain tasks. Recently, there has been an increased interest in augmenting this type of analysis with connectivity studies that seek to describe how brain regions interact and how these interactions depend on experimental conditions and behavioral measures. Often researchers discriminate between functional connectivity, the undirected association between two or more fMRI time series, and effective connectivity, the directed influence of one brain region on the physiological activity recorded in other brain regions. In this talk we argue that this distinction is not entirely clear or relevant. Instead, the validity of the conclusions made from any connectivity method will depend strongly on certain key assumptions that are often poorly specified and difficult to check. We illustrate how ideas from causal inference can provide a mathematical framework for determining these assumptions.

e-mail: mlindqui@jhsph.edu

### Causal Inference for fMRI Time Series Data with Systematic Errors of Measurement in a Balanced On/Off Study of Social Evaluative Threat

**Michael E. Sobel\***, Columbia University  
**Martin A. Lindquist**, Johns Hopkins  
Bloomberg School of Public Health

Functional magnetic resonance imaging (fMRI) has facilitated major advances in understanding human brain function. Neuroscientists are interested in using fMRI to study the effects of external stimuli on brain activity and causal relationships among brain regions, but have not stated what is meant by causation or defined the effects they purport to estimate. We construct a frame work for causal inference using blood oxygenation level dependent (BOLD) fMRI time series data. In the usual literature on causal inference, potential outcomes, assumed to be measured without systematic error, are used to define unit and average causal effects. However, in general the potential BOLD responses are measured with stimulus dependent systematic error. Thus we define unit and average causal effects that are free of systematic error, using a linear mixed model to estimate these effects. In contrast to the usual case of a randomized experiment where adjustment for intermediate outcomes leads to biased estimates of treatment effects (?), here the failure to adjust for systematic error leads to biased estimates. These results are important for neuroscientists, who typically do not adjust for systematic error. They should also prove useful to researchers in other areas where responses are measured with error and in fields where large amounts of data are collected on relatively few subjects. To illustrate our approach, we re-analyze data from a social evaluative threat task, comparing the findings with results that ignore systematic error.

e-mail: mes105@columbia.edu

### Data Adaptive Target Parameters in Causal Inference

**Alan E. Hubbard\***, University of California, Berkeley  
**Mark van der Laan**, University of California, Berkeley

This talk concerns inference for a parameter of interested (targeted parameter), which is not specified a priori, but is a random function of the data. Consider  $n$  i.i.d. copies of a random variable with typically unknown data-generating distribution. To define the statistical target, partition the sample in  $V$  sub-samples, and use this partitioning to define validation samples and corresponding training samples. We consider algorithms on training samples that produce parameters of interest and define our cross-validated statistical target parameter as the average of these training sample specific target parameters across the validation samples. We present estimators and asymptotics of adaptive parameters, providing opportunities for statistical learning from data that avoid the typical requirement that a parameter is defined a priori. The focus is on high-dimensional data problems where one wishes to explore parameters for which the data has the most information without paying too high a cost for this exploration, but still

derive consistent inference. Examples include estimation of parameters inspired by causal inference, such as average treatment effects. The general approach holds great promise for using algorithms to concentrate only on those parameters for which the data has sufficient information, while still reporting consistent inference.

e-mail: hubbard@berkeley.edu

## 94. ADVANCES IN TIME SERIES ANALYSIS OF BIOMEDICAL SIGNALS

### Spatial Identification of Epileptic Brain Regions

**Giovanni Motta\***, Columbia University

**Michael M. Haglund**, Duke University

**Daryl Hochman**, Duke University

The surgical outcomes of patients suffering from neocortical epilepsy are not always successful. It is known that the optical spectroscopic properties of brain tissue are correlated with changes in neuronal activity. The method of mapping these activity-evoked optical changes is known as imaging of intrinsic optical signals (ImIOS). Activity-evoked optical changes measured in neocortex are generated by changes in cerebral hemodynamics. ImIOS has the potential to be useful for both clinical and experimental investigations of the human neocortex. However, its usefulness for human studies is currently limited because intra-operatively acquired ImIOS data is noisy. In this paper we introduce a novel flexible tool, based on spatial statistical representation of ImIOS, that allows for source localization of the epilepsy regions. In particular, our model incorporates spatial correlation between the location of the epileptic region(s) and the neighboring regions, non-stationarity of the observed time series, and heartbeat/respiration cyclical components. To demonstrate how our method might be used for intra-operative neuro-surgical mapping, we provide an application of the technique to optical data acquired from a single human subject during direct electrical stimulation of the cortex.

e-mail: g.motta@stat.columbia.edu

### Time Series Analysis of Molecular Motor-Cargo Complexes

**John Fricks\***, The Pennsylvania State University

Linear molecular motors, such as kinesin and dynein, carry cargos along microtubules providing active transport within the cell. Kinesin and dynein tend to move in opposite directions along microtubules and yet may be connected to the same cargo. This gives rise to a number of questions: Are these opposing motors connected at the same time? Do these motors somehow cooperate or simply work against one another? Are there regulatory ramifications for having both types of motors? To study this system, we look at time traces of fluorescently labeled individual cargos at the time scale of tens of milliseconds to infer the underlying mechanism of transport with multiple motors.

e-mail: fricks@stat.psu.edu

### Penalized Multivariate Whittle Likelihood for Power Spectrum Estimation

**Robert T. Krafty\***, Temple University

**William O. Collinge**, University of Pittsburgh

Nonparametric estimation procedures that can flexibly account for varying levels of smoothness among different functional parameters, such as penalized likelihoods, have been developed in a variety of settings. However, geometric constraints on power spectra have limited the development of such methods when estimating the power spectrum of a vector-valued time series. I discuss a penalized likelihood approach to nonparametric multivariate spectral analysis through the minimization of a penalized Whittle negative log-likelihood. This likelihood is derived from the large-sample distribution of the periodogram and includes a penalty function that forms a measure of regularity on multivariate power spectra. The approach allows for varying levels of smoothness among spectral components while accounting for the positive definiteness of spectral matrices and the Hermitian and periodic structures of power spectra as functions of frequency. The method is illustrated through the spectral analysis of heart rate variability during different periods of sleep.

e-mail: krafty@temple.edu

### A Bayesian Model of Activation and Functional Connectivity for Event-Related fMRI

**Wesley K. Thompson\***, University of California, San Diego

Neuroscientists are increasingly focused on the determination of functional relationships among anatomically-distinct brain regions, termed functional connectivity (FC). FC has been evaluated using data obtained from fMRI studies. We propose a novel multivariate methodology for event-related fMRI that simultaneously smooths the BOLD responses and determines FC among several regions. Smoothing is accomplished via empirical bases obtained from functional principal components analysis. The coefficients of the basis are allowed to be correlated across regions, and the nature and strength of FC is derived from this correlation matrix. The model is implemented within a Bayesian framework by using a Markov Chain Monte Carlo (MCMC) sampling algorithm. We demonstrate our methodology on a sample of clinically depressed subjects and healthy controls in examining relationships among three brain regions implicated in depression and emotion during emotional information processing.

e-mail: wktwktwkt@gmail.com

## 95. FRONTIERS IN STATISTICAL GENETICS FOR COMPLEX TRAIT ASSOCIATION

### Genetic Architecture of Complex Traits: Implications for Discovery, Prediction and Prevention

**Nilanjan Chatterjee\***, National Cancer Institute, National Institutes of Health  
**JuHyun Park**, Dongguk University, South Korea

Large genome-wide association studies are now consistently pointing towards an extremely polygenic model for complex diseases. Such models may involve thousands of susceptibility markers, each conferring only a modest risk, but collectively they could be explaining substantial variation in disease-risks in populations. Further, a few large studies of gene-environment interactions indicate that genetic and environmental risk-factors may broadly act in a multiplicative fashion on the risk of a number of different cancers and possibly other diseases. In this talk, I will explore how under such emerging models for disease architecture, the performance of polygenic risk prediction models are expected to improve in the future with increasing sample size, incorporation of functional information and integration of next generation sequencing or genotyping technologies that can provide coverage for low frequency and rare variants. Further, using results from recent studies on bladder and breast cancers, I will illustrate potential implications for multiplicative gene-environment interactions for targeted prevention for these two malignancies both of which have modifiable risk-factors. These analyses will highlight both challenges and opportunities for using genetic information for personalized disease prevention.

e-mail: chattern@mail.nih.gov

### Statistical Approaches for Rare-Variant Association Testing in Families

**Michael P. Epstein\***, Emory University

With the emergence of next-generation sequencing studies and exome-chip technology, a tremendous number of rare-variant association tests have been developed to elucidate the genetic mechanisms of complex traits. However, the overwhelming majority of rare-variant tests assume either a case-control or population-based study design, with little development of rare-variant tests for family-based studies. Such development is important, since family-based studies (which were often avoided in the GWAS era) have gained increased relevance in the resequencing era as they can be used to study co-segregation patterns of causal variants and provide an opportunity to form test statistics that are robust to population stratification. In this talk, we describe novel statistical methods for family-based association testing of rare variants. We describe methods both for the analysis of complex diseases in affected sibships as well as for analysis

of disease-related quantitative traits. We will show that these methods are robust to population stratification and further describe screening procedures to improve power of some tests. We will illustrate our methods using a combination of simulated and real sequencing data. This is joint work with Drs. Karen Conneely and Glen Satten.

e-mail: mpepste@emory.edu

### A Novel Collapsing Method for Rare Copy Number Variants

**Jung-Ying Tzeng\***, North Carolina State University  
**Jin P. Szatkiewicz**, University of North Carolina, Chapel Hill  
**Patrick F. Sullivan**, University of North Carolina, Chapel Hill

CNVs play an important role in the etiology of multiple psychiatric disorders. Due to modest marginal effect size or rarity of the CNV, collapsing approaches could be important to study how CNVs impact risk for psychiatric disorders. In contrast to sequence variants, CNVs vary in size, type, dosage, and sequence level details of gene disruption. Because of its multi-faceted nature, CNV analysis is more challenging than SNP analysis and the most important consideration is the heterogeneous effects from a mixture of neutral, risk, and protective variants. Existing burden tests do not fully explore CNV-specific issues; their performance tends to be suboptimal due to ignoring heterogeneity and testing only one event at a time (e.g., duplication/deletion/both). We introduce a new collapsing method for CNVs that is robust to multiple types of heterogeneity. Our method is based on a similarity collapsing approach to collectively examine the effects of multiple CNV features (e.g. size, type, dosage, mixture effects) and weight CNVs by their frequencies and details of gene disruption. Multiple confounders can be simultaneously corrected. We demonstrate the robustness, validity and utility of the proposed approaches using real data applications and simulations.

e-mail: jytzeng@stat.ncsu.edu

### Testing Association without Calling Genotypes Allows for Systematic Differences in Read Depth and Sequencing Error Rate Between Cases and Controls

**Glen A. Satten\***, Centers for Disease Control and Prevention  
**Richard Johnston**, Emory University  
**Peizhou Liao**, Emory University  
**Yu Jiang**, Duke University  
**Andrew S. Allen**, Duke University  
**Yijuan Hu**, Emory University

The quality of genotype calling for next-generation sequence data depends on read depth. Loci with high coverage can typically be called reliably, while those with low coverage may be difficult to call. In an association study, if case participants are sequenced to a greater depth than controls, the difference in genotype quality can introduce a systematic bias. This can easily occur when historical controls are used. We propose directly comparing the proportion

of calls for the minor allele between cases and controls, rather than comparing genotypes. We show how this proposal can be used to perform both single-marker test and gene-level test of rare variants. We also show how this proposal allows the per-call read error rate to differ between cases and controls. Finally, we develop methods that allow for valid testing when screening out loci estimated to be monomorphic. Using simulated data, we demonstrate our proposals yield valid tests even in the presence of systematic differences in coverage rate between cases and controls, and show that in these situations, tests based on genotype have inflated size. We also show that power gains are possible using designs where we increase the number of controls while decreasing the read depth (while keeping total reads constant).

e-mail: gas0@cdc.gov

## 96. FUNCTIONAL DATA APPROACHES TO NEUROLOGICAL AND MENTAL DISEASE

### **Distance Splines, Nonparametric Functional Regression, and Multimodal Neuroimaging**

**Philip T. Reiss\***, New York University and Nathan Kline Institute

**Lei Huang**, Johns Hopkins University

**Huaihou Chen**, New York University

**David L. Miller**, University of St Andrews

The 'distance splines' technique, recently proposed for geographical applications with irregular boundaries, entails fitting a multidimensional spline smoother on a principal coordinate space. The same approach can be readily extended to general data objects with an associated distance measure. In particular, for functional data, this spline approach emerges as an alternative to the standard functional Nadaraya-Watson kernel estimator for nonparametric functional regression. Distance splines offer a flexible way to regress scalar responses on multiple functional or more general predictor objects, each defined on a metric space, without requiring these to be registered to a common domain. We illustrate the methodology using a multimodal imaging data set that includes both structural and functional brain images obtained from a community sample.

e-mail: phil.reiss@nyumc.org

### **Assessing Systematic Effects of Stroke on Motor Control Using Hierarchical Function-on-Scalar Regression**

**Jeff Goldsmith\***, Columbia University

**Tomoko Kitago**, Columbia University

This work is concerned with understanding common population-level effects of stroke on motor control while accounting for possible subject-level idiosyncratic effects. Upper extremity motor control for each subject is assessed through repeated planar reaching motions from a central point to eight pre-specified targets arranged on a circle. We observe the kinematic data for hand position as a bivariate

function of time for each reach. Our goal is to estimate the bivariate function-on-scalar regression with subject-level random functional effects while accounting for potential correlation in residual curves; covariates of interest are severity of motor impairment and target number. We express fixed effects and random effects using penalized splines, and allow for residual correlation using a Wishart prior distribution. Parameters are jointly estimated in a Bayesian framework, and we implement a computationally efficient approximation algorithm using variational Bayes. Simulations indicate that the proposed method yields accurate estimation and inference, and application results suggest that the effect of stroke on motor control has a systematic component observed across subjects.

e-mail: jeff.goldsmith@columbia.edu

### **Flexible Concurrent Regression Models for Functional Data**

**Janet Kim**, North Carolina State University

**Ana-Maria Staicu\***, North Carolina State University

**Arnab Maity**, North Carolina State University

We propose a generalization of the functional concurrent model, where both the response and the covariate are functional data observed on the same time domain. In contrast to typical functional concurrent models, we allow the relationship between the response and covariate to be nonlinear and vary with both the current time point and the current covariate. In this framework we develop methodology for estimation and inference of the unknown relationship, by allowing for correlated error structure as well as sparse and/or irregular sampling design. Formal hypothesis testing to investigate whether the relationship is non-linear is discussed. The proposed method is illustrated through simulation study and application to real world data.

e-mail: ana-maria\_staicu@ncsu.edu

### **Biosignatures Based on Imaging Data**

**Todd Ogden\***, Columbia University

**Adam Ciarleglio**, New York University

**Eva Petkova**, New York University

**Thaddeus Tarpey**, Wright State University

In many biomedical applications it is of interest to use imaging data or other very high dimensional data to predict some clinical outcome, to classify subjects, or to describe some patient attributes. Obtaining meaningful results in such a situation requires some form of dimension reduction while taking into account the structure of the data, a primary goal of functional data analysis in general. This talk will discuss various functional data analytic approaches for identifying relevant patient characteristics based on functional data observed at baseline that can help in classifying patients or predicting response.

e-mail: to166@columbia.edu

## 97. MODELING NEUROLOGICAL DISEASES WITH IMAGING DATA

### Developmental Disorders and Neuroimaging: Tools, Results and Issues

**Brian S. Caffo\***, Johns Hopkins  
Bloomberg School of Public Health

In this talk we discuss the use of neuroimaging and statistical tools to investigate developmental disorders focusing on attention deficit hyperactive disorder and autism. Focus will lie on resting state functional magnetic resonance imaging in the primary motor cortex. Statistical tools will be demonstrated to investigate, lateralization, brain connectivity and network shape. Tools will be evaluated on data collected from an individual lab and large public datasets.

e-mail: bcaffo@gmail.com

### Learning Brain Connectivity Network of Depression Via Multi-Attribute Canonical Correlation Graphs

**Jian Kang\***, Emory University  
Han Liu, Princeton University  
**DuBois F. Bowman**, Columbia University  
**Helen S. Mayberg**, Emory University

The rapid advancement of neuroimaging techniques provides a great opportunity for the study of major depressive disorder (MDD). Recent resting-state functional connectivity magnetic resonance imaging (fMRI) studies have shown that MDD is closely related to alteration in the functional brain network, i.e. significant differences in several regions and networks between MDD patients and health controls. In the paper, we propose a multi-attribute canonical correlation graph model to learn brain connectivity network of depression from resting-state fMRI data. Our method has ability to identify the brain connectivity network and strength for a whole brain analysis with a large number of brain regions. We apply the proposed method to a resting-state fMRI data of MDD patients and health controls. We identify several regions (e.g. Amygdala and Pallidum) where the MDD patients significantly reduce the functional connectivity to other regions.

e-mail: jian.kang@emory.edu

### Normalization Techniques for Statistical Inference from Magnetic Resonance Imaging

**Russell T. Shinohara\***, University of Pennsylvania  
**Elizabeth M. Sweeney**, Johns Hopkins University  
**Jeff Goldsmith**, Columbia University  
**Navid Shiee**, Henry M. Jackson Foundation  
**Farrah J. Mateen**, Harvard University  
**Peter A. Calabresi**, Johns Hopkins University  
**Samson Jarso**, Johns Hopkins University  
**Dzung L. Pham**, Henry M. Jackson Foundation  
**Daniel S. Reich**, National Institute of Neurological Disorders and Stroke, National Institutes of Health  
**Ciprian M. Crainiceanu**, Johns Hopkins University

While computed tomography and other imaging techniques are measured in absolute units with physical meaning, magnetic resonance images are expressed in arbitrary units that are difficult to interpret and differ between study visits and subjects. Much work in the image processing literature on intensity normalization has focused on histogram matching and other histogram mapping techniques, but with little emphasis on normalizing images to have biologically interpretable units. Furthermore, there are no formalized principles or goals for the crucial comparability of image intensities within and across subjects. To address this, we propose a set of criteria necessary for the normalization of images. We further propose simple and robust biologically motivated normalization techniques for multisequence brain imaging that have the same interpretation across acquisitions and satisfy the proposed criteria. We compare the performance of different normalization methods in thousands of images of patients with Alzheimer's Disease, hundreds of patients with multiple sclerosis, and hundreds of healthy subjects.

e-mail: rshi@upenn.edu

### Voxel-Wise Marginal Longitudinal Modelling of Brain Atrophy Data

**Bryan Guillaume**, University of Warwick and Université de Liège  
**Thomas E. Nichols\***, University of Warwick  
**Lourens Waldorp**, University of Amsterdam

The study of Alzheimer's Disease critically depends on the longitudinal modelling of local gray matter volume, to track how the disease progresses as reflected in this important marker. While longitudinal modelling with linear mixed effects models is a work-horse of biostatistics, these models have found little uptake in neuroimaging. With the exception of recent work by the Freesurfer group (which pools evidence over regions), all of the major brain imaging software packages have severe deficits in their longitudinal modelling tools: SPM models repeated measures covariance flexibly, but only once for the entire brain; FSL has no facility to model repeated measure covariance. We propose the use of another other work-horse biostatistical method, the GEE and the sandwich estimator. For Gaussian data and an identity working covariance matrix, this correspond to Ordinary Least Squares with a marginal model to get point estimates. Like others, we find the standard (large sample) results are quite poor for the small samples common in

neuroimaging, and evaluate a range of methods to improve finite sample performance, developing (to our knowledge) new estimate of the test statistic degrees of freedom when covariance is pooled over subjects within groups. We identify a method with good size and power characteristics and demonstrate its use with the ADNI dataset.

e-mail: t.e.nichols@warwick.ac.uk

## 98. MAKING SENSE OF SENSORS: STATISTICAL METHODS FOR WEARABLE COMPUTING

### **Activis: An R Package for Visualizing Functional Actigraphy Data**

**Abbass Sharif\***, University of Southern California  
**Juergen Symanzik**, Utah State University

Actigraphy is an emerging technology for measuring human activity/rest levels over time. An actigraph unit, a non-invasive watch-like device, collects actigraphy data almost continuously over time. So far, a few and limited visualization techniques have been provided by the manufacturers of actigraphs. In order to help better visualize such data, we developed an object-oriented R package that consists of a set of utility classes and methods for managing, storing, importing, and exporting actigraphy data and results. It implements univariate and multivariate explanatory data analysis (EDA) techniques to reduce noise and irregularities in visualizing large functional data sets, without any smoothing methods, in order to reveal interesting patterns and trends in these data. In this talk, we discuss the Activis package structure and object oriented design, and then present a case study to show how it could be utilized to visually analyze functional data in general and actigraphy data in particular.

e-mail: asharif@usc.edu

### **From Humans to Monkeys and Back: Physical Activity Patterns in Humans and Primates**

**Vadim Zipunnikov\***, Johns Hopkins University  
**Jeff Goldsmith**, Columbia University  
**Haochang Shou**, Johns Hopkins University  
**Ciprian Crainiceanu**, Johns Hopkins University

I will illustrate key statistical challenges of analyzing physical activity data by giving snapshots of three recent projects. First, I will talk about analysis of data collected on 700+ subjects wearing an Actiheart device that collects minute-by-minute activity counts and heart rate for one week as a part of the Baltimore Longitudinal Study of Aging. Secondly, I will talk about an experiment examining changes in activity in a monkey model of Parkinson's disease that can be used to guide studies of activity changes in human Parkinson's disease, as well as other related disorders associated with dopamine depletion. Finally, I will talk about Home and Away models for physical activity and inactivity in elderly adults. These models use both minute-by-minute activity and GPS data to identify activity hotspots, outside locations where alleviated physical activity is recorded.

e-mail: vzipunni@jhspsh.edu

### **Measurement Error Models for Physical Activity: Accelerometers and Self Report**

**John W. Staudenmayer\***, University of Massachusetts, Amherst

This talk will evaluate the amount and nature of measurement error and misclassification in several methods that are commonly used to assess physical activity / sedentary behavior. Our evaluations will be based on a recent validation study that simultaneously measured various aspects of physical activity / sedentary behavior using both monitor-based and interview-based methods. The study recruited 213 healthy adults and children, and those participants wore two physical activity measurement monitors (an Actigraph and an ActivPal) concurrently for seven days. The participants were also interviewed three times to complete three structured previous day physical activity recall surveys. This survey instrument is a modification of an established instrument, the 7-Day Physical Activity Recall. Both the monitors and the surveys produce estimates of several aspects of physical activity / sedentary behavior. This talk will consider and compare both parametric (mixed model) and non-parametric (deconvolution) measurement error models for these assessments. Misclassification models also will be used to assess aspects of physical activity / sedentary behavior that are summarized with discrete measurements.

e-mail: jstauden@math.umass.edu

### **Statistical Methods for Development and Temporal Organization of Repetitive Behavior**

**Nikolay Bliznyuk\***, University of Florida  
**Isaac H. Duerr**, University of Florida  
**Amber Muehleman**, University of Florida  
**Mark Lewis**, University of Florida

Repetitive behaviors, also known as stereotypies, are a common feature of a number of clinical disorders and are ubiquitous in normative development. However, little attention has been paid to their origin or temporal dynamics. We characterize these features in a mouse model of repetitive behavior in order to identify trajectories of development and temporal dynamics of developmental changes. We consider frequency- and pattern-based statistical models for spontaneous stereotypy measures and developmental trajectories in order to study how stereotypies develop over time. We focus on functional, point process and time series methods that allow one to cluster patterns of repetitive behavior and to model their temporal evolution over the developmental periods. Because of the lack of research on development of repetitive behaviors, the tools that we propose may prove instrumental in understanding, assessment and treatment of repetitive behavior in clinical populations.

e-mail: nbliznyuk@ufl.edu

## 99. SURVIVAL ANALYSIS

### A Semiparametric Bayesian Approach to Modelling Destructive Weighted Poisson Cure Rate Model

**Arpita Chatterjee\***, Georgia Southern University  
**Narayanaswamy Balakrishnan**, McMaster University

Nonparametric or semiparametric Bayesian models are becoming increasingly popular in the context of cure rate or long term survival models. These models are more robust than their parametric counterparts. Rodrigues et al. (2010) proposed a Bayesian hierarchical destructive Poisson cure rate model to analyze survival data with a surviving fraction. This model assumes that the original number of lesions caused by risk factors is not getting fully recovered by the treatment and thus, it undergoes a destructive process. Moreover, these unrepaired fractions of lesions are competing to give rise to a tumor. In this research we propose a semiparametric counterpart of such models by relaxing the distributional assumption on the unobserved lifetimes. We model the unknown survival distribution with a Weibull Dirichlet Process mixture model, mixing on both the shape and scale parameters of the Weibull kernel, which results in a flexible mixture that can model a wide range of distributional shapes. We finally apply this nonparametric model to a cutaneous melanoma data.

e-mail: [achatterjee@georgiasouthern.edu](mailto:achatterjee@georgiasouthern.edu)

### Support Vector Hazards Regression for Predicting Survival Outcome

**Xiaoxi Liu\***, University of North Carolina, Chapel Hill  
**Yuanjia Wang**, Columbia University  
**Donglin Zeng**, University of North Carolina, Chapel Hill

In biomedical studies, one important and challenging question is to predict the survival outcome using censored data. To overcome the potential misspecification of semiparametric methods, machine learning methods, including inverse probability weighted methods (Goldberg and Kosorok, 2013) and rank-based methods (Van Belle et al., 2011), have recently been proposed to derive nonparametric prediction rules. However, the former require either independent censoring or a correctly specified censoring distribution; while the rank-based methods only use feasible pairs of outcome data. In this paper, we develop a support vector hazards regression for predicting the survival outcome. Our method adapts support vector machines to predict dichotomous outcomes (event or no event) of the counting process among subjects at risk. Theoretically, we show that the decision rule is equivalent to maximize the discrimination power based on hazard rate functions. We establish the asymptotic properties of the proposed method, including universal consistency and learning rates. Numerical experiments demonstrate a superior performance of the proposed method to the existing learning methods. Two real data examples are used to illustrate the proposed method.

e-mail: [xiaoxi1@unc.edu](mailto:xiaoxi1@unc.edu)

### Semiparametric Extreme-Value Regression Model for Analyzing Biomarker-Defined Time-to-Event

**Noorie Hyun\***, University of North Carolina, Chapel Hill  
**Donglin Zeng**, University of North Carolina, Chapel Hill  
**David J. Couper**, University of North Carolina, Chapel Hill

In many longitudinal medical studies, the time to a disease event is determined by the value of some biomarker crossing a specified threshold. However, the exact time when the threshold is crossed is not observable and biomarker values are subject to substantial measurement error. Additionally, assuming a fixed threshold for all subjects may not be appropriate for some biomarkers. Medical researchers have showed that thresholds can vary across populations or from person to person. There is no existing method to overcome such challenges simultaneously. In this paper, we propose a semiparametric extreme-value regression model for the strongly skewed distribution of a biomarker. Our model is equivalent to modeling threshold-dependent time to a disease event via a Cox proportional hazards model, where the threshold-dependent event time is defined as the time of the biomarker values cross any given threshold. To account for the measurement error in the biomarker values, we incorporate additive model. We estimate the model parameters using the pseudo-likelihood by ignoring correlations within a subject and implement computation via the EM algorithm. The method is illustrated through an application to data from a diabetes ancillary study to the Atherosclerosis Risk in Communities (ARIC) Study.

e-mail: [noorie.hyun@gmail.com](mailto:noorie.hyun@gmail.com)

### Spatial Extended Hazard Model with Application to South Carolina Prostate Cancer Data

**Li Li\***, University of South Carolina

This paper quantifies racial cancer survival disparity in South Carolina through the consideration of a Bayesian semiparametric approach to the extended hazards model, with generalization to high-dimensional spatially-grouped data. The baseline hazard function is modeled using a novel penalized B-spline that a priori follows a parametric hazard function. County-level spatial correlation is accommodated marginally through the copula model of Li and Lin (2006), using a correlation structure implied by an intrinsic conditionally autoregressive prior. Efficient McMC algorithms are developed, especially applicable to fitting very large, highly-censored areal survival data sets. Per-variable tests for proportional hazards, accelerated failure time, and accelerated hazards are efficiently carried out with and without spatial correlation through Bayes factors. The resulting reduced, highly interpretable spatial models fit significantly better than the additive Cox model with spatial frailties.

e-mail: [lil@email.sc.edu](mailto:lil@email.sc.edu)

### **Local Polynomial Density Estimation with Interval Censored Data**

**Derick R. Peterson\***, University of Rochester  
**Mark J. van der Laan**, University of California, Berkeley

A survival time is interval censored if only its current status, an indicator of whether the event has occurred, is observed at a possibly random number of monitoring times. We provide estimators with pointwise confidence limits for all derivatives of the distribution of the time till event, assuming that the observed monitoring times are independent of the time of interest. Our estimator is a standard local polynomial regression smoother applied to the pooled sample of dependent current status observations. We show that the proposed estimator has a normal limiting distribution identical to that of a smoother applied to independent current status observations. Thus local bandwidth selection techniques and pointwise confidence limit procedures for standard nonparametric regression perform properly, despite the dependence in the pooled sample.

e-mail: peterson@bst.rochester.edu

### **Stacking Survival Models**

**Andrew Wey\***, University of Minnesota  
**John Connett**, University of Minnesota  
**Kyle Rudser**, University of Minnesota

For estimating conditional survival functions, non-parametric estimators can be preferred to parametric and semi-parametric estimators due to relaxed assumptions that enable robust estimation. Yet, even when misspecified, parametric and semi-parametric estimators can possess better operating characteristics in small sample sizes due to smaller variance than non-parametric estimators. Fundamentally, this is a bias-variance tradeoff situation in that the sample size is not large enough to take advantage of the low bias of non-parametric estimation. We extend the method of stacked regressions (Breiman, 1996) to the censored data setting. Stacking combines several survival models, e.g., parametric, semi-parametric, and non-parametric models, for estimating conditional survival functions by minimizing prediction error. By basing the weighted combination of survival models on prediction error, we achieve better performance for conditional survival function estimation. In particular, we demonstrate that stacking survival models can outperform the best performing model chosen by cross-validation through a simulation study and benchmark survival data sets.

e-mail: weyxx003@umn.edu

### **Semiparametric Approach for Regression with Covariate Subject to Limit Of Detection**

**Shengchun Kong\***, University of Michigan  
**Bin Nan**, University of Michigan

We consider generalized linear regression analysis with left-censored covariate due to the lower limit of detection. The complete case analysis by eliminating observations with values below limit of detection yields valid estimates for regression coefficients, but loses efficiency. Substitution methods are biased; maximum likelihood method relies on parametric models for the unobservable tail probability, thus may suffer from model misspecification. To obtain robust and more efficient results, we propose a semiparametric likelihood-based approach for the regression parameters using an accelerated failure time model for the covariate subject to limit of detection. A two-stage estimation procedure is considered, where the conditional distribution of the covariate with limit of detection given other variables is estimated prior to maximizing the likelihood function for the regression parameters. The proposed method outperforms the complete case analysis and the substitution methods in simulation studies. Technical conditions for desirable asymptotic properties are provided.

e-mail: kongsc@umich.edu

## **100. PERSONALIZED MEDICINE**

### **Combining Biomarkers to Optimize Patient Treatment Recommendations**

**Chaeryon Kang\***, Fred Hutchinson Cancer Research Center  
**Holly Janes**, Fred Hutchinson Cancer Research Center  
**Ying Huang**, Fred Hutchinson Cancer Research Center

Markers that predict treatment effect have the potential to improve patient outcomes. Oftentimes there are multiple markers of interest and the task is to derive marker combinations, such as the Oncotype DX score, that are optimized for treatment selection. However most methodology is designed to combine markers to predict outcome under a single treatment. We address the problem of combining markers for treatment selection, which requires modeling the treatment effect as a function of markers. Multiple models of treatment effect are fit iteratively by upweighting or boosting subjects potentially misclassified according to treatment benefit at the previous stage. The boosting approach is compared to existing methods in a simulation study based on the change in expected outcome under marker-based treatment. The boosting approach improves upon existing approaches to combining markers in some settings and has comparable performance in others. Our simulation study also provides insights as to the relative merits of the existing approaches. Application of the boosting approach to the Oncotype DX data produces marker combinations that may have improved performance for treatment selection.

e-mail: ckang2@fhcrc.org

### Simple Approximations to Optimal Treatment Regimes in Randomized Clinical Trial Data

**Jared C. Foster\***, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**Bin Nan**, University of Michigan

**Jeremy MG Taylor**, University of Michigan

We consider the use of randomized clinical trial data to identify simple sets of rules,  $A$ , based on patient information, which may be used to make treatment decisions. The rules are selected by maximizing the expected outcome under a treat-if-in- $A$  assignment regime, and potential rule sets are limited to simple regions of the covariate space, as it is desirable that treatment decisions be made with only a limited amount of patient information required. In particular, we consider a two-stage procedure. In stage 1, nonparametric regression is used to estimate treatment effects for each subject, and in stage 2, a systematic evaluation of many subgroups of a simple, pre-specified form is performed. Using a criterion which is based on the estimated treatment effects obtained in stage 1, the best of these subgroups is identified. We also briefly discuss methods for evaluating identified subgroups, with the goal of decreasing type-I error. As an illustration, the proposed methods are applied to data from an NICHD trial.

e-mail: jaredcf@umich.edu

### Regularized Outcome Weighted Subgroup Identification for Differential Treatment Effects

**Yaoyao Xu\***, University of Wisconsin, Madison

**Menggang Yu**, University of Wisconsin, Madison

**Yingqi Zhao**, University of Wisconsin, Madison

**Quefeng Li**, University of Wisconsin, Madison

**Jun Shao**, University of Wisconsin, Madison

Stratified or personalized medicine is a common theme in current medical research. To facilitate comparative intervention or treatment selection, it is important to identify subgroups that exhibit differential treatment effects. Existing approaches model outcomes directly and then define subgroups according to covariate-treatment interaction. However, outcomes are affected not only by the covariate-treatment interactions, but also by the main effects. Consequently, misspecification of the main effects part interferes with the covariate-treatment interaction estimation thus impedes valid predictive variable identification. We propose a method that uses a target function whose value directly reflect correct treatment assignment for patients. This can separate the covariates' main effects from the covariate-treatment interactions. The function uses patient outcomes as weights instead as modeling targets. Therefore, our method can deal with binary, continuous, time-to-event, and possibly contaminated outcomes in the same fashion. We first focus on identifying estimates from linear rules that characterize important subgroups with binary outcomes. We further consider estimation of differential comparative treatment effects for identified subgroups. We demonstrate the advantages of our method in both simulation studies and analyses of two real data sets.

e-mail: yaoyao@stat.wisc.edu

### Finding Optimal Treatment Dose Using Outcome Weighted Learning

**Guanhua Chen\***, University of North Carolina, Chapel Hill

**Donglin Zeng**, University of North Carolina, Chapel Hill

**Michael R. Kosorok**, University of North Carolina, Chapel Hill

Finding an optimal treatment dose is an important issue in clinical trials. Recently, there are increasing needs for considering individual heterogeneity in finding the optimal treatment dose. In particular, instead of determining a fixed dose for all patients, it is desirable to find a decision rule as a function of patient characteristics such that the expected clinical outcome in the population level is maximized. In this paper, we propose a randomized trial design for optimal dose finding and provide a corresponding analysis method. We show that our proposed dose finding method using randomized trial data can be regarded as an inverse probability weighting estimator of expected clinical outcome. Further, we show the estimation problem is equivalent to solve a weighted regression problem with a truncated L1 loss function. An efficient difference convex algorithm is proposed to solve the associated non-convex optimization problem. We derive the consistency of the estimated decision rule in the sense that the difference in expected clinical outcome of the estimated rule and the optimal treatment rule converges to zero when the sample size goes to infinity. In addition, the performance of the proposed methods and competitive methods are illustrated through both simulation examples and a real data example for dosage identification for Warfarin (an anti-thrombosis drug).

e-mail: guanhuac@live.unc.edu

### Assessing the Heterogeneity of Treatment Effects Via Potential Outcomes of Individual Patients

**Zhiwei Zhang\***, U.S. Food and Drug Administration

**Chenguang Wang**, Johns Hopkins University School of Medicine

**Lei Nie**, U.S. Food and Drug Administration

**Guoxing Soon**, U.S. Food and Drug Administration

There is growing interest in understanding the heterogeneity of treatment effects (HTE), which has important implications in treatment evaluation and selection. The standard approach to assessing HTE (i.e., subgroup analyses based on known effect modifiers) is informative about the heterogeneity between subpopulations but not within. It is arguably more informative to assess HTE in terms of individual treatment effects, which can be defined using potential outcomes. However, estimation of HTE based on potential outcomes is challenged by the lack of complete identifiability. This paper proposes methods to deal with the identifiability problem using relevant information in baseline covariates and repeated measurements. If a set

of covariates is sufficient for explaining the dependence between potential outcomes, the joint distribution of potential outcomes and hence all measures of HTE will then be identified under a conditional independence assumption. Possible violations of this assumption can be addressed by including a random effect to account for residual dependence or by specifying the conditional dependence structure directly. The proposed methods are shown to effectively reduce the uncertainty about HTE in an HIV trial.

e-mail: zhiwei.zhang@fda.hhs.gov

### **Identifying Subpopulations with Differential Risk Benefit Profiles**

**Junlong Li\***, Harvard School of Public Health

**Tianxi Cai**, Harvard School of Public Health

Accurate and individualized prediction of risk and treatment response plays a central role in successful disease prevention and treatment. Recent advancement in biological and genomic research has led to the discovery of a vast number of new markers predictive of disease outcomes. These new discoveries hold great potential for improving the prediction of clinical outcomes, and may lead to personalized, tailored medicine. To realize the goals of personalized medicine, significant efforts have been made on building risk prediction models and assessing subgroup-specific treatment effects. Many attempts have also been made to assess who may benefit most from a new treatment. However, most existing procedures focus on a single efficacy or adverse event outcome. In this talk, I'll discuss procedures that aim to identify subpopulations with differential risk benefit profiles with respect to the new treatment.

e-mail: juli@hsph.harvard.edu

### **Active Learning Clinical Trials for Personalized Medicine**

**Yingqi Zhao\***, University of Wisconsin, Madison

**Stanislav Minsker**, Duke University

**Guang Cheng**, Purdue University

Individualized treatment rules (ITRs) has become increasingly popular due to its adaptively to individual patient characteristics. Data from randomized clinical trials, which are known to be expensive to obtain, are utilized for estimating optimal ITRs. The main purpose of our talk is to propose the active learning method which can automatically and wisely recruit the most informative patients such that the learning cost can be significantly reduced without sacrificing any theoretical accuracy. We provide the risk bounds for the proposed active learning methods. Simulation studies and real data analyses are conducted to show that the proposed method performs favorably to competing methods, e.g., passive learning methods.

e-mail: yqzhao@biostat.wisc.edu

## **101. SPATIAL TEMPORAL MODELS**

### **A Bayesian Hierarchical Spatial Model for Dental Caries Assessment Using Non-Gaussian Markov Random Fields**

**Ick Hoon Jin\***, University of Texas

MD Anderson Cancer Center

**Ying Yuan**, University of Texas

MD Anderson Cancer Center

**Dipankar Bandyopadhyay**, University of Minnesota

Research of dental disease generates data with two levels of hierarchy: that of a tooth overall and that of the different surfaces of the tooth. The outcomes often exhibit spatial referencing among neighboring teeth and surfaces, i.e., the disease status of a tooth or surface might be influenced by the status of a group of neighboring teeth/surfaces. Assessments of dental caries (tooth decay) at the tooth level yield binary outcomes indicating the presence/absence of teeth, and trinary outcomes at the surface level indicating healthy, decayed, or filled surfaces. The presence of these mixed discrete responses complicates the data analysis under a unified framework. To mitigate complications, we developed a Bayesian two-stage model under suitable (spatial) Markov random field assumptions that accommodates the natural hierarchy within the mixed responses. In the first stage, we use an autologistic model to estimate the spatial dependence between existing and missing teeth. In the second stage, conditioned on a tooth being non-missing, we use a Potts model to analyze the spatial referencing for the tooth. To tackle the computational difficulty in the Bayesian estimation scheme that is caused by the intractable normalizing constant, we use a double Metropolis-Hastings sampler. We illustrate the proposed methodology using data from a clinical study at the Medical University of South Carolina.

e-mail: ijin@mdanderson.org

### **Spatial Analysis of Hotel Room Rate: Evidence from Star Rated Hotels in Beijing**

**Chuan Wang\***, University of Florida

**Yang Yang**, Temple University

Insufficient attention has been given to hotel-room-price attributions and its mechanism in the lodging research field till now. This research examines how site and situation factors differently affect lodging industry and room prices. Spatial Varying Coefficient Process (SVCP) is compared with Geographically Weighted Regression (GWR), showing that SVCP not only has an advantage in the aspect of statistical interpret-ability, but also stands out in the goodness of fit. However, considering the modeling of SVCP requires running of Markov Chain Monte Carlo (MCMC) which takes a long time, the validation of modeling becomes a challenging question. Thus a model is proposed to facilitate the validation procedure. Then the potential use of nonparametric Bayesian approach towards the spatial analysis of hotel room rate is discussed.

e-mail: wangchuan1989@gmail.com

### A Sparse Reduced Rank Framework for Group Analysis of Functional Neuroimaging Data

**Mihye Ahn\***, University of North Carolina, Chapel Hill  
**Haipeng Shen**, University of North Carolina, Chapel Hill  
**Weili Lin**, University of North Carolina, Chapel Hill  
**Hongtu Zhu**, University of North Carolina, Chapel Hill

In spatial-temporal neuroimaging studies, there is an evolving literature on the analysis of functional imaging data in order to learn the intrinsic functional connectivity patterns among different brain regions. However, there are only few efficient approaches for integrating functional connectivity pattern across subjects, while accounting for spatial-temporal functional variation across multiple groups of subjects. The objective of this paper is to develop a new sparse reduced rank (SRR) modeling framework for carrying out functional connectivity analysis across multiple groups of subjects in the frequency domain. Our new framework not only can extract both frequency and spatial factors across subjects, but also imposes sparse constraints on the frequency factors. It thus leads to the identification of important frequencies with high power spectra. In addition, we propose two novel adaptive criteria for automatic selection of sparsity level and model rank. Using simulated data, we demonstrate that SRR outperforms several existing methods. Finally, we apply SRR to detect group differences between controls and two subtypes of attention deficit hyperactivity disorder (ADHD) patients, through analyzing the ADHD-200 data.

e-mail: ahnm@email.unc.edu

### Bayesian Hierarchical Models for Two-Phase Studies

**Michelle E. Ross\***, University of Pennsylvania  
**Jon Wakefield**, University of Washington

This talk concerns the development of Bayesian methods for two-phase studies. Two-phase study designs are appealing from an efficiency perspective since they allow sampling to be concentrated in informative cells. A number of likelihood-based methods have been developed for the analysis of two-phase data, but I describe a Bayesian approach which has previously been unavailable. The benefits of a Bayesian approach include relaxation of the reliance on asymptotic inference, and the potential to model data with complex dependencies, for example through the introduction of random effects. In particular, we are interested in the use of two-phase studies in a spatial epidemiological context where one may wish to acknowledge confounding by location via the introduction of spatial random effects. The methods are illustrated using data collected on infant births in North Carolina.

e-mail: michross@upenn.edu

### Spatially Varying Distributed Lag Models

**Jongyu Baek\***, University of Michigan  
**Brisa Sanchez**, University of Michigan  
**Veronica Berrocal**, University of Michigan

Motivated by studies of the health effects of the built environment, we propose a spatially varying distributed lag model (SVDLM). The model is built upon a spatial model of Gelfand et al. (2003) and incorporates a Bayesian distributed lag (DL) constraint (Welty et al., 2009) on the coefficients of spatially lagged built environment covariates. In our motivating example, the lagged covariates are the number of food outlets within several “donut” shaped regions from schools, and the health outcome is the weight status of children attending the school. The SVDLMs allow us to allow us to study the spatial variation in the lagged effects of the built environment features, and how the effects of those features may dissipate with distance from schools. Further, given the complex nature of the proposed model, we develop an exploratory analysis tool to decide whether coefficients of covariates are constant, are heterogeneous, or vary and are correlated across space. Lastly, we propose a novel method for shrinkage slice sampling of multivariate posterior covariance samples using the Sweep operator (Beaton, 1964; Dempster, 1969) to enhance computational efficiency. We applied the SVDLM to examine the association between child’s BMI z-score and the of convenience stores across several distances from schools using a surveillance dataset for 5th grade children enrolled in public schools in the State of California.

e-mail: jongguri@umich.edu

### Efficient Data-Driven Knot Selection for Reduced Rank Spatial Models

**Casey M. Jelsema\***, National Institute of Environmental Health Sciences, National Institutes of Health  
**Shyamal D. Peddada**, National Institute of Environmental Health Sciences, National Institutes of Health

Modern advances have enabled the collection of massive spatial and spatio-temporal datasets (upwards of tens of thousands of observations). Reduced rank models are typically required for computational feasibility when analyzing such datasets. Multiple strategies exist for specifying reduced rank models (RRMs) for both spatial and spatio-temporal models. However, a common feature is to define a reduced-dimension latent process over a selected number of knot locations across the spatial domain. Most work on RRM focuses on estimation of parameters or specification of appropriate prior distributions. Thus far the selection of the knot locations has received relatively little attention. The methods which do address knot selection often rely on computationally intensive Bayesian methods. In this paper we propose an efficient, non-Bayesian, data-driven approach to knot selection for reduced rank spatial models. We illustrate our method through simulations and implement it on a real-world dataset.

e-mail: casey.jelsema@nih.gov

## **Spatiotemporal Hurdle Models for Zero-Inflated Count Data: Exploring Trends in Emergency Department Visits**

**Brian Neelon\***, Duke University

**Howard H. Chang**, Emory University

**Qiang Ling**, Emory University

**Nicole Hastings**, Duke University

Motivated by a study exploring spatiotemporal trends in emergency department use, we develop a class of two-part hurdle models for the analysis of zero-inflated areal count data. The models consist of a binary component that models the probability of any emergency department use and a truncated count component that models the number of emergency department visits given use. The models incorporate both patient- and region-level predictors, as well as spatially and temporally correlated random effects for each model component. We model the random effects via multivariate conditionally autoregressive priors, which induce dependence between the model components and provide spatial and temporal smoothing across adjacent spatial units and time periods, resulting in improved inferences. To accommodate potential overdispersion, we consider a range of parametric specifications for the positive counts, including truncated negative binomial and generalized Poisson distributions. We adopt a Bayesian modeling approach, and posterior computation is handled conveniently within standard Bayesian software. Our results indicate that the negative binomial and generalized Poisson hurdle models vastly outperform the Poisson hurdle model, demonstrating that overdispersed hurdle models provide a useful approach to analyzing zero-inflated spatiotemporal data.

e-mail: brian.neelon@duke.edu

## **102. STATISTICAL METHODS IN CANCER APPLICATIONS**

### **High-Dimensional Nonparametric Surface Estimation with Applications to Drug Combination Studies**

**Xuerong Chen\***, Georgetown University

**Hong-Bin Fang**, Georgetown University

**Ming Tan**, Georgetown University

Drug combination is a critically important approach in cancer and antiviral therapies. A main purpose is to find which combinations are additive, synergistic or antagonistic. To our best knowledge, all existing related studies are focused on two-drug combinations while oncological drug development and therapy often involve combinations of more than two drugs. Hence, it is urgent need to develop design and analysis methods for studies of three-drug combinations. In two-drug combinations, the nonparametric interaction index is often estimated by using thin plate spline technique. Unfortunately, this approach does not work well in three-drug combinations. In this paper, according to the geometrical features of the interaction index, we estimate the nonparametric interaction index surface by

additive model fitting and B-spline approximation. The asymptotic properties of the index estimator are developed. A study of three anticancer drugs, PD184, HA14-1 and CEP3891 is given to illustrate the proposed method.

e-mail: chenxr522@gmail.com

### **Meta-Analysis Sparse K-Means Framework for Disease Subtype Discovery**

**Zhiguang Huo\***, University of Pittsburgh

**George C. Tseng**, University of Pittsburgh

Disease phenotyping by omics data has become a popular approach that potentially can lead to better personalized treatment. Identifying disease subtypes via unsupervised machine learning is a first step towards this goal. In this paper, we extend a sparse K-means method towards a meta-analysis framework to identify novel disease subtypes when expression profiles of multiple cohorts are available. The lasso regularization and meta-analysis identify a unique set of gene features for subtype characterization. An additional pattern matching reward function guarantees consistent subtype signature patterns across studies. Simulations showed validity and better performance of the proposed method. In an application to three large breast cancer data sets, the identified disease subtypes from meta-analysis were characterized with improved accuracy and robustness compared to single study analysis. The model was applied to an independent METABRIC dataset and generated improved survival difference between subtypes. The performance evaluation showed that the proposed method is more accurate and robust than traditional method for single studies. These results provide a basis for diagnosis and development of individualized treatments for disease subgroups.

e-mail: xiaoguang1988@gmail.com

### **Identifying Driver Genes from Somatic Mutations: An Integrative Model-Based Approach**

**Keegan D. Korthauer\***, University of Wisconsin, Madison

**Christina Kendzierski**, University of Wisconsin, Madison

Identifying and prioritizing somatic mutations is an important and challenging area of cancer research that can provide new insights into gene function as well as new targets for drug development. Most methods for prioritizing mutations rely primarily on frequency-based criteria, where a gene is identified as having a driver mutation if it is altered in significantly more samples than expected according to a background model. Although useful, frequency-based methods are limited in that all mutations are treated equally. It is well known, however, that while some mutations have no functional consequence, others may have a major deleterious impact. Consequently, accounting for the likelihood of functional impact for individual mutations is critical to accurate inference. Also important is an accurate background model. Here we develop an integrative model-based approach for inferring driver gene status that incorporates both frequency and functional impact

criteria and accommodates a number of factors to improve the background model. Simulation studies demonstrate advantages of the approach, including a substantial increase in power over competing methods. Further advantages are illustrated in an analysis of data from The Cancer Genome Atlas (TCGA) ovarian project.

e-mail: kdkorthauer@wisc.edu

### Additive Regression Model with Frailty on Semi-Competing Risks Data

**Jinheum Kim\***, University of Suwon

**Youn Nam Kim**, Clinical Trials Center Severance Hospital

**Chung Mo Nam**, Yonsei University College of Medicine

We proposed an illness-death model with Lin and Ying's additive hazard and additive frailty for the regression analysis on semi-competing risks data. Comparing with the Cox-type model, the additive model is more natural and properly partitions the effect of the covariate on one transition into the other transition. In the proposed model, we adapted the additive frailty to describe the association between the covariates and failure time in terms of the risk difference rather than the risk ratio. For the inference, we considered a full maximum likelihood on the complete data and incorporated an EM algorithm and Gauss-Laguerre quadrature method. The proposed model was applied to the data from a national intergroup trial in the 1980's to study the effectiveness of two adjuvant therapy regimens for the improvement of surgical cure rates in stage III colon cancer. We compared the group treated with levamisole plus fluorouracil with the untreated group using the semi-competing risks model with cancer recurrence and death. Finally, we conducted simulations to evaluate the performance of the proposed model under several different scenarios.

e-mail: jkimdt65@gmail.com

### Investigating Herpes Simplex Virus Type 1 and KB Oral Cancer Using Fractional Factorial Designs for Drug Combination Determination

**Hongquan Xu**, University of California, Los Angeles

**Jessica Jaynes\***, University of Nevada, Las Vegas

**Xianting Ding**, Shanghai Jiao Tong University

**Weng Kee Wong**, University of California, Los Angeles

**Chih-Ming Ho**, University of California, Los Angeles

Experimental design and analysis is an effective and commonly used tool in scientific investigations and industrial applications. For drug combination determination, many challenges and complexities arise when trying to understand a system with multiple drugs (e.g., three or more drugs) because the underlying biological system is intrinsically complex and there are potential multiple drug interactions. We propose a new class of composite designs based on a two-level factorial design and a three-level orthogonal array to investigate two biological systems: Herpes simplex virus type 1, with six antiviral drugs, and KB oral cancer, with 11 anti-cancer drugs. We show how the sequential use of two-level and three-level fractional factorial designs can screen for important drugs and

drug interactions, as well as determine potential optimal drug dosages. By understanding the complex drug-drug interactions and drug-cell interactions, we identify optimal drug combinations with minimum additive dosages. These observations have practical implications in the understanding of antiviral, and anti-cancer drug mechanism that can result in better design of drug therapy.

e-mail: jessica.jaynes@unlv.edu

### Recursive Reclassification Using Genomic Markers

**Sean Devlin\***, Memorial Sloan-Kettering Cancer Center

**Irina Ostrovnaya**, Memorial Sloan-Kettering

Cancer Center

**Mithat Gönen**, Memorial Sloan-Kettering Cancer Center

Most cancers have a prognostication system that helps guide clinical care and patient counseling such as TNM staging for solid tumors and cytogenetic risk in hematologic malignancies. As new genomic marker panels are developed, it is of clinical interest to determine how these markers can help refine and improve the predictive accuracy of an existing classification system. In this talk, we outline a new method for incorporating genomic markers into a classification system based on an adaptive recursive partitioning algorithm that utilizes cross-validation for tuning parameter selection. This algorithm is particularly suited for refining the definition of prognostic categories but can be easily tailored for other reclassification needs. Simulation studies show adequate control of false positive reclassification and evaluate the overall accuracy of the algorithm as a function of the number and prevalence of the genomic markers.

e-mail: devlins@mskcc.org

### Impact Of Copula Directional Specification on Multi-Trial Evaluation of Surrogate Endpoints

**Lindsay A. Renfro\***, Mayo Clinic

**Hongwei Shang**, University of Connecticut

**Daniel J. Sargent**, Mayo Clinic

Evaluation of surrogate endpoints using patient-level data from multiple trials is the gold standard, where multi-trial copula models are used to quantify both patient-level and trial-level surrogacy. While limited consideration has been given in the literature to copula choice (e.g., Clayton), no prior consideration has been given to direction of multi-trial copula implementation (via survival versus distribution functions). We demonstrate that even with the correct choice of copula family, directional misspecification leads to biased estimates of patient-level and trial-level surrogacy. We illustrate with a simulation study and a re-analysis of disease-free survival as a surrogate endpoint for overall survival in early stage colon cancer.

e-mail: renfro.lindsay@mayo.edu

## 103. DIAGNOSTIC AND SCREENING TESTS

### **A New Diagnostic Accuracy Measure and Cut-Off Point Selection Criterion**

**Tuochuan Dong\***, State University of New York at Buffalo  
**Kristopher Attwood**, Roswell Park Cancer Institute  
**Lili Tian**, State University of New York at Buffalo

In diagnostic studies, there exist many measures for describing biomarkers' diagnostic accuracy. Most of the measures are only defined for standard two-class diseases although few of them have been extended to diseases with 3 or more classes. An important issue in diagnostic studies is to select the cut-off points. Some of the diagnostic measures, e.g. the generalized Youden index which is defined as optimizing over all the thresholds, can naturally generate cut-off points. This paper proposes a new measure for any  $k$ -class diseases ( $k \geq 2$ ) with an appealing geometric appeal. This new measure not only serves an overall accuracy measure but also can naturally provide cut-off points. Via geometric and probabilistic interpretation, it can be shown that the proposed measure has great advantages compared over other existing measures such as the generalized Youden index. Simulations are implemented to assess its performance for both three-class and four-class diseases. An Alzheimer's Disease example data is analyzed to illustrate the new measure and the corresponding cut-off point selection criterion.

e-mail: tuochuan@buffalo.edu

### **A Bayesian Missing Data Analysis Model for Estimating and Comparing Diagnostic Test Accuracy**

**Yi Hua\***, University of Illinois, Urbana Champaign  
**Chenguang Wang**, Johns Hopkins University

In order to get the FDA Premarket Approval for a diagnostic medical device, investigators often need to compare the device's sensitivity and specificity to a predicate device that is already on the market. Missing data, including missing reference standards and missing test results from any of the two devices, are commonly encountered during the investigation. The missing at random assumption, however, is lacking in interpretation in this non-monotone missingness setting. Second, the widely used conditional independence (CI) assumption is unverifiable when there is missing data and also causes bias in the estimation when misspecified. In this paper, we propose a Bayesian missing data analysis model that identifies the full data model by applying the available case missing value (ACMV) constraints. The model incorporates sensitivity parameters to explore the model robustness when ACMV constraints are relaxed. In addition, the proposed model allows a data driven quantified departure from the conditional independence assumption. We evaluate the approach via simulations and implement it on a recent diagnostic clinical trial.

e-mail: huayi056@gmail.com

### **Application of Latent Class Analysis for Screening Test of Adolescents Suicidal Behavior in United States (1991-2011 YRBSS Survey)**

**Hani Samawi\***, Georgia Southern University  
**Ryan Butterfield**, Odumosu and Butterfield, LLC.

According to World Health Organization (WHO) estimates, in the year 2000, approximately one million people died from suicide, and 10 to 20 times more people attempted suicide worldwide. They reported in 2005 that there were a total of 32,559 deaths from suicide in the United States of those 32,559 there were 4,474 deaths between the ages 5-24. From the Youth Risk Behavior Surveillance System (YRBSS) (CDC, 2009) it is reported that almost 9% of adolescents of high school age attempted suicide at some point in the 12 months prior to survey administration. Screening for a specific disease or condition is a fundamental component of human disease control and prevention. One of our primary goals is to use the latent class approach to estimate the sensitivity and the specificity of the screening test for Adolescents Suicidal Behavior in United States using data from 1991-2011 YRBSS Survey. Our preliminary results indicated the validity of using this technique. Also, we found that using the four suicide behavior indicators (Ideation, Planning, Attempt and Injury from attempt) as screening test has a high Sensitivity (0.9946) and Specificity (0.9233) from 2009 data.

e-mail: hsamawi@georgiasouthern.edu

### **Issues in Reviewing Precision Studies of Quantitative Measurement in Medical Device Submissions in FDA**

**Haiwen Shi**, U.S. Food and Drug Administration  
**Qin Li\***, U.S. Food and Drug Administration

Precision study, also called repeatability and reproducibility study, is an essential part in analytical study when reviewing submissions on diagnostic medical devices in CDRH of FDA. A new draft of CLSI guidance, EP5-A3, which mainly focuses on precision of quantitative measurement, is currently under development. In this talk, we are going to go through a couple of common issues in precision study of quantitative measurement that sponsors have trouble with. Our main focus will be the issue about whether or not we need, if needed then how, to calculate the confidence intervals of the coefficient of variation. Through literature reviews, we will provide some commonly used methods and plan to investigate and compare the properties of these methods by simulations. We also plan to test these methods on a real data set if available. In addition, issues on designing the precision study are planned to be discussed.

e-mail: haiwen.shi@fda.hhs.gov

### On the Relationship Between FROC and ROI Analyses for Detection-Localization Data

**Andriy I. Bandos\***, University of Pittsburgh  
**Nancy A. Obuchowski**, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University

Evaluation of the performance of medical diagnostic systems is essential for development, optimization, and regulatory approval purposes. Medical tests, such as diagnostic mammography, designed to locate possibly multiple lesions per patient are often evaluated in detection-localization studies. The Region of Interest (ROI) and the free-response ROC (FROC) are two statistical approaches used for analysis. These approaches use different data representation and analytical tools, and can lead to different conclusions. We developed a method equating FROC and ROI results on a conceptual, as well as a numerical level, and illustrated the results using two large multireader clinical studies of breast and colon cancer detection. Individual-reader AUC-like indices computed from the ROI and FROC datasets sometimes differed substantially, which can be interpreted as difference in clinical versus technical accuracy levels, respectively. However, we showed that when applied to the same ROI dataset, the difference in these accuracy indices is bounded typically by a small quantity. Furthermore, the calibrated results of the comparison of diagnostic modalities were very similar regardless of whether the same (ROI) dataset or separate (ROI and FROC) datasets were used.

e-mail: anb61@pitt.edu

### Comparison of Diagnostic Performance Levels Using Partial AUC

**Hua Ma\***, University of Pittsburgh  
**Andriy I. Bandos**, University of Pittsburgh  
**David Gur**, University of Pittsburgh

Evaluation of diagnostic performance is often based on the areas under entire ROC curves (AUC). The partial AUC (pAUC) focuses on the range of clinical interest. Its use, however, is limited partially due to the perceived need for larger sample sizes. We investigated properties of comparisons of two pAUCs both analytically and using an extensive simulation study of several families of non-crossing ROC curves. Our results demonstrate that for continuous data the properties of comparisons depend on curvature of the ROC curves, which can be viewed as a measure of informativeness. For concave binormal ROC curves, an increase in the range of interest often leads to an increase in pAUCs difference (whether standardized or not) thereby contributing to a frequent increase in statistical power. However, when ROC curves are flatter, the difference in standardized pAUC is relatively stable, and statistical power frequently decreases with increasing ranges of interest. In particular, differences in pAUCs could be more readily detectable than differences in full AUCs. Curves with low curvature (e.g., bigamma) are often visually similar to the binormal curves. Thus in practice, comparisons based on clinically relevant pAUCs could often require no larger sample sizes as compared with studies based on AUC.

e-mail: xuelang818@hotmail.com

### A Simplifying Reformulation of the Binormal Likelihood-Ratio Model

**Stephen L. Hillis\***, University of Iowa

A basic assumption for a meaningful diagnostic decision variable is that there is a monotone relationship between the decision variable and the likelihood of disease. This relationship, however, generally does not hold for the binormal model. As a result, receiver operating characteristic (ROC) curve estimation based on the binormal model produces improper ROC curves that have hooks, are not concave over the entire domain, and cross the chance line. To avoid this problem, Pan and Metz proposed basing ROC-curve estimation on the likelihood-ratio function of the binormal distribution, which I refer to as the binormal likelihood-ratio (binormal-LR) model. A disadvantage of the binormal-LR model is that the corresponding nondiseased and diseased distributions are unfamiliar and difficult to display. Furthermore, ROC-curve properties are difficult to derive. I show how that the binormal-LR model can be characterized as a bi-variable model that results in familiar distributions for the nondiseased and diseased populations. This reformulation results in better understanding of the model, meaningful plots of the latent decision variable, and easier derivation of formulas (e.g., AUC and pAUC).

e-mail: hillis@lisco.com

## 104. STATISTICAL METHODS FOR BIOMARKER DISCOVERY

### New Class of Bivariate Weibull Distributions to Accommodate the Concordance Correlation Coefficient for Left-Censored Data

**Uthumporn Domthong\***, The Pennsylvania State Hershey College of Medicine  
**Vernon M. Chinchilli**, The Pennsylvania State Hershey College of Medicine

In many clinical studies, Lin's concordance correlation coefficient (CCC) is a common tool to assess the level of agreement of a continuous response measured under two different conditions. However, the complicating feature is that the assay for measuring a specific biomarker typically cannot provide accurate numerical values below the lower limit of detection (LLD), which results in left-censored data. In addition, the CCC is based on a squared distance function, and it can be very sensitive to the effects of the outliers. In this work, we propose a new index for agreement in the presence of left-censored data. We construct a statistical model based on a new class of bivariate Weibull distributions that possesses the properties needed and allow for the left-censoring of low data points. Then, we take a parametric approach to derive maximum likelihood estimates of the means, variances, and covariance to construct the concordance correlation coefficient. Simulation studies confirm that the procedure is flexible, accurate, and relatively robust with respect to outliers. Finally, we used data from an ancillary study of the Assessment, Serial Evaluation, and Subsequent Sequelae of Acute Kidney Injury (ASSESS-AKI) Consortium for demonstration.

e-mail: uxd101@gmail.com

## **A Semi-Parametric ROC Method for Assessing Biomarkers Subject to Measurement Errors and Limit of Detection**

**Le Kang\***, U.S. Food and Drug Administration  
**Weijie Chen**, U.S. Food and Drug Administration  
**Lucas Tcheuko**, U.S. Food and Drug Administration

The receiver operating characteristic (ROC) curve is a widely used tool for assessing biomarkers. Due to imperfections of instruments, the measured biomarker levels are often corrupted with a random error. Estimation of the ROC curve that is free of measurement errors (ME) using repeated measurements is very useful for biomarker development as it reveals the true diagnostic potential of a biomarker after error reduction. Quantification of the ME is also useful in evaluating the repeatability of the biomarker. In addition, due to the limit of detection (LoD) of the instruments, the biomarkers are deemed unmeasurable when the true level is below certain threshold. Parametric methods that rely on the strong assumption of normality have been developed. In this work, we propose a semi-parametric approach that estimates the error-free ROC metrics with two or more replicates of measurements taking into account LoD. The proposed method requires weaker assumption and is more robust than the parametric methods. We have implemented an R package for our method. Extensive simulation studies indicate that our semi-parametric approach provides reliable estimates of the error-free ROC AUC in terms of bias and root mean square error, even when the normality assumption is violated.

e-mail: lekang@live.com

## **Clustering and DMR Identification Using Illumina Methylation Microarray**

**Jeff Campbell\***, Georgia Regents University  
**Duchwan Ryu**, Georgia Regents University  
**Varghese George**, Georgia Regents University  
**Hongyan Xu**, Georgia Regents University  
**Jaejik Kim**, Georgia Regents University

The identification of differentially methylated regions and the classification of samples according to the features of microarray data are of interest but methods for doing so are not well-developed. Methylation data have very high dimension with limited sample size, as well as large amounts of correlation and noise. We propose a functional data analysis method that uses Bayesian nonparametric regression with the sequence of measurements from each individual sample as the response to resolve these difficulties. From the use of Gibbs sampling, we (i) examine the effects of epigenetic factors on the methylation and (ii) compare the Bayes estimates of the methylation function to determine the epigenetic factors bring different functional pattern in some regions to detect the differentially methylated regions. In particular, we use the MCMC samples of smoothing penalties induced by different individual samples. To overcome the long computation of Bayesian method, we enhance our model by adopting latent variables, which represent a group of genomic sites of high correlation. We utilize matched Illumina 450k methylation samples from the Cancer Genome Atlas.

e-mail: jefcampbell@gru.edu

## **Confidence Metrics for Identification of Proteins, Post-Translational Modifications (PTMs) and Proteoforms**

**Naomi C. Brownstein\***, Florida State University  
**Nicolas L. Young**, Florida State University

The two main goals of top-down proteomics are protein identification and quantitation. Protein identification is traditionally defined as determination of the sequence of amino acids, which constitute the protein backbone. Mass spectrometry is employed to separate the protein into fragment ions, the mass-to-charge ratios of which are then measured. Results are compared to a database to determine the underlying protein sequence, and a measure of confidence in identification is reported. However, the presence of additional sources of variability, such as post-translational modifications (PTMs) and splice variants, complicates the problem of identification. Database-centered techniques are not designed to detect variations beyond the amino acid sequence, including the presence and location of PTMs. Consequently, identification of the unique exhaustively defined chemical species termed "proteoform" is difficult. Moreover, metrics for confidence of identified PTMs and proteoforms are needed individually and in the form of summary scores based on the biological question of interest. These problems are dependent on each other and on factors such as sequence identification and confidence, quantitation and signal-to-noise ratios. We define a statistical framework for multiple levels of identification and corresponding confidence metrics, which, at the simplest level, describe only the amino acid sequence and ultimately specify the proteoform.

e-mail: nbrownstein@magnet.fsu.edu

## **Feature Selection for Ranked-Based Classifiers Applied to Cancer Biomarker Discovery**

**Bahman Afsari\***, Johns Hopkins University  
**Luigi Marchionni**, Johns Hopkins University  
**Elana J. Fertig**, Johns Hopkins University  
**Ulisses Braga-Neto**, Texas A&M University  
**Donald Geman**, Johns Hopkins University

Finding robust, simple and mechanistical interpretable gene expression biomarkers have been a challenge in bioinformatics. Rank discriminants, functions on the ordering of a small set of genes, such as Top Scoring Pairs (TSP), have shown potential to overcome this challenge. A rank discriminant, kTSP, applies majority voting among comparisons of multiple ('k') pairs of genes yields powerful learning rules that often excel more complex classification methods. We propose an innovative gene selection for such rank discriminants to address current ad hoc selection of 'k' pairs. Our pair selection criterion is the maximization of the difference of the expected rank discriminant under two phenotypes relative to their variance, inspired by the t-test. We implement this approach using a greedy, two-step framework that reduces computation and mitigates over fitting. In the first step, for each fixed 'k' candidate, we find the set of pairs which maximizes an approximation of the criterion. In step two, among the candidates found in step one, we choose the exact maximizer of the criterion. The

average accuracy of the kTSP using this method is higher than that of SVM-RFE and PAM across 20 cancer datasets. This approach is implemented in an R package for kTSP, available on <http://astor.som.jhmi.edu/~marchion/software.html>.

e-mail: bahman.afsari@gmail.com

### **Meta-Analysis of Regulatory Network on Major Depressive Disorder by Liquid Association**

**Shuchang Liu\***, University of Pittsburgh

**Ying Ding**, University of Pittsburgh

**George C. Tseng**, University of Pittsburgh

Major depressive disorder (MDD) is a heterogeneous psychiatric illness with mostly uncharacterized pathology which may cause death by suicide. Evidence suggests that multi-system disturbances cumulatively influence overall risk of illness. Gene microarray can potentially overcome this complexity by measuring high-throughput gene expression in post mortem brain tissues. Genes exhibiting significantly correlated expression profiles are potentially to be functionally associated. In addition to pairwise correlation, the liquid association (LA) method by Ker-Chau Li was developed to detect gene triplets where two genes may be positively correlated when expression of a third gene (Z) is high but become non- or negatively correlated when expression of Z is low. In this talk, we develop and compare several meta-analysis frameworks for liquid association and show its application to eight MDD transcriptomic studies. The result shows that meta-analysis generates more consistent and verifiable findings of the regulatory networks. These results, together with traditional differential analysis and correlation network, provide novel insights to understand the genetics and neurobiology in MDD.

e-mail: silvia.shuchang.liu@gmail.com

### **Modeling Physical Mixtures of Test Samples to Improve Class Prediction**

**Niels R. Hansen\***, University of Copenhagen

**Martin Vincent**, University of Copenhagen

For class prediction of test samples the accuracy usually degrades if the test samples come from a different distribution than the samples used for training. This is shown to be a problem when molecular signatures from biopsies of metastases are used to predict the origin of the primary tumor. In this case the predictor is trained on primary tumor samples. One difference between the biopsies and the primary tumor samples is the presence of non-metastatic tissue in the biopsies. The molecular signature is therefore a mixture originating from the physical mixture of different cell types. Simple models are presented of how the physical mixture affects the molecular signature, and the models are shown to be able to improve prediction accuracy. The models can be combined with a predictor in different ways; by perturbing training samples to mimic test samples before training of the predictor, by perturbing the predictor directly, or by de-mixing the molecular signature obtained from the test samples. The different combinations are compared in terms of prediction accuracy, versatility and computational efficiency.

e-mail: Niels.R.Hansen@math.ku.dk

## **105. MODERN SURVIVAL ANALYSIS IN OBSERVATIONAL STUDIES**

### **More Efficient Estimator for Additive Hazard Model for Case-Cohort Studies**

**Jianwen Cai\***, University of North Carolina, Chapel Hill

**Soyoung Kim**, Fred Hutchinson Cancer Research Center

**David Couper**, University of North Carolina, Chapel Hill

Case-cohort study design is one of the common methods to save cost in large observational cohort studies. The design consists of a random sample of the entire cohort, named subcohort, and all the subjects with the disease of interest. One important advantage of the case-cohort study design is to re-use the same subcohort when several diseases are of interest. In multiple case-cohort studies, covariates collected on subjects with other diseases are available when estimating the risk effect on one disease. Usually, the analysis is done separately for each disease ignoring data collected on subjects with the other diseases. To make better use of available information, we propose more efficient estimators. We consider the additive hazards models for stratified case-cohort studies with rare and non-rare diseases. We consider both joint analysis and separate analysis. We propose an estimating equation approach with a new weight function. The proposed estimators are shown to be consistent and asymptotically normally distributed. Simulation studies show that the proposed methods using all available information gain efficiency. We apply our proposed method to the data from the Atherosclerosis Risk in Communities (ARIC) study.

e-mail: cai@email.unc.edu

### **Contrasting Group-Specific Cumulative Means Associated with Marked Recurrent Events in the Presence of a Terminating Event**

**Rick Ma**, Regeneron Pharmaceuticals

**Douglas E. Schaubel\***, University of Michigan

In many biomedical studies where the event of interest is recurrent (e.g., hospital admission), marks are observed upon the occurrence of each event (e.g., medical costs, length of stay). Few methods of analysis have been developed under a framework where subjects experience both marked recurrent events and a terminating event (e.g., death), a frequently occurring data structure. We propose two sets of semiparametric methods which contrast group-specific cumulative means, each computed as an integrated mark process. The first method utilizes a form of hierarchical modeling for the terminating event; the conditional recurrent event rate given survival; and for the mark, given an event has occurred. The second method casts the cumulative mean in terms of a counting process for the mark. In each case, large-sample properties are derived, with simulation studies conducted to assess finite sample properties. We apply the proposed methods to data obtained from the Dialysis Outcomes and Practice Patterns Study (DOPPS).

e-mail: deschau@umich.edu

## **Gateau Differential Based Boosting For Time-Varying Survival Models**

**Yi Li\***, University of Michigan  
**Ji Zhu**, University of Michigan  
**Kevin He**, University of Michigan

Survival models with time-varying effects provide a flexible framework for modeling the effects of covariates on event times. In view of existing work that is often focused on time-varying models with relatively low dimensionality, we propose a new Gateau differential-based boosting procedure for simultaneously selecting and automatically determining the functional form of covariates. Specifically, our procedure allows that in each boosting learning step only the best-fitting base-learner (and therefore the most informative covariate) is added to the predictor, and consequently encourages sparsity. In addition, our method controls smoothness, which is crucial for improving the predictive performance. The performance of the proposed method is examined by simulations and by applications to analyze the multiple myeloma data and national kidney transplant data.

e-mail: yili@med.umich.edu

## **Screening For Osteoporosis In Postmenopausal Women: A Case Study In Interval Censored Competing Risks Data**

**Jason Fine\***, University of North Carolina, Chapel Hill

Current US Preventive Services Task Force encourages osteoporosis screening using bone mineral density but does not specify a screening interval or ages to start and stop testing using an evidence based rationale. The current analysis explores these issues using data from the Study of Osteoporotic Fractures, the longest running cohort study of osteoporosis in the United States. Complications arise: time to osteoporosis in individuals free of osteoporosis, prior fracture, and previous preventive treatment, is subject to potentially dependent censoring by fracture and preventive treatment. Endpoint definition is addressed in a competing risks framework, with a certain cumulative incidence function correctly defining the risk of osteoporosis for the screening population. The analysis of this quantity is based on intermittent bone mineral density testing. Likelihood based inference, both full and "naïve", is investigated for such interval censored competing risks data, using a direct modelling strategy for the cumulative incidence functions. The screening interval is defined as a fixed time for a specified percentage of non-osteoporotic women to develop osteoporosis, accounting for the potentially dependent competing risks, which involves the use of so-called competing risks quantiles. The competing risks analysis illustrates how osteoporosis risks may be precisely quantified and used to develop evidence based policy for osteoporosis screening.

e-mail: jfine@bios.unc.edu

## **106. RECENT DEVELOPMENT ON PERSONALIZED MEDICINE**

### **Q-Learning with L1 Regularization**

**Min Qian\***, Columbia University

Recent research in treatment and intervention science is shifting from the traditional "one-size-fits-all" treatment to dynamic treatment regimes, which allow greater individualization in programming over time. A dynamic treatment regime is a sequence of decision rules that specify how the dosage and/or type of treatment should be adjusted through time in response to an individual's changing needs. Constructing an optimal dynamic treatment regime is challenging because the objective function is the expectation of a weighted indicator function that is non-concave in the parameters. In addition, there are many variables in the observed sample, yet cost and interpretability considerations imply that fewer rather than more variables should be included in the developed dynamic treatment regimes. To address these challenges we consider estimation based on L1 regularized Q-learning. This approach is justified via a finite sample upper bound on the difference between the mean response due to the estimated dynamic treatment regimes and the mean response due to the optimal dynamic treatment regime.

e-mail: mq2158@columbia.edu

### **Personalized Medicine and Artificial Intelligence**

**Michael R. Kosorok\***, University of North Carolina, Chapel Hill

Personalized medicine is an important and active area of clinical research involving high dimensional data. In this talk, we describe some recent design and methodological developments in clinical trials for discovery and evaluation of personalized medicine. Statistical learning tools from artificial intelligence, including machine learning, reinforcement learning and several newer learning methods, are beginning to play increasingly important roles in these areas. We present illustrative examples of issues and approaches in treatment of depression, cancer, and other diseases. The new approaches have significant potential to improve health and well-being.

e-mail: kosorok@unc.edu

### **Bayesian Methods for Dose-Finding with Targeted Agents in Early Phase Trials**

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

Patient heterogeneity is a central issue in clinical trials of targeted agents, which are designed to be most effective in patients who express specific biological targets. Early phase oncology trials of targeted agents differ from conventional trials of cytotoxic agents in that toxicities may be qualitatively different or less severe, efficacy may be characterized by both an early biological event and later clinical response, and patients who are positive for

the target may have higher response rates. In this talk, I will discuss some Bayesian sequentially adaptive designs that address these issues. To make things concrete, the talk will focus on a trial of standard chemo-radiation therapy with or without a new agent targeting the KRAS pathway in patients with locally advanced non-small-cell lung cancer. Several possible dose-finding designs for this trial will be discussed, depending on whether KRAS negative patients are included, known prognostic covariates are accounted for, dose-finding is done based on toxicity alone or both efficacy and toxicity, and chemo-radiation only arm is included. A phase I-II design will be discussed that chooses patient-specific doses, so-called individualized therapy, by accounting for effects of dose, prognostic covariates, biomarkers, dose-covariate interactions, and efficacy toxicity trade-offs.

e-mail: pftHall@yahoo.com

### Use of DNA Sequencing in Oncology Discovery Clinical Trials

**Richard Simon\***, National Cancer Institute, National Institutes of Health

Most cancer drugs are developed with defined molecular targets. These drugs are expected to only be effective for patients whose tumors are driven by de-regulation of a drug target. Because of the complexity of cancer biology, it is not clear from biological and pre-clinical studies how to define companion diagnostics for such drugs and information from early phase clinical trials is needed. I will review a new generation of oncology discovery trials that use DNA sequencing for a panel of genes to generate such information. These trials also have the potential for addressing some of the inefficiencies in developing drugs for extensively stratified diseases. The challenges in conducting these trials will be discussed.

e-mail: rsimon@nih.gov

## 107. CAUSAL INFERENCE IN THE ASSESSMENT OF SURROGATE MARKERS

### Measures of Surrogacy Using Principal Stratification

**Jeremy MG Taylor\***, University of Michigan  
**Anna Conlon**, University of Michigan  
**Michael R. Elliott**, University of Michigan

In clinical trials, a surrogate outcome variable ( $S$ ) can be measured before the outcome of interest ( $T$ ) and may provide early information regarding the treatment ( $Z$ ) effect on  $T$ . Using the principal surrogacy framework, we consider an approach that has a causal interpretation and develop a Bayesian estimation strategy for surrogate validation when the joint distribution of potential surrogate and outcome measures is multivariate normal. From the joint conditional distribution of the potential outcomes of  $T$ , given the potential outcomes of  $S$ , we propose surrogacy validation measures from this model. Because the model cannot be fully identified from the data, we use a Bayesian estimation approach to aid in the estimation of non-identified parameters and use prior distributions that are consistent

with reasonable assumptions in the surrogacy assessment setting. We explore the relationship between these surrogacy measures and the surrogacy measures proposed by Prentice (1989). We extend these ideas to address the scenario of an ordinal categorical variable as a surrogate for a censored failure time true endpoint. A Gaussian copula model is used to model the joint distribution of the potential and surrogate outcomes. The method is applied to data from an advanced colorectal cancer clinical trial.

e-mail: jmgt@umich.edu

### Assessing the Surrogacy Paradox

**Michael R. Elliott\***, University of Michigan  
**Anna Conlon**, University of Michigan  
**Yun Li**, University of Michigan  
**Jeremy MG Taylor**, University of Michigan

As described by Chen (2007) and elaborated by others including Ju (2010) and Vanderweele (2013), the surrogacy paradox occurs when a treatment has a positive causal effect on a surrogate, the surrogate and outcome have a positive association, but the causal effect of the treatment on the outcome is negative. Practically, such a situation can have disastrous effects, if a drug approved on the basis of its effect on a surrogate marker turns out to be harmful on average. We consider the surrogacy paradox in the principal surrogacy setting, both in the single trial setting and by extending into the meta-analytic setting, determining the posterior distribution of the probability that a positive treatment effect for a marker will be paired with a negative treatment effect for an outcome in a future trial.

e-mail: mreliot@umich.edu

### Direct Estimation of Joint Counterfactual Probabilities for the Assessment of Binary Surrogate Endpoints

**Marc Buyse\***, IDDI Inc.  
**Tomasz Burzykowski**, Hasselt University, Belgium  
**Ariel Alonso**, Maastricht University, The Netherlands  
**Geert Molenberghs**, Leuven University, Belgium

In a counterfactual framework for causal inference, the probabilities of the four potential outcomes for a single variable  $[(0,0), (0,1), (1,0) \text{ and } (1,1)]$  depend on a single parameter which we call the counterfactual odds ratio. For two binary variables, e.g. a binary surrogate and a binary true endpoint, the joint probabilities for the 16 potential outcomes depend on the counterfactual odds ratio for the surrogate and for the true endpoint and on other parameters that cannot be estimated from the observed data. Under simple assumptions, direct estimation of these joint probabilities is however possible, with confidence intervals obtained through bootstrapping. Measures of association and of predictive accuracy calculated from these joint probabilities are useful to investigate the validity of a potential surrogate. We will discuss the plausibility of the assumptions required, and illustrate the approach through several examples in ophthalmology and oncology.

e-mail: marc.buyse@iddi.com

## Evaluation of Surrogates of Protection in Pre-Clinical HIV Vaccine Trials

**Dustin M. Long**, West Virginia University  
**Michael G. Hudgens\***, University of North Carolina, Chapel Hill

A critical step toward developing a successful vaccine to control the HIV pandemic entails evaluation of vaccine candidates in non-human primates (NHPs). Historically, these studies have usually entailed challenges with very high doses, resulting in infection of all NHPs in the experiment. More recently, researchers have begun to conduct repeated low-dose challenge (RLC) studies in NHPs that may more closely mimic typical exposure in natural transmission settings. An objective of these studies is to determine immune biomarkers which are surrogate endpoints for infection. In this talk, different designs of RLC studies for assessing such surrogates of protection are considered.

e-mail: mhudgens@bios.unc.edu

## 108. NEW DEVELOPMENTS IN MULTIPLE COMPARISONS PROCEDURES AND VARIABLE SELECTION

### False Discovery Control in Large-Scale Spatial Multiple Testing

**Wenguang Sun\***, University of Southern California  
**Brian Reich**, North Carolina State University  
**Tony Cai**, University of Pennsylvania  
**Michele Guindani**, University of Texas MD Anderson Cancer Center  
**Armin Schwartzman**, North Carolina State University

This talk discusses a unified theoretical and computational framework for false discovery control in multiple testing of spatial signals. We consider both point-wise and cluster-wise spatial analyses, and derive oracle procedures that optimally control the false discovery rate, false discovery exceedance and false cluster rate, respectively. A finite approximation strategy is developed to mimic the oracle procedures on a continuous spatial domain. Our multiple testing procedures are asymptotically valid and can be effectively implemented using Bayesian computational algorithms for the analysis of large spatial data sets. Numerical results show that the proposed procedures lead to more accurate error control and better power performance than conventional methods. We demonstrate our methods for analyzing the time trends in tropospheric ozone in eastern US.

e-mail: wenguans@marshall.usc.edu

## Estimating the Evidence of Replicability in 'OMICS' Research

**Ruth Heller\***, Tel-Aviv University  
**Marina Bogomolov**, Technion

In many application fields such as genomics research, it is customary that a primary study of high dimension is followed by an independent study. The paramount importance of replicating findings has been well-recognized, yet there are no well-established formal statistical methods for evaluating whether findings have been replicated. Informally, findings are regarded as replicated if the follow-up study supports results from the primary study. For example in GWAS, associations of SNPs with phenotype may be regarded as replicated if for some SNPs the p-values for association with a phenotype were 'fairly small' in the primary study, and 'small' in the follow-up study. In this talk a formal statistical approach is proposed for identifying whether findings replicate from one study of high dimension to another. We show that existing meta-analysis methods are not appropriate for this problem, and suggest novel methods instead. We provide methods that are valid for dependent test statistics for both FWER and FDR control over false replicability claims. We demonstrate the usefulness of these procedures via simulations and real data examples.

e-mail: ruheller@post.tau.ac.il

### Statistics Coauthor and Citation Network

**Jiashun Jin\***, Carnegie Mellon University  
**Pengsheng Ji**, University of Georgia

We have collected data for the co-author network and citation network based on all published papers in the following 4 journals: Annals of Statistics, JASA, JRSS-B, and Biometrika in a ten-year period from 2003 to 2012. We propose a new spectral method which we call the SCORE for community detection and extraction. We find several meaningful communities including but are not limited to the objective Bayesian community, dimension reduction community, theoretical machine learning community, and the high dimensional data analysis community. We develop a theoretic framework and show that under mild conditions, SCORE gives consistent community detection both for undirected and directed networks.

e-mail: jjashun@stat.cmu.edu

### Adaptive Controls of FWER and FDR Under Block Dependence

**Wenge Guo**, New Jersey Institute of Technology  
**Sanat K. Sarkar\***, Temple University

Often in multiple testing, the hypotheses appear in non-overlapping blocks with the associated p-values exhibiting dependence within but not between blocks. We consider adapting the Bonferroni method for controlling the familywise error rate (FWER) and the Benjamini-Hochberg method for controlling the false discovery rate (FDR) to such dependence structure without losing their ultimate controls over the FWER and FDR, respectively, in a non-asymptotic setting. We present variants of conventional adaptive Bonferroni and Benjamini-Hochberg methods with proofs of their respective controls over the FWER and FDR. Numerical evidence is presented to show that these new adaptive

methods can capture the present dependence structure more effectively than the corresponding conventional adaptive methods. This paper offers a solution to the open problem of constructing adaptive FWER and FDR controlling methods under dependence in a non-asymptotic setting and providing real improvements over the corresponding non-adaptive ones.

e-mail: sanat@temple.edu

## 109. SPATIAL MODELS AND DYNAMICS APPLIED TO ENVIRONMENTAL SCIENCES AND PUBLIC HEALTH

### A Nonparametric Bayesian Model for Spatial Point Processes with Application to Raccoon Rabies Spread

**Gavino Puggioni\***, University of Rhode Island

**Luca Gerardo-Giorda**, Basque Center for Applied Mathematics, Spain

**Lance Waller**, Emory University

**Leslie Real**, Emory University

In order to limit the spread of several infectious diseases, surveillance plays a major role as a continuous and systematic collection, analysis and model of health-related data. It is crucial to monitor data over time and at different locations to allow outbreak detection, areas at risk and facilitate intervention. In this paper we propose a dynamic Bayesian non parametric approach to positive reporting of rabies in raccoons. Measurements were conducted from 1990 to 2007 at township reporting centers in the state of New York. The proposed model involves a dynamic density estimation problem, with the specification of a prior based on a Dirichlet Process mixture of bivariate normal distributions at each point in time. Temporal dependence is introduced through the atoms that evolve as dynamic linear models.

e-mail: puggioni@cs.uri.edu

### The Role of Weather in Meningitis Spread in Africa

**Yolanda Hagar\***, University of Colorado, Boulder

**Mary Hayden**, National Center of Atmospheric Research

**Abudulai Adams Forgor**, War Memorial Hospital, Ghana

**Tom Hopson**, National Center of Atmospheric Research

**Patricia Akweongo**, University of Ghana

**Abraham Hodgson**, Ghana Health Service

**Andrew Monaghan**, National Center of Atmospheric Research

**Christine Wiedinmyer**, National Center of Atmospheric Research

**Raj Pandya**, University Corporation for Atmospheric Research

**Vanja Dukic**, University of Colorado, Boulder

Bacterial (meningococcal) meningitis is a devastating infectious disease with outbreaks occurring annually during the dry season in locations within the "Meningitis Belt", a region in sub-Saharan Africa stretching from Ethiopia to Senegal. Meningococcal meningitis occurs from December to May in the Sahel with large epidemics every 5-10 years

and attack rates of up to 1000 infections per 100,000 people. High temperatures coupled with low humidity may favor the conversion of carriage to disease as the meningococcal bacteria in the nose and throat are better able to cross the mucosal membranes into the blood stream. Although the transmission dynamics are poorly understood, outbreaks regularly end with the onset of the rainy season and may begin anew with the following dry season. In this talk, we employ a semi-parametric spatial-temporal Poisson model for assessment of the association between number of reported meningitis cases across the meningitis belt and over time, with spatial-temporal data (weather, pollution, population variables). We will explore extensions of the Bayesian multi-resolution hazard (MRH) methodology to accommodate periods of sparse observations intrinsic to meningitis outbreaks. We present the analysis of monthly reported meningitis counts in over 500 districts and 20 countries in the meningitis belt, from 2008-2009.

e-mail: yolanda.hagar@colorado.edu

### A Spatial Point Process Model for Viral Infections

**Murali Haran\***, The Pennsylvania State University

**Joshua Goldstein**, The Pennsylvania State University

**John Fricks**, The Pennsylvania State University

**Francesca Chiaromonte**, The Pennsylvania State University

Understanding the progression of viral infections is of great interest to biologists. Here we use data from the imaging of cell cultures to study the spatial structure of a virus infection. Our study is motivated by an in-vitro cell culture study that identifies and locates cells infected with the human respiratory syncytial virus (RSV). A question of interest is: How does the presence of an infected cell impact infections in neighboring cells? To answer this question, we develop a new spatial point process model that allows for both attraction and repulsion among the cells in the RSV data. We describe a double Metropolis-Hastings algorithm for resolving the considerable computational challenges that arise when fitting our model to the data set.

e-mail: mharan@stat.psu.edu

### Using Genetic Sequences to Infer Population Dynamics: Phylodynamic Analysis of HIV Transmission in SE Michigan

**Edward L. Ionides\***, University of Michigan

Advances in methods for analysis of population-level genetic variation of pathogens can potentially provide useful information about characteristics of donors of infections. This complements conventional epidemiological surveillance of infectious disease, which is focused on identifying recipients of infection. Pathogen genetic sequences are increasingly available for a growing number of infectious diseases. We discuss recent methodological developments that use both sequence data and conventional epidemiological data for inference on dynamic models of disease transmission. As a specific example, we estimate the fraction of HIV transmission that occurs in the first year of the donor's infection. We find that combining conventional and genetic data gives substantial improvement over each alone.

e-mail: ionides@umich.edu

## 110. ADVANCES IN LONGITUDINAL STUDIES FOR PREDICTING CLINICAL OUTCOMES

### Multi-State Analysis of Serial Biomarkers, Non-Terminal, and Terminal Events

**Richard J. Cook\***, University of Waterloo

Markers of disease activity are routinely measured at follow-up visits in cancer clinical trials and interest often lies in modelling their profiles and relation to clinical endpoints. Many methods for the joint analysis of serial markers and clinical events (e.g. hierarchical mixed effect models) do not adequately address the fact that clinical endpoints such as progression are interval-censored, nor do they deal with the terminating effect of death on the marker process. We argue that multi-state modelling offers a convenient framework for coupling the serial marker values with models for clinical events such as progression and death. Simulation studies evaluate and highlight the utility of the proposed methods and data from a recent trial of patients with cancer metastatic to bone are used for illustration.

e-mail: rjcook@uwaterloo.ca

### Generalized Quasi-Likelihood Ratio Tests for Semiparametric Analysis of Covariance Models in Longitudinal Data

**Jin Tang\***, University of Georgia

**Yehua Li\***, Iowa State University

We model generalized longitudinal data from multiple treatment groups by a class of semiparametric analysis of covariance models, which take into account the parametric effects of time dependent covariates and the nonparametric time effects. In these models, the treatment effects are represented by nonparametric functions of time and we propose a generalized quasi-likelihood ratio test procedure to test if these functions are identical. Our estimation procedure is based on profile estimating equations combined with local linear smoothers. We find that the much celebrated Wilks phenomenon which is well established for independent data still holds for longitudinal data if working independence correlation structure is assumed in the test statistic. However, this property does not hold in general, especially when the working variance function is mis-specified. Our empirical study also shows that incorporating correlation into the test statistic does not necessarily improve the power of the test. The proposed methods are illustrated with simulation studies and a real application from heroin addiction treatments.

e-mail: yehuali@iastate.edu

### A Semi-Parametric Longitudinal Model for Predicting Clinical Outcomes

**Sanjoy Sinha\***, Carleton University

**Abdus Sattar\***, Case Western Reserve University School of Medicine

This research was motivated by a clinical study where longitudinal measurements were obtained from premature infants treated with supplemental oxygen. Both high and fluctuating levels of oxygen may lead to infant visual

impairment, or retinopathy of prematurity (ROP). ROP is of great concern as more than 50% of very premature infants, born 3 to 4 months early, are at risk for damage to the eyes. The goal of the study is to develop appropriate models to explore patterns in the longitudinal responses and to predict any future development of severe ROP. The analysis of the data is complicated by the fact that many of the longitudinal responses are missing due to a stochastic missing data mechanism. In this talk, I will present some novel statistical methods based on semi-parametric mixed effects models for analyzing such incomplete longitudinal data with missing responses.

e-mail: sinha@math.carleton.ca

### Predicting Outcomes Using Generalized Linear Mixed Models

**Sophia Rabe-Hesketh\***, University of California, Berkeley

**Anders Skrandal\***, Norwegian Institute of Public Health

An advantage of generalized linear mixed models for longitudinal data is that they can be used for predicting outcomes for individual subjects. When the model parameters have been estimated by maximum likelihood, prediction requires integration over the estimated posterior distribution of the random effects given the observed responses and estimated model parameters. When subjects are nested in clusters, such as doctors, three-level models can be used, and in this case, predictions borrow information from other subjects in the same cluster. Issues discussed include handling three-level data, expressing uncertainty of the predictions, and sensitivity of predictions to violations of model assumptions.

e-mail: sophiarh@berkeley.edu

## 111. NEW DEVELOPMENTS IN EDUCATION, CONSULTING, AND HEALTH POLICY

### The Use of Analogies to Help Clinicians and Investigators Better Understand the Principles and Practice of Biostatistics

**Martin L. Lesser\***, Feinstein Institute for Medical Research

**Meredith Akerman\***, Feinstein Institute for Medical Research

**Nina Kohn\***, Feinstein Institute for Medical Research

For the interaction between the biostatistician and the clinician or research investigator to be successful, not only is it important for the investigator to be able to explain biological and medical principles in a way that can be understood by the biostatistician, so, too, the biostatistician needs tools to help the investigator understand both the practice of statistics and specific statistical methods. In our practice, we have found it useful to draw analogies between statistical concepts and familiar medical or everyday ideas. These analogies, while not necessarily profound, help to stress a point or provide an understanding on the part of the investigator. For example, explaining the reason for using a non-parametric procedure (a general procedure used when the underlying distribution of the data is not known

or cannot be assumed) by comparing it to using broad spectrum antibiotics (a general antibiotic used when the specific bacteria causing infection is unknown or cannot be assumed) can be an effective teaching tool. We present a variety of useful (and hopefully amusing) analogies that can be adopted by statisticians to help investigators at all levels of experience better understand principles and practice of statistics.

e-mail: mlessner@nshs.edu

### **Distributed Data, Confidentiality and Specimen Pooling: Using an Old Tool for New Challenges**

**Paramita Saha Chaudhuri\***, Duke University

In the recent past, electronic health records and distributed data networks emerged as a viable resource for medical and scientific research. As the use of confidential patient information from such sources become more common, maintaining privacy of patients is of utmost importance. For a binary disease outcome of interest, we show that the techniques of specimen pooling could be applied for analysis of large and/or distributed data while respecting patient privacy. Aggregate level data are passed from the network to the analysis centre and can be used very easily with logistic regression for estimation of disease odds ratio associated with a set of categorical or continuous covariates. Pooling approach allows for consistent estimation of the parameters of logistic regression that can include confounders. Additionally, since the individual covariate values can be accessed within a network, effect modifiers can be accommodated and consistently estimated. Since pooling effectively reduces the size of the dataset by creating pools or sets of individual, the resulting dataset can be analyzed much more quickly as compared to the original dataset.

e-mail: paramita.sahachaudhuri@duke.edu

### **Analysis of Resting Metabolic Rate in a Latin Square Design with Repeated Measures**

**William D. Johnson\***, Pennington Biomedical Research Center

**Robbie Beyl**, Pennington Biomedical Research Center  
**Jeffrey Burton**, Pennington Biomedical Research Center

A study of two drugs known to increase metabolic rate was conducted with the levels of each drug administered as a placebo or one of two active doses. The primary aim was to investigate the concept that combined therapy may be superior to monotherapy. Eight subjects reported to Pennington Center Metabolic Laboratory on 8 occasions separated by  $7 \pm 2$  days. The 8 treatment combinations (absent double placebo) were arranged in an  $8 \times 8$  Latin square with the 8 occasions arranged sequentially across the columns and the 8 subjects randomly assigned to the 8 rows. Each row contained a unique sequence of treatments such that each subject received all treatment combinations. Before each visit subjects ate their normal diet then fasted from 9 pm the prior night, refrained from strenuous physical activity for 24 hours, and had no alcohol

or caffeine-containing beverages for 48 hours. During each visit subjects had their blood pressure, pulse, temperature, resting metabolic rate and respiratory quotient measured. After baseline measurements were taken, subjects were given a treatment combination in the form of 2 pills and all measurements were repeated 3 hours later. Statistical aspects of the analysis are discussed in this presentation.

e-mail: william.johnson@pbrc.edu

### **Small Area Estimation of Vaccination Coverage Rates by Combining Time Series and Cross Sectional Data**

**Santanu Pramanik\***, Public Health Foundation of India  
**Ramanan Laxminarayan**, Public Health Foundation of India

Information on population health indicators in India come from a number of surveys that vary in periodicity and depth. For instance, the most recent data on immunization and vaccination coverage indicators are derived from AHS-1, but these are conducted only in nine states of India. The most recent national survey of immunization coverage was conducted in 2009 (Coverage Evaluation Survey). Therefore, reliable immunization coverage data for the entire country since 2009 is lacking. We use an established approach of small area estimation to predict coverage rates of several vaccinations for the remaining states not covered by AHS-1 at the current time. To obtain the estimates we used a linear mixed model that combines data from five cross sectional surveys representing five different time points. Our model involves sampling error of the survey estimates, area specific random effects, autocorrelated area by time random effects and hence borrows strength across both small areas and time. Our model-based estimates are almost identical to the AHS-1 direct estimates for the nine states. This demonstrates the superiority of our model as AHS-1 estimates are highly precise because of their large sample size. The coverage inequality between rural and urban areas has been reduced significantly for most states in India.

e-mail: santanu.pramanik@phfi.org

### **Challenges Using Survey Data to Estimate Problem Gambling Prevalence in the SEIG-MA Project**

**Edward J. Stanek III\***, University of Massachusetts, Amherst

**Rachel A. Volberg**, University of Massachusetts, Amherst

**Robert J. Williams**, University of Lethbridge, Alberta, Canada

Massachusetts passed legislation in the fall of 2012 to allow construction of 3 casinos and a slot parlor in the state. In a progressive step, the legislation required the Social and Economic Impact of Gambling (SEIG) to be evaluated in Massachusetts (MA), including the prevalence of problem gambling, particularly in areas close to new casino sites. We describe the challenges in estimating the prevalence

of problem gambling based on an n=10,000 multi-mode address-based sample (ABS) survey, and an n=5000 online panel survey. The ABS survey interviews one adult per household based on a stratified sample of addresses from the US Postal system. The ABS interviews accrue from response to an online survey, a mail survey, or a telephone interview. The online panel survey is conducted by an independent survey organization. Challenges include defining a framework for estimation that accounts for the different survey modalities, defining populations for geographic target areas and linking survey data to these areas, accounting for survey weights and non-response, and accounting for important distributions of risk factors thought to be related to problem gambling prevalence. We describe the survey design and conduct, and review competing estimation approaches.

e-mail: stanek@schoolph.umass.edu

### **Practical and Statistical Challenges in Developing an HIV Drug Resistance Surveillance Protocol**

**Natalie Exner\***, Harvard University

**Marcello Pagano**, Harvard University

The World Health Organization is responsible for releasing guidance to low- and middle-income countries on HIV drug resistance surveillance in patients initiating antiretroviral treatment (baseline drug resistance) and in patients on treatment for at least 6 months (acquired drug resistance). The results of these surveys inform national treatment programs and the selection of first- and second-line therapies. This guidance may take the form of a set of generalizable survey protocols that can be modified to meet each country's programmatic needs. These protocols must be sufficiently flexible to adapt to the wide range of settings across all low- and middle-income countries globally. In practice, countries vary widely in number of infected patients, characteristics of their HIV epidemic, number and size of HIV clinics, quality of infrastructure, in-country technical capacity, and financial resources. We describe the practical and statistical challenges associated with developing generalizable surveillance protocols, and we describe our proposed, robust solutions. We focus on the creation of flexible and user-friendly tools for sample size calculation and the construction of didactic materials to encourage in-country data analysis.

e-mail: nexner@mail.harvard.edu

## **112. LATEST ADVANCES IN FUNCTIONAL AND IMAGING DATA ANALYSIS**

### **Online Functional Principal Component Analysis**

**David Degras\***, DePaul University

Online/sequential Principal Components Analysis (PCA) is a widespread technique in machine learning. The framework considered here bears important differences with standard online PCA: first, the observations are functional, i.e., continuous-time processes with smoothness properties, rather than vectors. In addition, the observations are only available for a limited time, i.e., only recent data are stored, as opposed to the usual case of an increasing data set. In this research we draw on perturbation theory and regularization methods to develop a new technique for online functional PCA. We compare the numerical performances of this approach to batch PCA and mainstream machine learning algorithms in a simulation study. Convergence results are derived in the case of stationary processes. We also apply this technique in combination with FPC regression to estimate real-time consumption in electricity usage data.

e-mail: ddegрасv@depaul.edu

### **Modeling Binary Functional Data with Application to Animal Husbandry**

**Jan Gertheiss\***, University of Göttingen

**Verena Maier**, Ludwig-Maximilians-University Munich

**Engel F. Hessel**, University of Göttingen

**Ana-Maria Staicu**, North Carolina State University

We observe a group of pigs over a period of about 100 days. On a very dense grid of time points, it is recorded when each pig is eating, leading to binary functional data for each pig and day. One aim of the data analysis is to find pig-specific eating profiles telling us when a certain pig is typically eating. These eating profiles can then be used for clustering/ comparing pigs, or the profiles may be related to the weight of the pig. In addition, there are measurements such as temperature and humidity available that may influence the pigs' behavior. For analyzing these data, we propose a functional logistic regression approach allowing us to model the binary but functional measurements by assuming an underlying smooth subject-specific profile. The method also allows to incorporate additional (potentially non-functional) covariates. Though the approach was originally designed for analyzing the pig data above, it is rather general and hence applicable to other types of binary functional data, too.

e-mail: jgerthe@gwdg.de

## Using Regression Models to Infer Active Connections in Cortex

**Mark A. Reimers\***, Virginia Commonwealth University

One of the aims in the BRAIN initiative is to understand the dynamics of circuits. Here we develop methods for analyzing voltage-sensitive dye measures of neural activity over the surface of mouse cortex across several temporal and spatial scales. We first simplify the problem by separating intrinsic dynamics from inter-regional communication: we try to fit a common model of cellular dynamics to each location on the cortical surface. Then we fit a constrained regression model, whose parameters represent the strength of direct inter-regional communication, on top of the earlier fit to infer active connections. We find that connections inferred from dynamic data often correspond to connections known from anatomical studies. This result suggests that it may be possible to infer active communication in the brain from high-density dynamic data.

e-mail: mreimers@vcu.edu

## Parametric Modulation of Functional MRI Signals: A Mixed Effect Model Approach

**Lei Huang\***, Johns Hopkins University

**Martin Lindquist**, Johns Hopkins University

**Philip Reiss**, New York University Child Study Center

**Ciprian Crainiceanu**, Johns Hopkins University

In conventional task-based functional magnetic resonance imaging (fMRI) studies, the blood-oxygen-level-dependent (BOLD) signals are modeled as the convolution between constant stimuli and haemodynamic response functions. However, there is evidence that the impact of the stimulus at the neural level varies across tasks. This necessitates an extension of the standard notion of parametric modulation for voxelwise linear modeling. This talk will present a pipeline to test and estimate the effect of heterogeneity of the stimuli, using a mixed effect model and exact restricted likelihood ratio test. We apply our method to an fMRI study and investigate the relationship between the estimates and reaction times.

e-mail: huangracer@gmail.com

## Pre-Processing of the Longitudinal Structural Brain Imaging Data: A Case Study

**Jacek Urbanek\***, Indiana University Fairbanks School of Public Health

**Jaroslav Harezlak**, Indiana University Fairbanks School of Public Health

**Elizabeth M. Sweeney**, Johns Hopkins Bloomberg School of Public Health

In a number of recent structural brain imaging studies of neurocognition, longitudinal brain imaging data have been collected, but there is no universally agreed upon method to estimate the rates of change of brain structure volumes. For example in the study of chronically infected HIV patients, we lack accurate measurements of brain tissue loss. Additionally, image pre-processing methods applied before the brain

tissue changes are estimated have significant influence on the outcome of the analysis. In this presentation we compare the brain image pre-processing methods dedicated to removal of extra-cranial tissue, longitudinal registration (co-registration) and intensity normalization. In particular, we compare intensity normalization based on the white matter and skull intensities; and co-registration with and without the extra-cranial tissue. Presented methods are evaluated quantitatively and their general utility for further studies is assessed. Finally, we address the question of longitudinal registration errors and the true disease-specific changes. Discussed results are based on a longitudinal brain imaging data collected from the chronically infected HIV patients.

e-mail: urbanek@agh.edu.pl

## Clustering of Ultra High Dimensional Longitudinal Data

**Seonjoo Lee\***, Columbia University

**Vadim Zipunnikov**, Johns Hopkins University

**Navid Shiee**, Amazon Inc.

**Daniel S. Reich**, National Institute of Neurological

Disorders and Stroke, National Institutes of Health

**Dzung L. Pham**, The Henry Jackson Foundation

**Brian S. Caffo**, Johns Hopkins University

**Ciprian M. Crainiceanu**, Johns Hopkins University

Longitudinal measurements of ultra high dimensional brain images are now realities of science. Indeed, there are hundreds of studies collecting multi-sequence multi-modality brain images at multiple time points on hundreds of subjects over many years. A fundamental problem is how to classify subjects according to their baseline and longitudinal changes of their brain images in the presence of strong spatio-temporal biological and technological measurement error. We propose a fast and scalable clustering approach by defining a distance metric between latent trajectories of brain images. Methods were motivated by and applied to a longitudinal brain morphology study of multiple sclerosis. Results indicate that there are two distinct patterns of ventricular change that are associated with multiple sclerosis outcomes.

e-mail: seonjool@gmail.com

## Effects of Registration on Statistical Analysis of MRI Data

**Ani Eloyan\***, Johns Hopkins University

**Haochang Shou**, Johns Hopkins University

**Russell T. Shinohara**, University of Pennsylvania

**Elizabeth M. Sweeney**, Johns Hopkins University

**Mary B. Nebel**, Johns Hopkins University

**Daniel S. Reich**, National Institute of Neurological

Disorders and Stroke, National Institutes of Health

**Martin A. Lindquist**, Johns Hopkins University

**Ciprian M. Crainiceanu**, Johns Hopkins University

The localization of lesions for people with multiple sclerosis is thought to be associated with the severity of adverse health effects as measured by disease disability scores. However, several factors hinder statistical analyses of such

associations using large magnetic resonance imaging datasets. For instance, the spatial registration algorithms developed for healthy individuals may be less effective on diseased brains, the interpretation of results requires the careful selection of confounders when modeling the association of disease severity and lesion localization and the choice of the method for correction of multiple comparisons in voxel-wise analysis of association has a prominent role in the interpretation of the results. In this talk, we discuss the performance of four registration algorithms. In addition, methods for dealing with confounding factors due to differences in disease duration and local lesion volume are introduced.

e-mail: aeloyan@jhsp.edu

## 113. BAYESIAN METHODS

### **Joint Models for Multivariate Longitudinal Measurements and a Binary Event: An Application to a Fetal Growth Study with Longitudinal Ultrasound Measurements**

**Sungduk Kim\***, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**Paul S. Albert**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

A large fetus is a major concern to obstetricians, affecting both maternal and fetal morbidity and mortality. Predicting large fetuses at birth has long been a challenge in obstetric practice. We use the data from a previous study conducted in Norway and Sweden from 1986-1989 in which each pregnant woman had four ultrasound exams at around 17, 25, 33 and 37 weeks of gestation. At birth, an infant is diagnosed with macrosomia if their birth weight > 4,000 grams. The focus of this paper is on developing a flexible class of joint models for the multivariate longitudinal ultrasound measurements that can be used for predicting a binary event at birth. A skewed multivariate scale-mixture normal distribution is proposed for the multivariate ultrasound measurements and the skewed generalized t-link is assumed for the link function relating the binary event and the underlying longitudinal processes. We consider a shared random effect to link the two processes together. The proposed model can accommodate not only a flexible error distribution for the ultrasound measurement, but also asymmetric link for the binary event. Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. Several variations of the proposed model are considered and compared via the deviance information criterion. The proposed methodology is motivated by and applied to a longitudinal fetal growth study.

e-mail: kims2@mail.nih.gov

### **Bayesian Peer Calibration Based on Network Position with Application to Alcohol Use**

**Miles Q. Ott\***, Carleton College

**Joseph H. Hogan**, Brown University

**Krista J. Gile**, University of Massachusetts, Amherst

**Crystal Linkletter**, Mathworks

**Nancy P. Barnett**, Brown University

Peers are often able to provide important additional information to supplement self-reported behavioral measures. The study motivating this work collected data on alcohol in a social network formed by college students living in a freshman dormitory. By using two imperfect sources of information (self-reported and peer-reported alcohol consumption), rather than solely self-reports or peer-reports, we are able to gain insight into alcohol consumption on both the population and the individual level, as well as information on the discrepancy of individual peer-reports. We develop a novel Bayesian comparative calibration model for continuous, count and binary outcomes that uses covariate information to characterize the joint distribution of both self and peer-reports on the network for estimating peer-reporting discrepancies in network surveys, and apply this to the data for fully Bayesian inference. We use this model to understand the effects of covariates on both drinking behavior and peer-reporting discrepancies.

e-mail: mott@carleton.edu

### **Modeling Long-Term HIV Dynamics with Left Censoring Measurements**

**Tao Lu\***, State University of New York at Albany

HIV dynamic model offers a different perspective of studying pathogenesis of HIV infection and developing treatment strategy for AIDS patients. In long-term HIV dynamics, viral load often rebound or resurgence in later stage of treatment. The failure is primarily due to reduced drug efficacy. To characterize long-term AIDS treatment, time varying drug efficacy is incorporated into the ODE model. As a result, the ODE system does not have an analytical solution and has to be solved numerically. Furthermore, due to technology constraints, in practice, the measurement of viral load is usually censored at the detection limit. The statistical inference may be biased without accounting for the censored data. We propose the idea of coupling nonlinear mixed-effects ODE model with Stochastic Approximation EM (SAEM) algorithm to take into account both long-term HIV dynamics and under detection limit data. The proposed method is able to handle ODE without close form solution in the longitudinal setting which is often encountered in long-term clinical trial. We illustrate the performance of the proposed method in both simulation and real case examples. We hope these results inspire more research to clarify biological mechanism of HIV infection as well as develop better cures.

e-mail: stat.lu11@gmail.com

### **A Bayesian Missing Data Framework for Generalized Multiple Outcome Mixed Treatment Comparisons**

**Hwanhee Hong\***, University of Minnesota  
**Haitao Chu**, University of Minnesota  
**Jing Zhang**, University of Minnesota  
**Bradley P. Carlin**, University of Minnesota

Bayesian statistical approaches to mixed treatment comparisons (MTCs) are becoming more popular due to their flexibility and interpretability. Many randomized clinical trials report multiple outcomes with possible inherent correlations. Moreover, MTC data are typically sparse (though richer than standard meta-analysis, comparing only two treatments) and researchers often choose study arms based on previous trials. In this paper, we summarize existing hierarchical Bayesian methods for MTCs with a single outcome, and introduce novel Bayesian approaches for multiple outcomes simultaneously, rather than in separate MTC analyses. We do this by incorporating partially observed data and its correlation structure between outcomes through contrast- and arm-based parameterizations that consider any unobserved treatment arms as missing data to be imputed. We also extend the model to apply to all types of generalized linear model outcomes, such as count or continuous responses. We offer a simulation study under various missingness mechanisms (e.g., MCAR, MAR, and MNAR) providing evidence that our models outperform existing models in terms of bias and MSE, then illustrate our methods with two real MTC datasets. We close with a discussion of our results and a few avenues for future methodological development.

e-mail: hong0362@umn.edu

### **Clustering Significant Regions of Brain Activation Using fMRI Meta Data**

**Meredith Ray\***, University of South Carolina  
**Hongmei Zhang**, University of Memphis  
**Jian Kang**, Emory University

We developed a Bayesian clustering method for identifying groups of significant regions of brain activation (foci). The data used are meta data originated from functional magnetic resonance imaging (fMRI), which has the ability to measure the intensity of blood flow and oxygen to a location within the brain that was activated by a given thought or emotion. Meta-analyses are used to summarize fMRI studies to increase the sample size and therefore testing power and reproducibility. We considered two levels of clustering, latent foci center and study activation center, utilizing the Dirichlet process built into a spatial Poisson point process describing the distribution of foci. Intensity was modeled as a function of distance between the focus and the center of the cluster of foci using Gaussian kernels. Simulation studies are conducted to evaluate the sensitivity and robustness of the proposed method with respect to cluster identification and underlying data distributions. The method is illustrated using a meta data of emotion foci.

e-mail: mere2110@yahoo.com

### **Bayesian Factorizations of Big Sparse Tensors**

**Jing Zhou\***, University of North Carolina, Chapel Hill  
**Anirban Bhattacharya**, Duke University  
**Amy H. Herring**, University of North Carolina, Chapel Hill  
**David B. Dunson**, Duke University

It has become routine to collect data that are structured as multiway arrays (tensors). One motivating example is the National Birth Defects Prevention Study (NBDPS), which was designed to evaluate environmental, behavioral, biomedical, and sociodemographic associated with the occurrence of congenital malformations. Such data can be placed into a large contingency table (tensor) with cell counts defined as the number of subjects with each possible combination of covariates. There is literature on low rank tensor factorization often relying on parallel factor analysis (PARAFAC), which expresses a rank  $k$  tensor as a sum of rank one tensors. However, when observations, as in our NBDPS study, are only available for a tiny subset of the cells of a big tensor, the low rank assumption is not sufficient, and PARAFAC has poor performance. We induce an additional layer of dimension reduction by allowing the effective rank to vary across dimensions of the table. Taking a Bayesian approach, we place priors on terms in the factorization and develop an efficient Gibbs sampler for posterior computation. The methods are shown to have excellent performance in simulations, and results in birth defects epidemiology are presented.

e-mail: jingzhou@live.unc.edu

## **114. MULTIVARIATE SURVIVAL ANALYSIS**

### **Inference on Quantile Residual Life for Semi-Competing Risks Data**

**Wen-Chi Wu\***, University of Pittsburgh  
**Jong-Hyeon Jeong**, University of Pittsburgh

For randomly censored data, the residual life function at a given time determines a life distribution of a subject survived up to that time point. In the situation where the data are censored, or where the underlying distribution is skewed, the quantile residual life function is preferred. A large number of studies regarding the quantile residual lifetime has been conducted in the univariate settings by many professionals. However, patients may often experience multiple types of events. The conditional quantile residual lifetime to the terminal event, i.e. time to mortality, given the occurrence of the nonterminal event, i.e. time to morbidity, beyond time  $t$  might be of upmost interest. Such situation applies to a semi-competing risks setting subject to right censoring. An estimator for conditional quantile residual life function at a fixed time point is obtained by inverting the estimating equation. The estimator is shown to be asymptotically consistent and converges to a zero-mean Gaussian process. Finally, the covariate effects at specific pairs of event times are evaluated based on a log-linear regression on conditional quantile residual lifetime. Simulation studies demonstrate the performance of the estimator for a moderate sample size. The proposed method will be applied to a breast cancer dataset from a phase III clinical trial.

e-mail: wenchi.wu82@gmail.com

### **Model Selection and Goodness-of-Fit Test Procedures for Copula Models**

**Antai Wang\***, New Jersey Institute of Technology

In this talk, we first present an important formula of marginal survival functions for a given Archimedean copula model. Based on this formula, we propose new model selection and test procedures for Archimedean copula models and demonstrate the effectiveness of our strategies using simulations and real data examples.

e-mail: aw224@njit.edu

### **Analysis of Recurrent Events Data Based on Accelerated Recurrence Time Model**

**Xiaoyan Sun\***, Emory University

**Limin Peng**, Emory University

**Yijian Huang**, Emory University

**Amita K. Manatunga**, Emory University

**Hui-Chuan Lai**, University of Wisconsin, Madison

The accelerated recurrence time (ART) model extends the modeling strategy of censored quantile regression to recurrent events settings, offering easy physical interpretations of time to expected recurrence frequency and the flexibility to investigate evolving covariate effects. In this work, we develop a new ART regression procedure by utilizing a mean zero stochastic process associated with recurrent events counting process. The new method enables more efficient and stable computation as compared to existing methods. Moreover, it can readily accommodate the more general recurrent events scenario, where the observation of recurrent events is subject to left censoring, for example, due to delayed entry as occurred in many observational studies. We derive the asymptotic properties of the proposed estimator, and develop inference procedures. Particularly, we present a new sample-based procedure for covariance estimation, which is much computationally faster than bootstrapping-based inference and can easily be adapted to single event survival settings. Simulation studies demonstrate satisfactory sample performance of our proposals. We illustrate the utility of the proposed method via an application to a dataset from the US Cystic Fibrosis Foundation Patient Registry (CFFPR).

e-mail: xsun33@emory.edu

### **Simple Two-Stage Semiparametric Estimation of the Positive Stable Shared Frailty Model**

**Yu Han\***, University of Rochester

**Changyong Feng**, University of Rochester

**Xin Tu**, University of Rochester

Positive stable shared frailty Cox proportional hazards model is widely used in the analysis of correlated survival data due to its attractive features. However, the estimation of the parameters in the model is not so straightforward. In this paper, we devise a simple two-stage estimation procedure for such model, which can be easily implemented with the standard statistical packages. The large sample properties of the proposed estimators are developed and simulation studies show that the estimation approach has comparable efficiency with existing methods.

e-mail: yu\_han@urmc.rochester.edu

### **Nonparametric Estimation of Joint Distribution of Time from Umbilical Cord Blood Transplantation to First Infection and Gap Times Between Recurrent Infections**

**Chi Hyun Lee\***, University of Minnesota

**Xianghua Luo**, University of Minnesota

**Chung-Yu Huang**, Johns Hopkins University

**Todd DeFor**, University of Minnesota

Infection is one of the most common complications after hematopoietic cell transplantation. Many patients experience infectious complications repeatedly after transplant. Existing statistical methods for recurrent gap time data typically assume that patients are enrolled due to the occurrence of an event of the same type as the recurrent event or assume that all gap times, including the first gap, are identically distributed. Applying these methods on the post-transplant infection data will inevitably lead to incorrect inferential results because the time from the transplant to the first infection has a different biological meaning than the gap times between recurrent infections. Some inefficient methods may include using the univariate survival analysis methods on the first infection only data or using the bivariate serial event data methods on the data up to the second infections. In this paper, we propose a nonparametric estimator of the joint distribution of time from transplant to the first infection and the gap times between following infections. The proposed estimator takes into account the potentially differential distribution of the two types of times and fully utilizes the data of recurrent infections from patients. Asymptotic properties of the proposed estimators are established.

e-mail: leex5865@umn.edu

### **Safe Trials for Equivalence of Two Survival Functions: Alternative to the Tests Under Proportional Hazards**

**Elvis Martinez\***, Florida State University

**Wenting Wang**, University of Texas

MD Anderson Cancer Center

**Debajyoti Sinha**, Florida State University

**Stuart Lipsitz**, Harvard Medical School

**Richard Chappell**, University of Wisconsin, Madison

For either the equivalence trial or the non inferiority trial with survivor outcomes from two treatment groups, the most popular testing procedure is the extension (e.g, Wellek, 1993) of log-rank based test under the proportional hazards model (PHM). We show that the actual type I error rate for the popular procedure of Wellek (1993) is higher than the intended nominal rate when survival responses from two treatment arms satisfy the proportional odds survival model (POSM). When the true model is POSM, we show that the hypothesis of equivalence of two survival functions can be formulated as a statistical hypothesis involving only the survival odds-ratio parameter. We further show that our new equivalence test, formulation, and related procedures

are applicable even in the presence of additional covariates beyond treatment arms, and the associated equivalence test procedures have correct type I error rates under the PHM as well as the POSM. These results show that use of our test will be a safer statistical practice for equivalence trials of survival responses than the commonly used log-rank based tests.

e-mail: elvism@stat.fsu.edu

### **Composite Likelihood for Joint Analysis of Multiple Multistate Processes Via Copulas**

**Liqun Diao\***, University of Rochester

**Richard J. Cook**, University of Waterloo

A copula-based model is described which enables joint analysis of multiple progressive multistate processes. Unlike intensity-based or frailty-based approaches to joint modeling, the copula formulation proposed herein ensures that a wide range of marginal multistate processes can be specified and the joint model will retain these marginal features. The copula formulation also facilitates a variety of approaches to estimation and inference including composite likelihood and two-stage estimation procedures. We consider processes with Markov margins in detail, which are often suitable when chronic diseases are progressive in nature. We give special attention to the setting in which individuals are examined intermittently and transition times are consequently interval-censored. A simulation study gives empirical insight into the relative efficiency of the different methods of analysis and an application involving progression in joint damage in psoriatic arthritis provides further illustration.

e-mail: Liqun\_Diao@URMC.Rochester.edu

## **115. STATISTICAL ANALYSIS IN THE PRESENCE OF MISSING DATA**

### **Variable Selection and Prediction with Incomplete High-Dimensional Data**

**Ying Liu\***, Columbia University

**Yang Feng**, Columbia University

**Yuanjia Wang**, Columbia University

**Melanie Wall**, Columbia University

Variable selection in high-dimensional setting is a vital tool for analyzing many epidemiological and survey studies that collect large number of predictor variables. A complication that often arises in practice is variables with missing information. There is a wealth of existing literature on variable selection with complete data. Application of these methods to real world studies with missing data requires listwise deletion which we show in this work can substantially reduce prediction power. The problem of information loss is especially severe in high-dimensional data setting where a large proportion of subjects have missing data on at least one variable. A few recent works that address missing data cannot adequately account for large numbers of variables with arbitrary missing patterns. In this work, Multiple Imputation Random Lasso (MIRL) is introduced to select important variables in the presence of

missing data. The proposed method combines penalized regression techniques with multiple imputation and stability selection. Bootstrapping of multiply imputed data is used to construct an importance measure and random lasso regression is applied to simultaneously estimate regression coefficients and perform variable selection. The final estimates are the averages of coefficients across samples and the important variables are chosen and ranked according to stability selection criterion (Meinshausen & Bühlmann, 2010). Extensive simulation studies were conducted to compare the proposed method with several alternatives under various missing data mechanism, proportion of missing, number of noise variables and degrees of correlation among variables. MIRL outperforms other methods in high-dimensional scenarios in terms of prediction error and variable selection, and it shows greater advantages when the correlation is large and missing proportion is large. Lastly, MIRL is shown to outperform other methods when it is applied to a large epidemiological study of Eating and Activity in Teens where 80% of individuals would be eliminated if listwise deletion is used.

e-mail: summeryingl@gmail.com

### **Quantile Regression in the Presence of Monotone Missingness with Sensitivity Analysis**

**Minzhao Liu\***, University of Florida

**Michael Daniels**, University of Texas, Austin

In this paper, we develop methods for longitudinal quantile regression when there is monotone missingness. In particular, we propose pattern mixture models with a constraint that provides an interpretation form for the marginal quantile regression parameters. Our approach allows sensitivity analysis and informative priors which are an essential component in inference for incomplete data. To facilitate computation of the likelihood, we propose a novel way to obtain analytic forms for required integrals. Both frequentist and Bayesian inferences are illustrated. The model is applied to data from a clinical trial on weight management.

e-mail: liuminzhao@ufl.edu

### **Improving the Robustness of Doubly Robust Estimators**

**Peisong Han\***, University of Waterloo

**Lu Wang**, University of Michigan

Doubly robust estimators are widely used in statistical analysis with missing data due to their double protection on estimation consistency. These estimators employ one model for the missingness probability and one model for the data distribution, and are consistent if either model is correctly specified. To improve over double robustness, we propose an estimation procedure that allows multiple models for both the missingness probability and the data distribution, and the resulting estimator is consistent if any one of these

multiple models is correctly specified. We provide an easy-to-implement algorithm for the computation. The proposed estimator, unlike many existing estimators based on the inverse probability weighting method, is also robust against extreme values of the estimated missingness probability.

e-mail: peisonghan@uwaterloo.ca

### **Nonparametric Manova Approaches for Non-Normal Multivariate Outcomes with Missing Values**

**Fanyin He\***, University of Pittsburgh

**Sati Mazumdar**, University of Pittsburgh

**Gong Tang**, University of Pittsburgh

**Stewart J. Anderson**, University of Pittsburgh

Comparisons between groups play a central role in clinical research. As these comparisons often entail many potentially correlated response variables, the classical multivariate general linear model has been accepted as a standard tool. However, automatic deletion of cases with missing values in response variables is a shortcoming of standard software when performing multivariate tests. This paper addresses issues of missing data, focusing on the extension of a non-parametric multivariate Kruskal-Wallis (MKW) test. Our proposed permutation-based method retrieves information in partially observed cases. An R-based program was written to compute p-values of MKW tests for group comparisons. We performed a sensitivity analysis to compare the performance of the standard test utilizing only complete cases and the proposed test. Results show that the proposed method provides higher power level, encompassing a broad spectrum of multivariate effect sizes. The proposed method provides a unique and flexible tool for simultaneously comparing multiple outcomes with missing values across groups. An illustrative example using a small sample dataset from a psychiatric clinical trial is provided.

e-mail: fah11@pitt.edu

### **Model Independent Diagnostic for Multiple Imputations**

**Irina Bondarenko\***, University of Michigan

**Trivellore Raghunathan**, University of Michigan

Multiple imputation is a popular approach to analyze data with missing values. Software packages are widely available to impute missing values. However, diagnostic tools to check validity of imputations are limited, and frequently depend on the imputation model. We developed a set of diagnostic tools that by comparing certain conditional distributions of observed and imputed values allows to access validity of imputations under MAR assumption. This method does not require knowledge of the exact model used for imputations. Proposed diagnostics are useful to identify whether a variable should be included in the imputation process, and recognize a need for non-linear transformation of the variable entering imputation model. The method is illustrated using a data set with large number of variables of different types with varying amount of missing values. Performance of the method is assessed on the simulated dataset.

e-mail: ibond@umich.edu

### **Simple Relaxed Conditional Likelihood**

**John J. Hanfelt**, Emory University

**Lijia Wang\***, Emory University

When the data are sparse but not exceedingly so, we face a tradeoff between bias and precision that makes the usual choice between conducting either a fully unconditional inference or a fully conditional inference unduly restrictive. We propose a method to relax the conditional inference that relies upon commonly available computer outputs. In the rectangular array asymptotic setting, the relaxed conditional maximum likelihood estimator has smaller bias than the unconditional estimator and smaller mean square error than the conditional estimator. We apply our method to a study of neuropsychiatric symptoms in patients with mild cognitive impairment, stratified by several demographic factors.

e-mail: jjajia.best@gmail.com

## **116. TOOLS FOR LONGITUDINAL DATA ANALYSIS**

### **Longitudinal Outcome Evaluation of a Pilot Study of Provider Delivered Care Management**

**Hsiu-Ching Chang\***, BlueCross BlueShield of Michigan

In 2010, the health plan collaborated with five Michigan physician organizations to conduct a pilot Provider Delivered Care Management (PDCM) program. This study presented a longitudinal program evaluation on the effectiveness in controlling the health-related costs between PDCM and the health plan internal care management (HPDCM) program. Target into the care management programs occurs on an open basis, meaning each member would possibly be targeted at different time and drop out as time progresses. Furthermore, members change care relationships with their primary care physicians over time (membership turnover) and a large number of monthly primary outcomes (costs) are zero. To fully model these features, a multilevel zero-inflated correlated model that explains the interrelatedness of the member outcomes and membership turnover was employed. The Differences-in-Differences (DiD) estimator was then used to fairly test the impact of PDCM. For all study subjects, we obtained their monthly medical claims from June 2009 to Dec 2012, including pre target data and post target data. The study population consisted of 505,927 monthly observations representing 15,477 members across 2,442 primary care practices. The evaluation results can be useful for planner working with health care intervention that offers longitudinal service.

e-mail: air.chang@gmail.com

### **Properties and Applications of Multivariate Antedependence Models**

**Chulmin Kim\***, University of West Georgia

Attributes in the biological studies are measured on each subject over repeated time yielding longitudinal data. Modeling covariance has been considered importantly to analyze longitudinal data. The goal of covariance modeling is to obtain as parsimonious the covariance presentation as possible which also fits the data well. Antedependence

(AD) models are generalization of Autoregressive (AR) models that allow the variances and same lag correlations to vary over time so they can be very useful for covariance structure for longitudinal data. We generalize the AD models to multivariate AD (MAD) models and study some of their properties. Examples in biological studies and in sports analysis are given to illustrate the usefulness of those properties of MAD.

e-mail: ckim@westga.edu

### **Ante-dependence Models for Skew Normal Longitudinal Data**

**Shu-Ching Chang\***, University of Iowa  
**Dale Zimmerman**, University of Iowa

This paper explores the problems of fitting ante-dependence (AD) models to continuous non-Gaussian longitudinal data. AD models impose certain conditional independence relations among the measurements within each subject. The models are parsimonious and useful for data exhibiting time-dependent correlations. Since the relation of conditional independence among variables is rather restrictive, we consider AD multivariate skew normal models. The multivariate skew normal distribution not only shares some nice properties with multivariate normal distributions but also allows for any value of skewness. We derive necessary and sufficient conditions on the shape and covariance parameters for multivariate skew normal variables to be AD(p) for some p. Likelihood-based estimation as well as likelihood ratio hypothesis tests for the order of ante-dependence and for zero skewness under the models are presented. Numerical results show that the proposed models may provide reasonable fits to some continuous non-Gaussian longitudinal data set.

e-mail: sherrie.schang@gmail.com

### **Bayesian Shared Parameter Models for Dyadic Longitudinal Data with Intermittent Dropouts**

**Jaeil Ahn\***, Georgetown University  
**Ying Yuan**, University of Texas  
MD Anderson Cancer Center  
**Wenyi Wang**, University of Texas  
MD Anderson Cancer Center

Longitudinal dyadic data with intermittent missing data is difficult to analyze due to the complicated inter-and-outer correlations within and between dyads, as well as non-ignorable intermittent missing data. Based on a shared parameter model, we propose an approach to analyze longitudinal dyadic data with non-ignorable intermittent dropouts. We factorize the joint distribution of the measurement and dropout processes into three components: the marginal distribution of random effects, the conditional distribution of the dropout process allowing intermittent missing given the random effects, and the conditional distribution of the measurement process given the random effects and missing data patterns. The proposed model accounts for the dyadic interplay using the concept

of actor and partner effects as well as dyad-specific random effects. We evaluate the performance of the proposed method using a simulation study, and apply our method to a longitudinal dyadic data set that arose from a metastatic breast cancer study.

e-mail: jaeilahn@hotmail.com

### **An R Package for Sensitivity Analysis on Longitudinal Data with Non-Ignorable Intermittent Missingness**

**Jing Wang\***, The George Washington University  
**Chenguang Wang**, Johns Hopkins University

Missingness in longitudinal studies can be monotone (dropout) or intermittent. Handling non-ignorable intermittent missing data is methodologically and computationally challenging. Thus, the R package *dpmis* is developed to address the non-ignorable intermittent missing data in longitudinal studies. It comprises of functions for sequentially and semi-parametrically modeling the observed data and for sensitivity analysis with sensitivity parameters that identify the full data model based on the observed data model's fitting results (Daniel Scharfstein et. al., working paper). Print, summary, and plot functions are included to summarize parameter estimations and sensitivity analysis results. Finally, an applied example using the functions of *dpmis* in a sleep drug clinical trial data is provided.

e-mail: fionaw0701@gmail.com

### **A Novel Mixture Model Estimates Time to Onset of Disease or Drug Effects and its Association with Key Covariates**

**Mengyuan Xu\***, National Institute of Environmental Health Sciences, National Institutes of Health  
**Yin Yao**, The National Institute of Mental Health, National Institutes of Health

We developed a new statistical tool, the mixture model, to analyze longitudinal data. This method can model onset of disease, or of drug effects. We further investigate the association between the onset of disease or drug effects and contributing factors such as genotype, age, and disease subtype. The longitudinal phenotype data needed to implement this model typically have three parts: the incubation period, disease onset, and the disease manifesting period. The mixture model uses an expectation-maximization (EM)-based approach to estimate the distribution of disease onset. A solution for a weighed logistic regression on the outcome variables was used to estimate onset distribution. A log likelihood ratio test was used to evaluate the significance of the association between onset distribution and genotype or environmental factors. Extensive simulation studies, followed by application to longitudinal data from the Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) study were conducted to evaluate model performance and the model's overall utility.

e-mail: mengyuantracy@hotmail.com

### **A Two-Part Mixture Model for Zero-Inflated Longitudinal Measurements with Heterogeneous Random Effects**

**Huirong Zhu\***, University of Texas Health Science Center at Houston

**Sheng Luo**, University of Texas Health Science Center at Houston

**Stacia M. DeSantis**, University of Texas Health Science Center at Houston

Longitudinal zero-inflated count data arise frequently as outcomes in substance use research when assessing the effects of behavioral and pharmacological interventions. Zero-inflated count models (e.g., zero-inflated Poisson or zero-inflated negative binomial) with random effects have been developed to analyze this type of data; however, there may be considerable intervention- and/or covariate-specific heterogeneity in subject-specific random effects, which are often simply assumed to follow a bivariate normal distribution. We review these commonly used zero-inflated models and extend them to a broader class that accommodates between-subject heterogeneity via modeling the random effects covariance structure as function of clinically relevant covariates. We show via simulation that ignoring intervention and covariate-specific heterogeneity can produce biased estimates for both treatment effects, and random model parameters. Importantly, we show that this may lead to incorrect inference regarding treatment effects, which is rectified by correctly modeling the random effects covariance structure. The methodological development is motivated by and applied to the Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) study, the largest clinical trial of alcohol dependence in United States with 1383 individuals.

e-mail: Huirong.Zhu@uth.tmc.edu

## **117. ANALYSIS OF DATA FROM CLINICAL TRIALS**

### **Finding the Optimal Allocation in Sequential Binary Response Experiments with Two Possibly Correlated Endpoints**

**Lu Wang\***, University of Texas Health Science Center at Houston

**Hongjian Zhu**, University of Texas Health Science Center at Houston

Response-adaptive randomization (RAR) has long been proposed to solve the problem of randomly assigning an inferior treatment to volunteers in clinical trials. The basic idea is to skew allocation probability according to the previous treatment assignments and responses in order to meet certain objectives such as maximize power. Most attention has been focused on clinical trials with only one endpoint. Comparison of two or more samples with multiple endpoints is a common statistical problem in biomedical research. For clinical trials with two possibly correlated endpoints, we establish two formal optimization criteria to

find optimal allocation strategies for two specific questions: (1) How do we maximize power of test of homogeneity with two binary responses? And (2) for fixed power, how do we minimize expected treatment failures? We also assign different weights to the treatment failures to emphasize the endpoints of different importance. We find the optimal allocation and implement it through a randomized sequential design and simulation results will give support to our conclusion.

e-mail: lu.wang@uth.tmc.edu

### **A General Class of Correlation Coefficients Between Binary and Continuous Variables for the $2 \times 2$ Crossover Design**

**Luojun Wang\***, Penn State Hershey College of Medicine  
**Vernon Chinchilli**, Penn State Hershey College of Medicine

In many clinical studies, the Pearson correlation coefficient and the Kendall correlation coefficient are two popular statistics for assessing the correlation between two variables in a bivariate sample. We indicate how both of these statistics are special cases of a general class of correlation statistics that is parameterized by  $\lambda \in [0, 1]$ . The Pearson correlation coefficient is characterized by  $\lambda = 1$  and the Kendall correlation coefficient by  $\lambda = 0$ , so they yield the upper and lower extremes of the class, respectively. In our current work, we describe how this class of correlation statistics can be modified when a bivariate sample consists of (1) an ordinal or binary variable and (2) a continuous variable. We demonstrate this use with two real data examples and we investigate its properties via a small simulation study.

e-mail: vicwong@psu.edu

### **Weighted and Replicated Estimator for Comparing Dynamic Treatment Regimens with a Binary Outcome Using Smart Data: Practical Issues and a Simulations-Based Sample Size Calculator**

**Kelley M. Kidwell\***, University of Michigan  
**Inbal Nahum-Shani**, University of Michigan  
**Connie Kasari**, University of California, Los Angeles  
**Daniel Almirall**, University of Michigan

Personalized medicine is emerging as the main paradigm guiding modern medicine. Chronic diseases and other disorders, including depression, autism, and cancer, often require sequences of treatments that are personalized to the individual. These sequences of treatments, which are tailored to individual responses, characteristics and behaviors, are known as dynamic treatment regimens (DTRs). The sequential multiple assignment randomized trial (SMART), a multi-stage trial, was developed for the purpose of addressing many critical questions in the development of DTRs. Often investigators may be interested in conducting a SMART to compare DTRs (or to find the best DTR) where the overall outcome is binary such as response, success or recurrence rate. This talk discusses the practical implementation and application of a weighted and replicated regression estimator (WRRE), which can be used to simultaneously compare all DTRs embedded within a

SMART, with respect to a binary outcome. The WRRE is easy-to-use with standard software; and it permits investigators to incorporate baseline covariates for improving the statistical efficiency in the comparison of embedded DTRs. We also present software that implements a simulations-based sample size calculator for designing SMART studies with a binary outcome. The methodology is illustrated using data from a SMART in autism.

e-mail: kidwell@umich.edu

### **Marginal Meta Analysis for Combining Randomized Clinical Trials with Rare Binary Outcomes - Reevaluating the Safety Concern of Avandia**

**Yi Huang\***, University of Maryland, Baltimore County  
**Elande Baro**, University of Maryland, Baltimore County  
**Guoxing Soon**, U.S. Food and Drug Administration

Meta-analysis is commonly used in the safety evaluations of medical drugs and medical devices by combining multiple RCT with binary outcomes, especially when the adverse events are rare. Among the commonly used fixed effect meta-analysis estimators, Mantel-Haenszel, and Peto, rare events and the violation of treatment homogeneity assumption typically threaten their validity. Since the low power of rare events often leads to fail to reject the homogeneity tests, combining safety signals across trials with hidden heterogeneity may damage its interpretation as population average effect. Also, various add-hoc continuation correction methods under those estimators may induce bias. Non-collapsibility issue of odds ratio makes its interpretability worse. To improve the interpretability and validity of Meta analysis estimator in those safety studies, here we proposed a relatively more flexible homogeneity assumption and a marginal meta-analysis estimator for combining those RCTs estimating marginal causal effects consistently. Our estimator is particular appealing in the applications with rare events, even though it could be applied generally. Systematic simulation studies shows that the proposed estimator performs reasonably well under various rationales of general safety studies. The method is illustrated by re-evaluating the safety issue on Avandia drug.

e-mail: yihuang@umbc.edu

### **Design Issues and their Effect on Power and Sampling Frequency Requirements for N-of-1 Clinical Trials**

**Yanpin Wang\***, Scripps Health  
**Andrew Viterbi**, Scripps Health  
**Nicholas Schork**, Scripps Health

N-of-1 or single subject clinical trials seek to identify the optimal treatment or intervention strategy for an individual based on the objective, empirical evaluation of the efficacy and side effects profiles of two (or more) treatments provided to that individual. The design of such trials is typically rooted in a simple crossover design with, possibly, multiple evaluation periods. The effect of serial correlation between measurements, the number of evaluation periods, the use of washout periods, heteroscedasticity (i.e., variance

heterogeneity) and carry-over phenomena on the power of such studies is crucially important for putting the yield and feasibility of N-of-1 trial designs into context. We evaluate these effects on the power of different designs for N-of-1 trials using standard likelihood-based theory assuming an autoregressive, lag 1, serial correlation structure between the observations. We show that the influence of serial correlation and heteroscedasticity on power can be substantial but can be mitigated through the use of washout and multiple evaluation periods. We also show that the detection of carry-over effects is heavily influenced by design considerations as well.

e-mail: yanpin@scripps.edu

### **Using Internal Pilots to Design Cluster Randomized Trials with Unequal Cluster Sizes**

**Ashutosh Ranjan\***, University of Alabama, Birmingham  
**Christopher S. Coffey**, University of Iowa  
**Leslie A. McClure**, University of Alabama, Birmingham

Cluster randomized trial, which randomizes groups of individuals to an intervention, are common in health services research where one evaluates improvement in a subject's health by intervening at an organizational level. Many such trials have unequal cluster sizes that may lead to an underpowered study. Researchers may account for this imbalance by performing a cluster-level analysis weighting by cluster size or minimum variance weights. A possibly better approach would be to make the adjustment in the design phase of the study. We use an internal pilot design (two-stage design allowing sample size modification without interim data analysis) to incorporate the cluster size variability at the time of re-estimating the number of clusters and study the characteristics of design in maintaining the type I error rate and power. Using such an approach also allows us to overcome the potential misspecification of intra-cluster correlation at the planning stage. The results from simulations indicate that an internal pilot design performs better when the coefficient of variation of cluster size is greater than 0.30. Further, minimum variance weights perform better than cluster size weights by maintaining a stricter control of type I error but may lead to a slightly conservative estimate of power.

e-mail: ash3@uab.edu

### **Designing Balanced Patient-Specific Treatment Stimuli for Post-Stroke Language Interventions**

**Minming Li\***, University of Massachusetts, Amherst  
**Edward J. Stanek III**, University of Massachusetts, Amherst  
**Jacquie Kurland**, University of Massachusetts, Amherst

Aphasia is a chronic, post-stroke language impairment affecting over one million people currently living in the U.S. Intensive Language Action Therapy (ILAT) has been used to treat chronically aphasic patients with positive results. A randomized controlled trial of a training approach was performed in order to study the effectiveness of the treatment on patients following a stroke. During baseline testing, each patient was asked to name pictures of 218 objects and 100 actions, with the test repeated three times. The results were used to construct sets of pictures for training and intervention. Of the 80 pictures

consistently not-named per patient, half were used in an intensive treatment program. The matched sets of 'to be trained' or 'untrained' pictures were formed by matching psycholinguistic characteristics of the pictured concepts on average. We describe the results of an alternative principal component analysis approach used to match characteristics of pictures in forming the sets for home testing. Finally, we compare the two approaches, and describe some additional insights gained by the principal component analyses.

e-mail: minmingli@schoolph.umass.edu

## 118. HUMAN HEALTH AND ENVIRONMENTAL STATISTICS AT THE U.S. EPA'S OFFICE OF RESEARCH AND DEVELOPMENT

### Exploring Chemically Induced Change in Neuronal Networks

**Diana Hall\***, University of North Carolina, Chapel Hill

Thousands of chemicals are utilized in commerce for which adequate toxicity data is lacking. Unfortunately, standard toxicity assays fail to keep pace with the rate at which new chemicals become commercially available. Therefore, high throughput in-vitro toxicity screening methods are needed. In particular, methods for detecting neurotoxicity using multi-well Micro-Electrode Array (MEA) technology are under development. Electrophysiological data gathered from neuronal cells cultured on MEAs include firing and bursting rates, synchrony and network connectivity. Harnessing appropriate statistical techniques to quantify chemically induced changes in neuronal cultures is key to the success of in-vitro neurotoxicity screening. The purpose of this talk is to present the scope and characteristics of MEA data, as well as methods for its visualization and statistical analysis.

e-mail: dianaransomhall@yahoo.com

### Development and Evaluation of Two Reduced Form Versions of a Deterministic Air Quality Model for Ozone and Particulate Matter

**Kristen M. Foley\***, U.S. Environmental Protection Agency

**Sergey L. Napelenok**, U.S. Environmental Protection Agency

**Sharon B. Phillips**, U.S. Environmental Protection Agency

**Carey Jang**, U.S. Environmental Protection Agency

Due to the computational cost of running regional-scale numerical air quality models, reduced form models (RFM) have been proposed as computationally efficient simulation tools for characterizing the pollutant response to many different types of emissions reductions. The U.S. Environmental Protection Agency has developed two types of reduced form models based upon simulations of the Community Multiscale Air Quality (CMAQ) modeling system. One is based on statistical response surface modeling (RSM) techniques using a multidimensional kriging approach to approximate the nonlinear chemical and physical processes. The second approach is based on sensitivity calculations from the Higher-Order Decoupled Direct Method in 3 dimensions (DDM RFM) and uses a Taylor series approximation for the nonlinear response of the pollutant

concentrations to changes in emissions from specific sectors and locations. Both types of reduced form models are used to estimate the changes in ozone and PM<sub>2.5</sub> across space associated with emissions reductions of NO<sub>x</sub> and SO<sub>2</sub> from power plants and other sectors in the eastern United States. This study provides a direct comparison of the RSM and DDM RFMs in terms of computational cost, model performance against brute force runs, and model response to changes in emissions inputs.

e-mail: foley.kristen@epa.gov

### Fully Bayesian Analysis of High-Throughput Targeted Metabolomics Assays

**James L. Crooks\***, U.S. Environmental Protection Agency

**Denise K. MacMillan**, U.S. Environmental Protection Agency

**Jane E. Gallagher**, U.S. Environmental Protection Agency

High-throughput metabolomic assay that screens hundreds of metabolites have recently become available. Such assays provide a window into understanding changes to biochemical pathways due to chemical exposure or disease. Software included with the kits supports basic analyses, but is less suited to complex experimental designs, and lacks multiple-testing adjustment and other tools for promoting reproducibility. Fortunately, the smaller number of features measured by such assays relative to genomic assays makes a fully Bayesian approach practical on standard computing hardware. We present a novel model for analyzing the output of high-throughput metabolomic assays in a statistically rigorous manner. The model includes flexible spline-based normalization, as well as Bayesian variable selection and multiplicity adjustment. The model can be used with metabolite concentrations treated as either dependent or independent variables. Examples of both are given using data from the Mechanistic Indicators of Childhood Asthma (MICA) study targeting 186 metabolites associated with a variety of metabolic disorders. Disclaimer: This abstract does not necessarily reflect U.S. EPA policy.

e-mail: jimcrooks1975@gmail.com

### Implications of Nonlinear Concentration Response Curve for Ozone Related Mortality on Risk Assessment

**Ana G. Rappold\***, U.S. Environmental Protection Agency

**James L. Crooks**, U.S. Environmental Protection Agency

Predicted rise in temperatures and decrease in global air circulation are expected to have an impact on ground level ozone concentrations. The largest impacts are anticipated to occur in the form of the frequency and duration of summertime regional pollution episodes impacting the higher end of the ozone distribution. These events are likely to induce nonlinear increases in ozone levels. We examine the impact of nonlinear concentration-response curve for ozone related mortality to evaluate the sensitivity of the health burden with respect to changes at the high end of the ozone distribution. A flexible Bayesian hierarchical model

was developed to allow for a nonlinear ozone risk curve with a shape parameter controlling the prior belief about the monotonicity of the risk through the prior distribution of basis function parameters. When probability mass of each basis function parameter is shifted toward the positive half line, the probability that the overall nonlinear trend is monotone then also increases. We used polynomial spline basis functions and their derivatives to evaluate the relative risk of ozone at all percentiles of the distribution. The model is applied to the mortality time series data from major US urban centers between 1987 and 2000 from the US National Morbidity, Mortality, and Air Pollution Study. Disclaimer: This abstract does not necessarily reflect U.S. EPA policy.

e-mail: rappold.ana@epa.gov

### **Modeling the Effect of Temperature on Ozone-Related Mortality**

**Ander Wilson\***, North Carolina State University  
**Ana G. Rappold**, U.S. Environmental Protection Agency  
**Lucas M. Neas**, U.S. Environmental Protection Agency  
**Brian J. Reich**, North Carolina State University

Climate change is expected to alter the distribution of ambient ozone levels and temperatures, which in turn may impact public health. Much research has focused on the effect of short-term ozone exposure on mortality and mobility, but less is known about the joint effects of ozone and temperature. The extent of the health effects of changing ozone levels and temperatures will depend on whether these effects are additive or synergistic. In this paper we propose a spatial, semi-parametric surface model to estimate the joint ozone-temperature risk surfaces in 95 US urban areas. Our methodology restricts the ozone-temperature risk surfaces to be monotone in ozone and allows for both non-additive and non-linear effects of ozone and temperature. We use data from the National Mortality and Morbidity Air Pollution Study (NMMAPS) and show that the proposed model fits the data better than additive linear and non-linear models. We then examine the synergistic effect of ozone and temperature both nationally and locally. [Disclaimer: This work does not necessarily reflect EPA policy.]

e-mail: anderwilson@gmail.com

## **119. POWER ANALYSIS FOR MIXED MODELS: WHERE WE STAND**

### **Quick (But Accurate) Power and Precision Approximation Using Generalized Linear Mixed Model Software**

**Walter W. Stroup\***, University of Nebraska, Lincoln

Littell (1977) and O'Brien (1983) presented methods to compute power using linear model software. Stroup (2002) and Littell, et al. (2006) extended the approach to linear mixed model (LMM) software. Stroup (2013) extended these methods to generalized linear mixed models (GLMMs). These methods create an exemplary data set (O'Brien), and determine power from non-central F distribution with GLMM non-centrality computations. Simulations for LMMs and Poisson and binomial GLMMs show near-exact accuracy

for the exemplary data method. GLMMs with negative binomial data show the exemplary data method is accurate, though not near-exact. GLMM-based methods allow quick comparisons of competing designs (split-plot, clustered or incomplete block designs) that use the same resources but deploy them differently, leading to trade-offs in power and precision. In addition, they allow power assessment in the presence of serial or spatial correlation. They allow the same assessment for non-Gaussian as well as Gaussian data. More importantly, they reveal where non-GLMM-based power analyses can result in catastrophically inappropriate sample size assessment. This presentation includes GLMM power and precision basics and examples of planning multi-level designs with Gaussian and non-Gaussian response variables.

e-mail: wstroup1@unl.edu

### **Sample Size for Fixed Effect Inference in Longitudinal and Multilevel Mixed Models**

**Yueh-Yun Chi\***, University of Florida

Mixed models are commonly used to analyze multilevel (clustered) and longitudinal data when correlations exist among observations. The popularity demands accurate methods to ensure adequate study planning and effective sample size determination. We seek to provide practical methods that allow applying multivariate (e.g., MANOVA) sample size methods to fixed effect inference in mixed models. Recognizing sufficient conditions that lead to inference equivalence allows using existing software. The results apply to a wide range of common mixed model scenarios in longitudinal, split-plot and randomized block designs. Simple approximations allow accounting for modest amounts of missing data. However, finding effective sample size with complex patterns of missing data remains a challenge for future research.

e-mail: yychi@ufl.edu

### **Optimal Combination of Number of Participants and Number of Repeated Measurements in Longitudinal Studies with Time-Varying Exposure**

**Donna Spiegelman\***, Harvard School of Public Health  
**Jose Barrera-Gomez**, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain  
**Xavier Basagana**, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

We explored the values of the number of participants and the number of repeated measurements that maximize the power to detect the hypothesized effect, given the total cost of the study, under two models, one that assumes a transient effect of exposure and one that assumes a cumulative effect. Results were derived for a continuous response variable with damped exponential covariance, and a binary time-varying exposure. Under certain assumptions, we derived simple formulas for the approximate solution to the problem when the response covariance structure is compound symmetry. Results showed the importance of the exposure intraclass correlation in determining the optimal combination of the number of participants and the number of repeated measurements, and therefore the optimized power.

Incorrectly assuming a time-invariant exposure leads to inefficient designs. We also analyzed the sensitivity of results to dropout, mis-specification of the response correlation structure, allowing a time-varying exposure prevalence and potential confounding impact. Software is available to determine the optimal combination of the number of participants and the number of repeated measurements in this setting.

e-mail: stdls@hsph.harvard.edu

## 120. NEW DEVELOPMENTS IN ESTIMATING CAUSAL EFFECTS OF TIME-VARYING TREATMENTS

### Double Robust Estimation Strategies for Longitudinal Censored Data

**Mireille E. Schnitzer\***, Université de Montréal  
**Judith J. Lok**, Harvard School of Public Health

In longitudinal data structures arising from observational or experimental studies, subject attrition is a common occurrence. If the goal is estimation of the parameters of a marginal complete-data model for the outcome, biased inference will result from fitting the model of interest with only uncensored subjects. Inverse probability weighting is a popular method that relies on correct estimation of the probability of censoring to produce consistent estimation, but is an inefficient procedure prone to instability in small samples. We developed a useful implementation of the Bang and Robins (2005) theoretical framework for the estimation of the marginal complete-data model which adjusts for longitudinal missing at random censoring using sequential augmented regressions. In addition, we describe a closely related nonparametric approach using targeted maximum likelihood estimation (van der Laan and Rubin, 2006). These methods are illustrated through an investigation of determining predictors of clinical events in virologically suppressed HIV positive patients in the ALLRT study.

e-mail: mireille.schnitzer@gmail.com

### Nonparametric Smoothing for Causal Inference with Continuous Treatments

**Edward H. Kennedy\***, University of Pennsylvania  
**Marshall M. Joffe**, University of Pennsylvania

We consider the problem of estimating the causal effect of a continuous (or many-valued) treatment, such as dosage or duration, as well as estimating how the effect of such a treatment is modified by covariates. Information about effect modification can be very useful for exploring causal mechanism (i.e., how treatments work) and also for tailoring or personalizing future treatments to subjects who benefit most. Quite a lot has been written about this problem for settings in which the treatment effect is assumed to be parametric or low-dimensional; however, in many cases such assumptions may not be tenable. Our contribution in this work is to develop a doubly robust approach for estimating the effect of a continuous treatment while allowing the causal model to have unrestricted or infinite-dimensional

components. Specifically, we propose a class of kernel-based estimators under a varying coefficient model for the effect of treatment, discuss asymptotic and finite-sample properties of these estimators, and illustrate the methodology by estimating the effect of erythropoietin dosage among patients with kidney disease.

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

### Overcoming Challenges Associated with Artificial Censoring in Structural Nested Failure Time Models

**David M. Vock\***, University of Minnesota

Time-dependent confounding is a common problem when trying to assess the causal effect of a time-varying intervention on a time-to-event outcome. Structural nested failure time models (SNFTM) estimated by G-estimation have been proposed to overcome this problem. An inherent drawback of SNFTM estimated by G-estimation is the use of artificial censoring, a technique where some subjects who are observed to fail are treated as censored. The use of artificial censoring can lead to loss of information and non-continuous and non-smooth estimating functions which can be difficult to solve using traditional root finding algorithms. The presence of highly unusual future covariate and treatment trajectories given past covariate information in particular can lead to substantial loss of information. We show that censoring those patients with unusual covariate and treatment trajectories and then using inverse probability of censoring weighted estimators can lead to improved estimation.

e-mail: vock@umn.edu

### Inference for Causal Effects of Time-Varying Treatment in the Presence of Selective Measurement Error

**Marshall M. Joffe\***, University of Pennsylvania

It is sometimes the case that reliable information on treatment and covariates is present only in a subset of the person-intervals in a given data set. When this subset is identifiable, we can use selective ignorability assumptions for inference about the effect of the treatment using structural nested models and G-estimation. The approach using selective ignorability will provide valid tests of the null hypothesis of no treatment effect, and will reduce sensitivity of inference to the missingness under alternatives compared to standard G-estimation approaches. We use influence functions to provide further theory and justification for our methods. We motivate and illustrate our approach using observational data from the United States Renal Data Systems to estimate the effect of erythropoietin alpha (EPO) on hematocrit levels among hemodialysis patients. In this database, information on EPO is less reliable when subjects are hospitalized than when they are not. We show how our approach leads to less sensitivity of inference.

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

## 121. INSIDE THE BIOSTATISTICAL COLLABORATIVE PROCESS

### Mass Spectrometry-Based Metabolomics to Understand Human Health and Disease

**Andrew Patterson\***, The Pennsylvania State University

Metabolomics, the process of measuring metabolite concentration changes in biofluids such as urine, serum, sweat, or saliva or that of intact cells or tissues, is the terminal part of the omics cascade (genomics, transcriptomics, proteomics) necessary for providing a holistic view (DNA, RNA, protein and metabolites) of human health and disease and represents the final output to cellular responses to endogenous and exogenous stimuli. As a discipline, metabolomics is rapidly establishing itself as an indispensable tool for providing unprecedented views of the metabolic response to disease, drug treatment and chemical toxicants. Metabolomics can also give rise to clues about mechanisms of disease and/or toxicity in addition to discovery of biomarkers. Continued development and refinement of analytical platforms and the application of new bioinformatics strategies has accelerated the widespread use of metabolomics and allows further integration of small molecules into systems biology.

email: adp117@psu.edu

### Kernel Machines for Metabolomics Data Analysis

**Xiang Zhan**, The Pennsylvania State University

**Debashis Ghosh\***, The Pennsylvania State University

**Andrew Patterson**, The Pennsylvania State University

The kernel machine approach has become quite popular for the analysis of various types of genomic data. In this talk, we extend its application to the analysis of data from metabolomics experiments. In particular, we deal with the issue of the metabolites being present on only a subset of samples. This will be done using a novel kernel machine approach that we term the stratified kernel machine. This modeling methodology will allow us to incorporate both presence/absence information as well as quantitative information in a manner that does not require formulation of parametric models. Estimation and inference proceeds along existing lines. We will also discuss extensions of the methodology to multiple metabolites. Simulation studies and real data analyses demonstrate the power of the methodology. Application to data from a liver cancer metabolomics experiment will be given.

email: ghoshd@psu.edu

## 122. NON-PARAMETRIC METHODS

### On Inverse Probability Weighted Estimators in the Presence of Interference

**Lan Liu\***, University of North Carolina, Chapel Hill

**Michael G. Hudgens**, University of North Carolina, Chapel Hill

**Sylvia Becker-Dreps**, University of North Carolina, Chapel Hill

In the last decade a growing body of research has studied how to draw inference about the causal effect of a treatment in the presence of interference. In the observational setting where individuals are not randomly assigned treatment, inverse probability weighted (IPW) estimators have been proposed assuming that individuals can be partitioned into groups such that there is no interference between individuals in different groups. Unfortunately this assumption (sometimes referred to as "partial interference") may not hold in many settings. An additional potential drawback of the existing IPW estimators is for certain scenarios they may have large variance. Therefore, in this paper, we consider some alternative IPW estimators that might be employed when interference is present. Specifically, first we propose generalized IPW estimators and two stabilized IPW estimators that do not require the partial interference; rather any form of interference between individuals is permitted. Second, we derive the asymptotic distribution of the IPW estimators and Hajek estimators and propose consistent variance estimators assuming partial interference. Empirical results are presented demonstrating one of the Hajek estimators can have substantially smaller variance. The different estimators are illustrated using data from a recent study examining the effects of rotavirus vaccination in Nicaragua.

email: liu1815@gmail.com

### Association of Time to Recovery and a Subsequent Depressive or Mania Episode

**Xiaotian Chen\***, University of Pittsburgh

**Yu Cheng**, University of Pittsburgh

We aim to develop inference procedures for the bivariate distribution function of the gap times that are subject to competing-risk censoring. Our work was motivated by a study of bipolar disorder, where the patients who managed to recover from their symptomatic entry may later have developed a new episode of depression or mania. The investigators were interested in quantifying the association between time to recovery and time to recurrence. The estimation of the bivariate distribution of the gap times with independent censoring has been well studied. For example, Wang and Wells (1998), Lin et al. (1999), and Una-Alvarez and Meira-Machado (2008) proposed nonparametric estimators for the bivariate distribution or survival function. However, all these existing methods cannot be applied to failure times that are censored by competing causes. Cheng et al. (2007) developed nonparametric estimators for the bivariate cumulative incidence function (CIF), a quantity that is proper

for gap times from different causes. Therefore, we extend their work to handle successive event times with competing-risk censoring, and propose several nonparametric estimators of the CIF. The performances of the estimators are compared through simulation studies and their practical utility is illustrated in an analysis of the bipolar disorder study.  
email: justinchen87@gmail.com

### **Bayesian Doubly Semiparametric Proportional Hazards Model with Commensurate Priors that Facilitate Borrowing from a Nonexchangeable Data Source**

**Thomas A. Murray\***, University of Minnesota  
**Brian P. Hobbs**, University of Texas  
MD Anderson Cancer Center  
**Bradley P. Carlin**, University of Minnesota

In time-to-event settings, parametric models for the hazard function can inadequately characterize the complexity of the data, thereby missing important features. Furthermore, standard linear regressors may insufficiently characterize the effect of continuous covariates in a proportional hazards model. In this work, we develop a doubly semiparametric proportional hazards model that flexibly models both the baseline log-hazard function and covariate effects with penalized splines. Additionally, we construct priors that can impose shape constraints on covariate effects and that can borrow strength from a supplemental dataset when available and warranted. We then apply our methods to colorectal cancer data from two clinical trials. We borrow strength from the first trial while making flexible inference about the progression free survival curve of the second trial, and also estimate the effect of aspartate transaminase on the overall survival function in the second trial subject to a non-decreasing shape constraint.

email: 8tmurray@gmail.com

### **Bivariate Penalized Splines for Regression**

**Ming-Jun Lai**, University of Georgia  
**Lily Wang\***, University of Georgia

In this work, we are interested in smoothing data over complex irregular boundaries or interior holes. We propose bivariate penalized spline estimators over triangulations using energy functional as the penalty. We establish the consistency and asymptotic normality for the proposed estimators, and study the convergence rates of the estimators. A comparison with thin-plate splines is provided to illustrate some advantages of this spline smoothing approach. The proposed method can be easily applied to various smoothing problems over arbitrary domains, including irregularly shaped domains with irregularly scattered data points.

email: lilywang@uga.edu

### **Signed Rank with Responses Missing at Random** **Huybrechts F. Bindele\***, University of South Alabama

This paper is concerned with the study of the signed-rank estimator of the regression coefficients under the assumption that some responses are missing at random in the regression model. Strong consistency and asymptotic normality of the proposed estimator are established under mild conditions. To demonstrate the performance of the signed-rank estimator, a simulation study is conducted under different settings of errors' distributions, and shows that the proposed estimator is more efficient than the least squares estimator whenever the error distribution is heavy tailed or contaminated. When the errors follow a normal distribution, the simulation experiment shows that the signed-rank estimator is more efficient than its least squares counterpart whenever a large proportion of the responses are missing.

email: hbindele@southalabama.edu

### **Combination of Nonparametric Regression Based Classifiers for Breast Tissue Diagnosis from Raman Spectra**

**Jing Guo\***, University of Kentucky  
**Richard Charnigo**, University of Kentucky  
**Cidambi Srinivasan**, University of Kentucky  
**Ramachandra Dasari**, Massachusetts Institute of Technology  
**Maryann Fitzmaurice**, Case Western Reserve University  
**Abigail Haka**, Cornell University

Breast cancer is the most commonly diagnosed cancer among females in the United States. About one in eight U.S. women will develop invasive breast cancer over the course of her lifetime. Mammography serves only as a screening tool and cannot distinguish between malignant and benign lesions. Thus biopsies are usually performed on patients with breast abnormality found through screenings. However, due to current limitations, the entire diagnosis processes take months and sometimes require repeat, even surgical biopsies. Raman spectroscopy can provide timely feedback on quantitative chemical information about breast tissue, and is a less invasive technique than current diagnosis procedures. A previous study (Haka et al, 2005) showed promising results (sensitivity of 94% and specificity of 96%) by classifying pathologies with the aid of basis Raman spectra acquired from the individual chemical constituents of breast tissue. In this present study, we propose an innovative stacking-type method to combine different nonparametric regression based classifiers that rely on basis spectra, derivatives of basis spectra and distances between a patient's spectra and prototypical spectra for various diagnoses, with the aim to improve the sensitivity and specificity for breast tissue diagnosis.

email: livelyjingle.g@gmail.com

**Gene-Trait Similarity U Test**

**Changshuai Wei\***, Michigan State University  
**Qing Lu**, Michigan State University

Group test has been shown to be suitable for sequencing studies, by jointly testing multiple variants to increase the power and reduce the dimensionality. However, limitations exist for the current group test methods, such as inability to handle multiple traits, dependency on distribution assumption, poor performance for small sample size, and computational inefficiency. To overcome these limitations, in this paper, we proposed a non-parametric gene-trait similarity U test, referred to as GTSU. GTSU first summarizes the genetic information and multiple traits into the genetic similarity and trait similarity, and then combines the two similarities in the framework of weighted U statistic. To evaluate the performance of GTSU, we conducted extensive simulation studies and compare it with three popular parametric methods (SKAT, SKATO and AdjSKAT). GTSU showed significant advantage over the existing methods across all the simulation scenarios (e.g. different sample sizes, different trait distributions). Moreover, GTSU (R coded) is computationally more efficient than the existing methods. We also implemented GTSU in C++ to further improve its computational efficiency. At last, we applied GTSU to the multiple traits analysis of Dallas Heart Study. GTSU can identify association of ANGPTL4 with 5 metabolic related traits, while SKAT, AdjSKAT and SKATO could not.

email: cwei@epi.msu.edu

**123. VARIABLE SUBSET SELECTION****Time-Varying Networks Estimation and Dynamic Model Selection**

**Xinxin Shu\***, University of Illinois, Urbana-Champaign  
**Annie Qu**, University of Illinois, Urbana-Champaign

In many biomedical and social science studies, it is very important to identify and predict the dynamic changes of associations among network data over time. We propose a varying-coefficient model to incorporate time-varying network data, and impose a piecewise penalty function to capture local features of the network associations. The advantages of the proposed approach are that it is nonparametric and therefore flexible in modeling dynamic changes of association for network data problems, and capable of identifying the time regions when dynamic changes of associations occur. To achieve local sparsity of network estimation, we implement a group penalization strategy involving overlapping parameters among different groups. However, this imposes great challenges in the optimization process for handling large-dimensional network data observed at many time points. We develop a fast algorithm, based on the smoothing proximal gradient method, which is computationally efficient and accurate. We illustrate the proposed method through simulation studies

and children's attention deficit hyperactivity disorder fMRI data, and show that the proposed method and algorithm efficiently recover the dynamic network changes over time. The proposed approach works especially well when networks are sparse.

email: shu11@illinois.edu

**Simultaneous Variable Selection for Joint Models of Longitudinal and Survival Outcomes**

**Zangdong He\***, Indiana University School of Medicine and Fairbanks School of Public Health  
**Wanzhu Tu**, Indiana University School of Medicine, Fairbanks School of Public Health and Regenstrief Institute, Inc.  
**Sijian Wang**, University of Wisconsin, Madison  
**Haoda Fu**, Eli Lilly & Company  
**Zhangsheng Yu**, Indiana University School of Medicine and Fairbanks School of Public Health

Joint models with longitudinal and survival outcomes are used with increasing frequency in clinical investigations. Variable selection in joint model settings, however, has not been developed. Herein, we proposed a doubly penalized likelihood (DPL) method with adaptive LASSO and adaptive selection operator (ASO) penalty functions to simultaneously select fixed and random effects. Reparameterization of covariance matrix by Cholesky decomposition was used to ensure its positive-definiteness. Likelihood was penalized by row vector L-2 norm of the decomposed matrix. ASO was then incorporated into the penalized likelihood (PL) to enhance selection accuracy. To correct the estimation bias due to the penalty, we proposed a two-stage procedure to reduce bias. For computation, we approximated the PL by Gaussian quadrature and optimized it by expectation-maximization (EM) algorithm. Simulation showed that the procedure has excellent selection result and that the two-stage estimation substantially reduced the bias. To illustrate, we analyzed a real data set with brain natriuretic peptide as the longitudinal outcomes and death as the survival outcomes from an electronic medical record database.

email: zanghe@iupui.edu

**Local Feature Selection in Varying-Coefficient Models**

**Lan Xue**, Oregon State University  
**Xinxin Shu**, University of Illinois, Urbana-Champaign  
**Peibei Shi\***, University of Illinois, Urbana-Champaign  
**Colin O. Wu**, National Heart, Lung and Blood Institute, National Institutes of Health  
**Annie Qu**, University of Illinois, Urbana-Champaign

We propose new varying-coefficient model selection and estimation methods based on the spline approach which is capable of capturing time-dependent covariate effects. Traditional model selection approach focuses on the entire region for vary-coefficients. However, in many scientific problems, signals associated with relevant predictors are time-dependent and detecting relevant covariate effects in the local region is more scientifically relevant than those of the entire region. Our method conducts local signal detection with a new penalty function by utilizing local-

region information for varying-coefficients. We provide the asymptotic theory of model selection consistency on detecting local signals and establish the optimal convergence rate for the varying-coefficient estimator. Our simulation studies indicate that the proposed model selection incorporating local features outperforms the global feature model selection approaches. The proposed method is also illustrated through a longitudinal growth and health study from National Heart, Lung and Blood Institute. We are able to detect the specific age region in which height covariates is significantly associated with blood pressure for adolescent girls. Our conclusion that the effect of height on blood pressure attenuates as children growing up is also consistent with existing scientific findings.

email: pshi2@illinois.edu

### **Structured Feature Selection for Longitudinal Biomarker Data**

**Anthony V. Pileggi\***, Emory University  
**Brent A. Johnson**, Emory University  
**DuBois Bowman**, Columbia University

Alzheimer's Disease (AD) is a serious mental illness that affects an estimated 5.3 million Americans. Notoriously difficult to diagnose, AD could only be confirmed posthumously until recently. Through the longitudinal Alzheimer's Disease Neuroimaging Initiative (ADNI) study, researchers have identified several AD biomarkers, including: variations of the APOE gene, amyloid-beta concentration in the brain (detected via PET imaging), decreases in brain volume (i.e., tissue degeneration), and cognitive decline as measured by various clinician-administered mental evaluations. We consider a novel approach to small-scale multimodal (e.g., imaging, genetics, demographics, etc.) feature selection in the longitudinal setting, with a flexible predictor-structure. Based on a GEE formulation of the generalized Dantzig selector, our method correctly accounts for within-subject temporal correlation while simultaneously performing variable selection. Further, the flexibility associated with penalizing linear combinations of the regression coefficients (i.e., fusion penalties) can help overcome the lack of a "grouping effect" (e.g., ridge regression, elastic net). We apply our method to a subset of ADNI subjects, and suggest a data-driven strategy for constructing an appropriate penalty matrix.

email: avpileg@emory.edu

### **Parsimonious Covariate and Conditional-Mean Model Selection with Multiple Candidate Predictors**

**Greg DiRienzo\***, State University of New York at Albany

An objective methodology to select a parsimonious set of important covariates and a conditional-mean model when faced with multiple candidate predictor variables is proposed and evaluated. The methodology can be used to fine-tune a well-established covariate screening method such as ISIS/SCAD, or used in its own right within a forward stepwise algorithm. The methods employ (i) bootstrap estimates of prediction error for working models; (ii) an

objective model comparison strategy; and (iii) multiple hypothesis testing that is valid under mis-specified models. The methods are analytically and numerically shown to work well in the sense that the probability that the final model selected contains one or more unimportant variables is asymptotically bounded at a preselected level for arbitrary data-generating distributions. This methodology is illustrated with a dataset consisting of birth certificate information and mortality records from year 2001 from the US-DHHS. It is shown how the instantaneous daily mortality hazard can be modeled flexibly by allowing both the set of important predictors and their effect on the hazard to change arbitrarily thru time.

email: adirienzo@albany.edu

## **124. HIGH DIMENSIONAL DATA IN GENETICS AND GENOMICS**

### **Strategies for Developing Prediction Models from Genome-Wide Association Studies**

**Jincao Wu\***, National Cancer Institute,  
National Institutes of Health  
**Ruth M. Pfeiffer**, National Cancer Institute,  
National Institutes of Health  
**Mitchell H. Gail**, National Cancer Institute,  
National Institutes of Health

Genome-wide association studies (GWASs) have identified hundreds of single nucleotide polymorphisms (SNPs) associated with human diseases. It is not entirely clear how best to incorporate these SNPs in risk prediction models to ensure good predictive performance. We studied various aspects of model building to improve disease prediction, as measured by the area under the receiving operating characteristic curve (AUC), including: (1) How well does a one-phase procedure that selects SNPs and estimates odds ratios on the same data perform? (2) How should training data be allocated between SNP selection (Phase 1) and estimation (Phase 2) in a two-phase procedure? (3) How many SNPs should be selected? (4) Is multivariate estimation preferred to univariate when SNPs are correlated? We investigate model building strategies based on both the simulated and real GWAS data for Crohn's disease. We found that the most critical aspect of prediction model building was initial SNP selection, to which considerable effort should be devoted. The single-phase procedure yields larger AUCs than those in the two-phase model, which yields unbiased estimates. Despite the fact that there are thousands of SNPs with small effect sizes that could potentially improve prediction, one should only select tens or hundreds of SNPs.

email: celiawjc@gmail.com

### **A Penalized Multi-Trait Mixed Model for Association Mapping in Pedigree-Based GWAS**

**Jin Liu\***, University of Illinois, Chicago

**Can Yang**, Yale University

**Xingjie Shi**, Shanghai University of Finance and Economics, China

**Cong Li**, Yale University

**Jian Huang**, University of Iowa

**Hongyu Zhao**, Yale University

**Shuangge Ma**, Yale University

We consider genome-wide association studies (GWAS) of multiple highly correlated quantitative traits from pedigree data in this paper. In the analysis of GWAS, penalized regression has been found to be useful for identifying multiple associated genetic markers where linear mixed models are commonly used to account for complicated dependence among samples. Therefore, penalized linear mixed model is a natural choice that combines the advantages of both modeling approaches for GWAS data analysis. For GWAS of multiple quantitative traits that are highly correlated, analyzing each trait separately is sub-optimal. In this manuscript, we propose a penalized multi-trait mixed model (penalized-MTMM) which simultaneously accounts for both the within-trait and between-trait variance components to jointly analyze multiple traits. Our method not only accounts for the relatedness among study subjects, but also borrows information across traits through joint analysis of these traits using group penalties. We have evaluated the performance of penalized-MTMM in simulation studies and through its application to a GWAS data from the Genetic Analysis Workshop (GAW) 18.

email: gordonliu810822@gmail.com

### **A Mixture of Experts Approach for the Analysis of SNP Data**

**Julia Schiffner\***, Heinrich-Heine-Universitaet Duesseldorf

**Holger Schwender**, Heinrich-Heine-Universitaet Duesseldorf

Single nucleotide polymorphism (SNP) data allow to gain insight into the molecular background of diseases. Often, in order to find sets of SNPs that are associated with a disease a case-control setting is considered. A common approach is to regard the case-control status as binary response and apply a classification method combined with some dimensionality reduction technique, e.g., partial least squares (PLS). Originally, PLS has been developed for quantitative responses but can be applied to classification problems by embedding it into the logistic regression framework. The resulting model is linear in the SNP variables, but often the relationship between SNPs and disease status is assumed to be more complex, with interactions playing an important role. Moreover, the class of diseased individuals may be heterogeneous in the sense that quite different genetic profiles can lead to an increased disease risk. For these reasons we propose a mixture of experts approach where the gating as well as all local expert models are PLS logistic regression models. This mixture model is nonlinear and

takes potential heterogeneity and interactions into account. The performance of the proposed approach is assessed on simulated and real-world data and compared to standard methods.

email: schiffner@math.uni-duesseldorf.de

### **Joint Estimation of Multiple Dependent Gaussian Graphical Models**

**Yuying Xie\***, University of North Carolina, Chapel Hill

**Yufeng Liu**, University of North Carolina, Chapel Hill

**William Valdar**, University of North Carolina, Chapel Hill

Gaussian graphical models are widely used to represent conditional dependence among random variables. In this paper we propose a novel estimator for data arising from a group of GGMs that are themselves dependent. A motivating example is that of modeling gene expression collected on multiple tissues from the same individual: a multivariate outcome that is affected by dependencies defined at the level of both the tissue and the whole body. Existing methods that assume independence among graphs are not applicable in this setting. To estimate multiple dependent graphs, we decompose the problem into two graphical layers: the systemic layer, which is the network affecting all outcomes and thereby induces cross-graph dependency, and the category-specific layer, which represents the graph-specific variation. We propose a new graphical EM technique that estimates the two layers jointly, establish the consistency and selection sparsistency of the proposed estimator, and conform by simulation that the EM method is superior to a naive one-step method. Lastly, we apply our graphical EM technique to mouse genetic data and obtain biologically plausible results.

email: xyy@email.unc.edu

### **Gateau Differential Boosting for Analysis of Gene Effects and Gene-Gene Interaction**

**Kevin He\***, University of Michigan

**Yi Li**, University of Michigan

**Ji Zhu**, University of Michigan

GWAS studies aim to detect the relationship between genetic factors and a given disease. Because many common diseases are affected by certain combinations of only a few genotypes, an urgent need exists for methods for selecting the important genes and gene-gene interactions and building an appropriate model for the effect of the genes on the disease. Penalized method (e.g., group LASSO and grouped LARS) is an established strategy for selection of grouped variable. However, for analysis of gene-gene interaction, the three-level genotype factors and their interactions can create many parameters, and problems with overfitting arise. Penalized method using the entire dictionary may be computationally infeasible. We propose a new boosting method based on Gateau

differential. To allow interaction terms to enter the model more easily, an asymmetric hierarchy is considered such that any factor/interaction of factors in the active set can form a new interaction with any other single factor, even when the single factor is not yet in the active set. To add more flexibility, another option considered is to provide all possible second-order interactions as well as main-effect terms as candidate factors.

email: kevinhe@umich.edu

### **Concordant Integrative Analysis of Multiple Gene Expression Data Sets**

**Fanni Zhang\***, The George Washington University  
**Yinglei Lai**, The George Washington University

The sample size of a microarray experiment is still relatively small due to its relatively high cost. Many gene expression datasets have been collected for the same or similar study. We expect obtain more efficient analysis results if these datasets can be efficiently integrated. It is essential to evaluate whether multiple datasets are genome-wide concordant before the data integration. A mixture model based method has been proposed for the concordant integrative analysis when there are two microarray datasets. It is necessary to extend this approach for multiple datasets. The general statistical framework for our integrative analysis is the partial concordance/discordance (PCD) model. Its related statistical estimation difficulty is that its parameter space increases exponentially with the number of datasets. Since the complete concordance model (CC) and the complete independence (CI) model are two basic statistical frameworks that can be derived from the PCD model, we propose a two-level mixture model to approximate the PCD model. It combines the CC and CI models and its parameter space increases linearly with the number of datasets. We implement an expectation-maximization algorithm for the parameter estimation. Simulation studies are conducted. The method is applied to three microarray gene expression datasets for lung cancer studies.

email: fnzhang@gwmail.gwu.edu

### **D\_CDF Test of Negative Log Transformed P-Values with Application to Genetic Pathway Analysis**

**Hongying Dai\***, Children's Mercy Hospital  
**Richard Charnigo**, University of Kentucky

In genetic pathway analysis and other high dimensional data analysis, thousands and millions of tests could be performed simultaneously. P-values from multiple tests are often presented in a negative log-transformed format. We construct a contaminated exponential mixture model for  $-\ln(p)$  and propose a  $D\_CDF$  test to determine whether some are from tests with underlying effects. By comparing the cumulative distribution functions (CDF) of  $-\ln(p)$  under mixture models, the proposed method can detect the cumulative effect from a number of variants with small effect sizes. Weight functions and truncations can be incorporated to the  $D\_CDF$  test to improve power and better control the correlation among data. By using the modified maximum likelihood estimators (MMLE), the  $D\_CDF$  tests have very tractable limiting distributions under the null hypothesis. A copula based procedure is proposed to address the

correlation issue among p-values. We also develop power and sample size calculation for the  $D\_CDF$  test. The extensive empirical assessments on the correlated data demonstrate that the (weighted and/or c-level truncated)  $D\_CDF$  tests have well controlled Type I error rates and high power for small effect sizes. We applied our method to gene expression data in mice and identified significant pathways related the mouse body weight.

email: hdai@cmh.edu

## **125. TOOLS FOR SURVIVAL ANALYSIS**

### **Robust Prediction of Cumulative Incidence Function Under Non-Proportional Subdistribution Hazards**

**Qing Liu\***, University of Pittsburgh  
**Chung-Chou H Chang**, University of Pittsburgh

Prediction of cause-specific cumulative incidence probabilities is of primary interest to clinical researchers when conducting statistical analysis involving competing risks. The Fine-Gray proportional subdistribution hazards model is widely used to incorporate multiple continuous and discrete prognostic factors, yet it is not applicable in situations which the assumption of proportional subdistribution hazards (PSH) is not valid. In this study, we investigated the properties of the partial likelihood estimators of the Fine-Gray model under nonproportional hazards and proposed a robust risk prediction procedure which is not sensitive to the PSH assumption. The proposed method is easy to implement because it bypasses the modeling of time-varying covariate effects. For analyses of data in which independent censoring is present, we modified our method by incorporating weights to adjust for the possible bias from effects of censoring. We evaluated the prediction performance of our procedures in simulations by estimating and comparing the Brier scores to the nonparametric estimates and the Fine-Gray model with time-covariate interactions. We also demonstrated an application of our procedures in predicting the absolute risk of locoregional recurrence for breast cancer patients, given a set of prognostic factors in which not all of them satisfy the PSH assumption.

email: qil18@pitt.edu

### **Dynamics Model of Diabetes Disease Progression to End-Stage-Renal Disease and Mortality**

**Ying Jiang**, University of Saskatchewan, Canada  
**Nathaniel Osgood**, University of Saskatchewan, Canada  
**Roland Dyck**, University of Saskatchewan, Canada  
**Hyun J. Lim\***, University of Saskatchewan, Canada

The increase in Type 2 diabetes (T2DM) is a growing public health problem worldwide. Public health agencies from many countries are starting to use computational modeling to assist policy makers to better understand the health care system's structure, to project possible future burdens of disease, and to evaluate health care goals and

planning strategies. Using a health administrative database in Saskatchewan, Canada, we built diabetic end-stage-renal disease (ESRD) dynamic models, and validated the models by comparing the model-predicted data with historical data. Before building dynamic models, hazard rates were estimated from competing risks survival analysis, including a Cox cause-specific model, a sub-distribution model, and a piecewise exponential model. Then using the estimated hazard rates, we built two dynamic models: a system dynamics model (SDM) and an agent-based model (ABM). In this example the ABM demonstrated a better match between historical and model predicted data compared to the SDM. Dynamic modeling approaches could be used in the context of public health to understand complex dynamics of health system and evaluate policy interventions.  
email: hyun.lim@usask.ca

### **On the Consistency of Maximum Likelihood Estimators for the Three Parameter Lognormal Distribution**

**HaiYing Wang\***, University of New Hampshire  
**Nancy Flournoy**, University of Missouri, Columbia

The three parameter log-normal distribution is a popular non-regular model, but whether the maximum likelihood estimators for its parameters are consistent or not has been speculated about since the 1960s. We give a rigorous proof for the consistency of the maximum likelihood estimates for the three parameter log-normal distribution which solves a problem that has been recognized and unsolved for 50 years.  
email: whygps@gmail.com

### **Regression When the Predictor may be Censored**

**David Oakes\***, University of Rochester

There is a voluminous literature on regression models where the response variable may be censored. Situations where a predictor variable is subject to right censoring have received much less attention. However the increasing use of biomarkers in clinical studies has focused attention on this area. Unlike in situations with censored response variables, complete case analyses, which simply omit observations with censored values of the predictor, typically do not give biased estimates of the regression parameter. However there may be a substantial loss of efficiency, as the censoring differentially affects the observations with high values of the predictor and therefore high leverage. in the estimation. We briefly survey previous work and explore some approaches to recovery of available information from the censored observations using parametric and nonparametric methods.  
email: oakes@bst.rochester.edu

### **Nonparametric Discrete Survival Function Estimation with Uncertain Endpoints Using an Internal Validation Subsample**

**Jarcy Zee\***, University of Pennsylvania Perelman School of Medicine  
**Sharon X. Xie**, University of Pennsylvania Perelman School of Medicine

When a true survival endpoint cannot be assessed for some subjects, an alternative endpoint that measures the true endpoint with error may be collected, which often occurs when obtaining the true endpoint is too invasive or costly. We develop an estimated likelihood function for the situation where we have both uncertain endpoints for all participants and true endpoints for only a subset of participants. We propose a nonparametric maximum estimated likelihood estimator of the discrete survival function of time to the true endpoint. We show that the proposed estimator is consistent and asymptotically normal. We demonstrate through extensive simulations that the proposed estimator has little bias compared to the naïve Kaplan-Meier survival function estimator, which uses only uncertain endpoints, and more efficient with moderate missingness compared to the complete-case Kaplan-Meier survival function estimator, which uses only available true endpoints. Finally, we illustrate the proposed method with data from an Alzheimer's disease progression study and discuss its broad biomedical applications.

email: jarcyzee@mail.med.upenn.edu

### **A New Flexible Association Measure for Semi-Competing Risks**

**Jing Yang\***, Emory University  
**Limin Peng**, Emory University

In many biomedical studies, it is of interest to assess the dependence between semi-competing risks events (e.g. disease and death), which may help understand the impact of some important disease landmark event. In this work, we propose a new association measure based on the stochastic view of the non-terminating and terminating events involved in the semi-competing risks setting. The proposed measure is well tailored to the data structure of semi-competing risks and can accommodate the exploration of the potential changing pattern of association in the identifiable region of semi-competing risks data. We develop a nonparametric estimation procedure for the proposed association measure by adopting a working quantile residual life model. The estimation method can readily be extended to adjust for confounders for the semi-competing risks dependence of interest. We establish the asymptotic properties of the proposed estimator, and develop inferences accordingly. The proposed methods can be implemented based on standard statistical software without involving smoothing or resampling. Our proposals are illustrated via simulation studies and an application to real data.

email: jyang89@emory.edu

## **Pseudo-Value Approach for Comparing Survival Medians for Dependent Data**

**Kwang Woo Ahn\***, Medical College of Wisconsin  
**Franco Mendolia**, German Aerospace Center, Institute of Aerospace Medicine, Germany

Survival median is commonly used to compare treatment groups in cancer-related research. The current literature focuses on developing tests for independent survival data. However, researchers often encounter dependent survival data such as matched pair data or clustered data. We propose a pseudo-value approach to test the equality of survival medians (univariate analysis) for both independent and dependent survival data. The Type I error and power of the proposed method are examined by a simulation study, in which we examine independent and dependent data. In the simulation study, the proposed method was compared with two existing methods: i) Brookmeyer and Crowley test, *JASA* 1982; 77:433-440; and ii) Rahbar et al. A nonparametric test for equality of survival medians, *Stat. in Med.* 2012; 31:844-854) for independent and dependent survival data. The proposed method and the two existing methods performed equivalently for independent survival data. However, for dependent data, while the two existing methods ignoring dependency were found to be too conservative, the proposed method controlled Type I error satisfactorily and had higher power than the two existing methods. The proposed method is illustrated by a study comparing survival median times for bone marrow transplants.

email: kwoohn@mcw.edu

## **126. META-ANALYSIS**

### **Estimation Of Treatment Effects In Matched-Pair Cluster Randomized Trials By Calibrating Covariate Imbalance Between Clusters**

**Zhenke Wu\***, Johns Hopkins Bloomberg School of Public Health

**Constantine E. Frangakis**, Johns Hopkins Bloomberg School of Public Health

**Thomas A. Louis**, Johns Hopkins Bloomberg School of Public Health and U.S. Census Bureau

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

We address estimation of intervention effects in experimental designs in which (a) interventions are assigned at the cluster level; (b) clusters are selected to form pairs, matched on observed characteristics; and (c) intervention is assigned to one cluster at random within each pair. One goal of policy interest is to estimate the average outcome if all clusters in all pairs are assigned control versus if all clusters in all pairs are assigned to intervention. In such designs, inference that ignores individual level covariates can be imprecise because cluster-level assignment can leave substantial imbalance in the covariate distribution between

experimental arms within each pair. However, most existing methods that adjust for covariates have estimands that are not of policy interest. We propose a methodology that explicitly balances the observed covariates among clusters in a pair, and retains the original estimand of interest. We demonstrate our approach through the evaluation of the Guided Care program.

email: zhwu@jhsph.edu

### **A Unification of Models for Meta-Analysis of Diagnostic Accuracy Studies without a Gold Standard**

**Yulun Liu\***, University of Texas Health Science Center at Houston

**Yong Chen**, University of Texas Health Science Center at Houston

**Haitao Chu**, University of Minnesota

Many statistical methods for meta-analysis of diagnostic accuracy studies have been discussed in the presence of a gold standard. However, in practice, the selected reference test may be imperfect due to measurement error, non-existence, invasive nature, or expensive cost of a gold standard. It has been suggested that treating an imperfect reference test as a gold standard can lead to substantial bias in the estimation of diagnostic test accuracy. Recently, two models have been proposed to account for the imperfect reference test, namely, a multivariate random effects model and a hierarchical summary receiver operating characteristic (HSROC) model. Both models are very flexible in accounting for heterogeneity in tests across studies as well as the dependence between tests. In this paper, we show that these two models, although with different formulations, are closely related and are equivalent in the absence of study-level covariates. Furthermore, we provide the exact relations between the parameters of these two models and assumptions under which two models can be reached to identical submodels. The established relations between two models are empirically validated by a meta-analysis of the performance of the Papanicolaou and histology tests for the detection of cervical neoplasia.

email: Yulun.Liu@uth.tmc.edu

### **Meta-Analysis Methods for Combining Multiple Expression Profiles: Comparisons, Statistical Characterization and an Application Guideline**

**Lun-Ching Chang\***, University of Pittsburgh

**Hui-Min Lin**, University of Pittsburgh

**George C. Tseng**, University of Pittsburgh

As high-throughput genomic technologies become more accurate and affordable, an increasing number of data sets have been accumulated in the public domain and genomic information integration and meta-analysis have become routine in biomedical research. In this paper, we focus on microarray meta-analysis, where multiple microarray studies with relevant biological hypotheses are combined in order to improve candidate marker detection. Many methods have been developed and applied, but their

performance and properties have only been minimally investigated. There is currently no clear conclusion or guideline as to the proper choice of a meta-analysis method given an application; the decision essentially requires both statistical and biological considerations. Here we perform a comprehensive comparative analysis for twelve microarray meta-analysis methods through simulations and six large-scale applications using four statistical evaluation criteria. We elucidate hypothesis settings behind the methods and further apply multi-dimensional scaling (MDS) and an entropy measure to characterize the meta-analysis methods and data structure, respectively. The aggregated results provide an insightful and practical guideline to the choice of the most suitable method in a given application.

email: lunching@gmail.com

### **Investigation on Adaptively Weighted Evidence Aggregation Meta-Analysis Methods**

**Shaowu Tang\***, University of Pittsburgh  
**George C. Tseng**, University of Pittsburgh

In genomic data analysis, it is common to combine the p-values of thousands of genes simultaneously and it is not realistic to assume all the alternative hypotheses are the same for differentially expressed genes. The conventional meta-analysis methods such as Fisher's or Stouffer's methods assign equal weights to each study, which are simple in nature but cannot always achieve high power for a variety of alternative hypotheses. Intuitively more weights should be assigned to the studies with high power to detect the difference between different conditions. In this presentation, a general class of adaptively weighted meta-analysis methods, of which the AW-Fisher's method is a special case, were proposed to find for each gene in which studies it is differentially expressed. For each gene, the best binary 0/1 weights were determined by minimizing the p-value of all possible weighted test statistics. By using the order statistics technique, the searching space for adaptive weights reduces to linear instead of exponential and an integral form was derived to compute the corresponding p-value of the AW-test statistic. Some properties of AW methods such as consistency and asymptotical Bahadur optimality (ABO) have also been investigated. Simulations for  $K = 2$  were performed to verify our findings.

email: sht41@pitt.edu

### **Bayesian Hierarchical Models for Network Meta-Analysis Incorporating Nonignorable Missingness**

**Jing Zhang\***, University of Minnesota  
**Haitao Chu**, University of Minnesota  
**Hwanhee Hong**, University of Minnesota  
**James D. Neaton**, University of Minnesota  
**Guoxing Greg Soon**, U.S. Food and Drug Administration  
**Beth A. Virnig**, University of Minnesota  
**Bradley P. Carlin**, University of Minnesota

Network meta-analysis (NMA) expands the scope of a conventional pairwise meta-analysis to simultaneously handle multiple treatment comparisons, synthesizing both direct and indirect information and thus strengthening inference. Since most of the trials only compare two of the treatments of interest, the typical data in a NMA managed as a trial-by-treatment matrix is extremely sparse, like an incomplete block structure with seriously missing data problems. The most popular methods up to date for NMA belong to the contrast-based (CB) method which requires the data are missing completely at random (MCAR). We show two issues of the CB method: different estimations of odds ratios (ORs) are obtained with different baseline selections; different event rates back translated from ORs are obtained with different reference selections even if the same baseline is used. Zhang et al. (Clinical Trials, 2013) proposed an arm-based (AB) method under missing at random (MAR) mechanism to overcome these issues. However, in randomized clinical trials (RCTs), nonignorable missingness or missing not at random (MNAR) may happen due to deliberate choices. We thus extend the AB method to incorporate MNAR using  $\text{selection}$  models, which is applied to a smoking cessation data, and whose performance is evaluated through simulation studies. In addition we find that the AB method outperforms the CB method in terms of both bias and MSE in various simulation settings.

email: jingzhang2773691@gmail.com

### **Plug-In Tests for Non-Equivalence of Means of Independent Normal Populations**

**Sungwoo Choi\***, University of Maryland Baltimore County  
**Junyong Park**, University of Maryland Baltimore County

We consider the problem of testing the non-equivalence of several independent normal population means. In classical testing problems, it is a well known problem to test the equality of several means using ANOVA. Instead of determining the exact homogeneity or equality, one may consider more flexible homogeneity which allows a predetermined level of difference. This problem is known as testing the non-homogeneity of populations. We propose the plug-in statistics for two different measures of variability: the sum of the absolute deviations and the maximum of the absolute deviations. For each test, the least favorable configuration (LFC) to ensure the maximum rejection probability under the null hypothesis is investigated. Furthermore, we demonstrate the numerical studies based on both simulation and real data to evaluate the plug-in tests and compare these with the range test.

email: schoi8@umbc.edu

## 127. STATISTICAL METHODS FOR HANDLING MISSING DATA

### Censoring Adjustment Methods for Source Apportionment Models

**Jenna R. Krall\***, Johns Hopkins Bloomberg School of Public Health

**Charles H. Simpson**, Havoc Engineering

**Roger D. Peng**, Johns Hopkins Bloomberg School of Public Health

Sources of particulate matter (PM) air pollution are generally inferred from PM chemical constituent concentrations using source apportionment models. Concentrations of PM constituents are often censored below minimum detection limits and most source apportionment models cannot handle missing data. It is not known how source estimation is affected by the method chosen to impute censored data and a comprehensive assessment would help guide treatment of missing constituent data. We demonstrated that standard methods used to adjust censored data cause bias in source estimation and we developed a likelihood-based approach that addresses these limitations. We compared our likelihood-based method to standard censoring adjustment methods when estimating sources in New York City. We found source means and standard deviations differed by censoring adjustment method. We provide general guidance for adjusting censored PM constituent data in source apportionment, which is necessary for estimation of PM sources and their subsequent health effects.

email: jkrall@jhsph.edu

### A Hot Deck Imputation Procedure for Multiply Imputing Nonignorable Missing Data: The Proxy Pattern-Mixture Hot Deck

**Danielle M. Sullivan\***, The Ohio State University

**Rebecca R. Andridge**, The Ohio State University

Hot deck imputation is a common method for handling item nonresponse in surveys, but most implementations assume data are missing at random (MAR). We combine the distance-based donor selection method of Siddique and Belin (2008) with the proxy pattern-mixture (PPM) model (Andridge and Little 2011) to create a hot deck-based method for imputing a continuous partially missing outcome  $Y$  that does not assume data are MAR. The PPM model reduces a set of fully observed covariates to a single variable  $X$  called the proxy, estimated from a regression analysis of respondent data. We use  $X$  to define distances between donors and donees under different missingness assumptions. Missingness in  $Y$  is assumed to depend on  $X+LY$ , where  $L$  is a sensitivity parameter that controls the missingness mechanism. This allows for a sensitivity analysis comparing MAR ( $L=0$ ) to various levels of MNAR ( $L=1$ ,  $L=\infty$ ). In addition, we propose two donor quality metrics, since for strong MNAR mechanisms there may not be close donors available. We explore bias and coverage of the PPM hot deck through simulations and apply the method to data from the Ohio Medicaid Assessment Survey.

email: sullivan.467@osu.edu

### Analysis of Incomplete Derived Responses: Multiple Imputation for Body Mass Index Data

**Jiwei Zhao\***, University of Waterloo

**Richard Cook**, University of Waterloo

**Changbao Wu**, University of Waterloo

The body mass index (BMI) is defined as an individual's weight (kg) divided by their height squared ( $m^2$ ). It serves as useful measure of health and risk for many diseases. It is routinely categorized as normal ( $BMI \leq 25$ ), overweight ( $25 < BMI \leq 30$ ) and obese ( $BMI > 30$ ). This variable may be impossible to compute because of missing height (Pattern 1), weight (Pattern 2), or both (Pattern 3). Analysis based on the categorical BMI variable only, individuals with missing data in Patterns 1 and 2 are ignored, and efficiency may be lost. A challenge is how to extract the information contained in partially observed weight/height data from Patterns 1 and 2. We propose a multiple imputation (MI) approach to address this problem. Our response model is a polychotomous logistic regression model based on the categorized BMI variable, and a bivariate imputation model is adopted involving weight and height. We derive the asymptotic variance of our proposed MI-based estimator for the case that the imputation model is correctly specified. The efficiency gain is explored by the asymptotic relative efficiency (ARE) comparing our proposed estimator and the one based only on subjects with complete data. Finally, we illustrate various methods through application to a recent youth smoking study.

email: jiwei2012zhao@gmail.com

### Longitudinal Latent Variable Models Given Incompletely Observed Biomarkers and Covariates

**Chunfeng Ren\***, Virginia Commonwealth University

**Yongyun Shin**, Virginia Commonwealth University

In this paper, we analyze a two-level latent variable model for longitudinal data from the National Growth of Health Study where surrogate outcomes or biomarkers and covariates are subject to missingness at any of the levels. A conventional method for efficient handling of missing data is to reexpress the desired model as a joint distribution of variables, including the biomarkers, which are subject to missingness conditional on all of the covariates that are completely observed, and estimate the joint model by maximum likelihood, which is then transformed to the desired model. The joint model, however, identifies more parameters than desired, in general. We show that the over-identified joint model produces biased estimation of the latent variable model, and describe how to impose constraints on the joint model so that it has a one-to-one correspondence with the desired model for unbiased estimation. The constrained joint model handles missing data efficiently under the assumption of ignorable missing data and is estimated by a modified version of the expectation-maximization (EM) algorithm.

email: renc@vcu.edu

### Clustering Incomplete Data Using Normal Mixture Models

**Chantal Larose\***, University of Connecticut  
**Dipak Dey**, University of Connecticut  
**Ofer Harel**, University of Connecticut

Model-based clustering provides a flexible, easily-describable framework to describe how data groups together. Normal mixture models build that framework under the assumption that the data comes from a Normal distribution. Existing methods for Normal mixture model clustering require complete data. Multiple imputation is a simulation-based approach to analyzing missing data which bypasses many of the disadvantages present in other methods for handling missing data. We present a methodology for clustering incomplete data using Normal mixture models. Our method combines the disparate results produced by multiple imputation into a single cluster solution. We illustrate our method with a simulation study using Fisher's Iris dataset, then demonstrate the utility of the method on yeast gene expression data.

email: chantal.larose@uconn.edu

### Causal Inference in Longitudinal Studies with Dropout and Truncation by Death

**Michelle Shardell\***, University of Maryland  
**Gregory Hicks**, University of Delaware  
**Luigi Ferrucci**, National Institute on Aging,  
 National Institutes of Health

Motivated by aging research, we propose an estimator of the effect of a time-varying exposure on an outcome in longitudinal studies with dropout and truncation by death. We also show that the estimator's causal interpretation is a function of principal strata effects. Simulations show that the estimator is unbiased when weights or outcome regressions are correct. We apply the method to a longitudinal study of vitamin D and gait speed among older adults.

email: mshardel@epi.umaryland.edu

### The Effect of Imputing a Complex Outcome on the Rejection Rate of Pearson's Chi-Square Test of Independence and a Permutation-Based Correction Factor

**Megan J. Olson Hunt\***, University of Pittsburgh  
**Gong Tang**, University of Pittsburgh

There exist contexts where it is meaningful to combine two binary outcomes,  $A$  and  $B$ , into a third variable,  $Y$ , known as a complex outcome. Ultimately, Pearson's test for independence between  $Y$  and another variable,  $T$  (treatment, e.g.), is of interest. Consider the case where  $A$  is subject to missingness and subsequently  $Y$  is as well. When data are missing completely at random, there exist two valid imputation procedures: imputing  $Y$  conditionally on  $T$ , denoted  $Y | T$ , and imputing  $A$  conditionally on  $B$  and  $T$ , denoted  $A | (B, T)$ . Simulation confirms single imputation

based on  $A | (B, T)$  is more efficient than based on  $Y | T$ . Under the null, imputation leads to a higher rejection rate in Pearson's test than the nominal alpha, and thus correction is required. Because a closed-form solution is not clearly tractable, a permutation-based method that determines the corrected critical value by estimating the empirical distribution of the test statistic under the null of independence between  $(A, B)$  and  $T$  is proposed. Simulation confirms this approach yields the nominal alpha level under the null, and additionally that  $A | (B, T)$  results in a test with higher power than  $Y | T$ .

email: Megan.Olson.Hunt@gmail.com

## 128. LONGITUDINAL DATA ANALYSIS

### Generalized p-Values for Testing Zero-Variance Components in Linear Mixed-Effects Models

**Haiyan Su\***, Montclair State University  
**Xinmin Li**, ShanDong University of Technology  
**Hua Liang**, The George Washington University  
**Hulin Wu**, University of Rochester

Linear mixed-effects models are widely used in analysis of longitudinal data. However, testing for zero-variance components of random effects has not been well resolved in statistical literature, although some likelihood-based procedures have been proposed and studied. In this article, we propose a generalized p-value based method in coupling with fiducial inference to tackle this problem. The proposed method is also applied to test linearity of the nonparametric functions in additive models. We provide theoretical justifications and develop an implementation algorithm for the proposed method. We evaluate its finite-sample performance and compare it with that of the restricted likelihood ratio test via simulation experiments. An application of real study using the proposed method is also provided.

email: suh@mail.montclair.edu

### Sufficient Dimension Reduction for Longitudinal Data

**Xuan Bi\***, University of Illinois, Urbana-Champaign  
**Annie Qu**, University of Illinois, Urbana-Champaign

Correlation structure contains important information about longitudinal data. Existing sufficient dimension reduction approaches assuming independence may lead to substantial loss of efficiency. We apply the quadratic inference function to incorporate the correlation information and apply the transformation method to recover the central subspace. The proposed estimators are shown to be consistent and more efficient than the ones assuming independence. In addition, the estimated central subspace is also efficient when the correlation information is taken into account. We compare the proposed method with other dimension reduction approaches through simulation studies, and apply this new approach to longitudinal data for an environmental health study.

email: xuanbi2@illinois.edu

### **Real Time Monitoring of Progression Towards Renal Failure in Primary Care Patients**

**Peter J. Diggle**, Lancaster University, United Kingdom and University of Liverpool, United Kingdom

**Ines Sousa**, University of Minho, Portugal

**Ozgur Asar\***, Lancaster University, United Kingdom

Renal disease can be asymptomatic for many years, but early detection and treatment can slow the rate of progression towards renal failure. Analysis of routinely collected biomarkers of kidney function can assist early detection. Current UK guidelines use the estimated glomerular filtration rate (eGFR) as an overall measure of kidney function and recommend that a patient who are losing kidney function at a rate of at least 5% per year, as measured by their eGFR, should be referred to a specialist treatment centre. In this study, we consider use of dynamic linear modelling to obtain the predictive distribution of the underlying rate of change in kidney function. Our model assumes that kidney function within any one patient evolves according to a continuous-time, non-stationary stochastic process, accommodates between-patient heterogeneity by a combination of baseline covariates and a random patient-specific intercept, and treats eGFR as a noisy measurement of a patient's underlying kidney function. Our overall aim is to incorporate model-based predictions into a real-time surveillance system that can alert general practitioners to the possible need for the referral of their patient to a specialist treatment centre.

email: o.asar@lancaster.ac.uk

### **AR(1) Latent Class Models for Longitudinal Count Data**

**Nicholas Henderson\***, University of Wisconsin, Madison

**Paul Rathouz**, University of Wisconsin, Madison

In a variety of applications involving longitudinal or repeated-measurements data, it is desired to uncover natural groupings or clusters which exist among study subjects. We propose a method to address this goal when the data in question are counts. Under our approach, we assume that the subject-specific observations are generated from a first-order autoregressive process which is appropriate for counts. One advantage of this is that the marginal distribution of the response can be expressed in closed form which allows us to bypass common computational issues associated with random effects models. To further improve computational efficiency, we outline a quasi-EM procedure for estimating the model parameters where, within each EM iteration, the maximization step is approximated by solving an appropriately chosen set of estimating equations. We explore the effectiveness of the procedure through simulations based on a four-class model placing a special emphasis on posterior classification. Finally, we illustrate our approach using data on children of subjects from the National Longitudinal Study of Youth.

email: nhenders@stat.wisc.edu

### **Time-Varying Coefficient Models to Identify and Model Time-Clusters in Recurrent Event Data**

**Xiaoxue Li\***, University of Pittsburgh

**Stewart J. Anderson**, University of Pittsburgh

**Saul Shiffman**, University of Pittsburgh

In Ecological momentary assessment (EMA) studies, subjects are measured instantaneously and repeatedly over time in their everyday life. We investigated one such study involving the instantaneous assessments of location, mood, activities, food and drink at times smoking episodes occurred in subjects who were either daily smokers (DS) or intermittent smokers (ITS). We noticed that reported cigarettes tend to be clustered in time for some ITS. We assume that these time-clusters are generated by the combination of a stochastic process and a Poisson Point process, which model the underlying smoking behaviors. We use inter-events gap time intervals as outcomes and propose a time-varying coefficient model to identify time-clusters and model their impact on subjects' future smoking behavior at the same time while simultaneously adjusting for covariates that may affect outcome. Hence, differences in smoking patterns between ITS and DS are tested by adding group as a fixed covariate.

email: xil55@pitt.edu

### **Identifying Multiple Change-Points in a Linear Mixed Effects Model**

**Yinglei Lai\***, The George Washington University

**Paul S. Albert**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Although change-point analysis methods for longitudinal data have been developed, it is often of interest to detect multiple change-points in longitudinal data. In this paper, we propose a linear mixed effects modeling framework for identifying multiple change-points in longitudinal Gaussian data. Specifically, we develop a novel statistical and computational framework that integrates the Expectation-Maximization (E-M) and the Dynamic Programming (DP) algorithms. We conduct a comprehensive simulation study to demonstrate the performance of our method. Our method is illustrated with an analysis of data from a trial evaluating a behavioral intervention for the control of type I diabetes in adolescents with HbA1c as the longitudinal response variable.

email: ylai@gwu.edu

### Regression Analysis of Mixed Recurrent-Event and Panel-Count Data

**Liang Zhu\***, St. Jude Children's Research Hospital

**Xingwei Tong**, Beijing Normal University

**Jianguo Sun**, University of Missouri, Columbia

**Kumar Srivastava**, St. Jude Children's Research Hospital

**Wendy Leisenring**, Fred Hutchinson Cancer Research Center

**Leslie Robinson**, St. Jude Children's Research Hospital

In event history studies concerning recurrent events, two types of data have been extensively discussed. One is recurrent-event data, and the other is panel-count data. In the former case, all study subjects are monitored continuously; thus, complete information is available for the underlying recurrent-event processes of interest. In the latter case, study subjects are monitored periodically; thus, only incomplete information is available for the processes of interest. In reality, however, a third type of data could occur in which some study subjects are monitored continuously, but others are monitored periodically. When this occurs, we have mixed recurrent-event and panel-count data. This paper discusses regression analysis of such mixed data and presents two estimation procedures for the problem. One is a maximum likelihood estimation procedure, and the other is an estimating equation procedure. The asymptotic properties of both resulting estimators of regression parameters are established. Also, the methods are applied to a set of mixed recurrent-event and panel-count data that arose from a Childhood Cancer Survivor Study and motivated this investigation.

email: liang.zhu@stjude.org

## 129. PREDICTION AND PROGNOSTIC MODELING

### Predicting Probabilities of Competing Risk Outcomes Under Informative Censoring, with Application to Safety and Efficacy of Initial Art in HIV-Positive Patients

**Judith J. Lok\***, Harvard School of Public Health

**Michael D. Hughes**, Harvard School of Public Health

Two typical outcomes in HIV clinical trials are virologic failure and treatment limiting adverse events. Because many treatments are successful in limiting the probability of virologic failure, preventing treatment limiting adverse events on initial treatment may be the next target. We are interested in predicting the competing risks virologic failure and treatment limiting adverse events on initial treatment. Treatment discontinuation for pregnancy, disallowed medications, clinical events, and death are additional competing risks. When patients are lost to follow-up we censor them, which possibly leads to informative censoring. To account for informative censoring, we use a discrete version of Inverse Probability of Censoring Weighting. Most analyses to predict a patient's competing risks outcomes based on covariates have focused on the cause specific hazard. We focus on cumulative incidence functions, the probabilities of each of the outcomes over time, because of their clinical interest. To predict these from baseline characteristics, we use a logit model. We illustrate our

method by investigating how the probabilities of virologic failure and treatment limiting adverse events on initial treatment during the first two years in the ACTG5095 study depend on baseline characteristics, comparing a 3-drug and a 4-drug Efavirenz-containing regimen.

email: jlok@hsph.harvard.edu

### The Optimality of a Pseudo-Likelihood Approach to Bayesian Classification

**Josephine K. Asafu-Adjei\***, Harvard School of Public Health

**Rebecca A. Betensky**, Harvard School of Public Health

Despite the relatively high accuracy of the naive Bayes classifier, which assumes conditional independence among outcomes, there may still be several instances where this classifier does not have the same classification performance as the traditional Bayes classifier, which utilizes the joint distribution of the examined outcomes. Therefore, we introduce a 'pseudo-likelihood' classifier that instead takes into account all bivariate relationships that may exist among the different outcomes. In this paper, we first describe the necessary and sufficient conditions under which the pseudo-likelihood classifier is optimal, i.e., has the same classification performance as the traditional Bayes classifier. We then discuss sufficient conditions for which the pseudo-likelihood classifier, and not naive Bayes, is optimal in the case of normal data. Through simulation studies involving normal data, we show that our proposed classifier yields more accurate results than the naive Bayes classifier for various sample sizes, mean values, and correlation structures. We then demonstrate the increase in classification accuracy of our approach relative to naive Bayes in the context of high dimensional biomedical data studies.

email: jasafuad@hsph.harvard.edu

### Local Likelihood-Based Estimation for Quantile Classification in Binary Regression Models

**John D. Rice\***, University of Michigan

**Jeremy M. G. Taylor**, University of Michigan

One common application of binary response regression methods is classification based on an arbitrary probability threshold dictated by the particular application rather than by statistical considerations. Since this is given to us a priori, it is sensible to incorporate the threshold into our estimation procedure. Specifically, for the linear logistic model, we solve a set of locally weighted score equations, using a kernel-like weight function centered at the threshold. The bandwidth for the weight function is selected by cross validation of a novel hybrid loss function that combines classification error and Kullback-Leibler divergence. Although inspired by local likelihood methodology, this work shares more in common with robust estimation, but differs from previous approaches in both areas in its focus on prediction, specifically classification into high- and low-risk groups. Simulation results are given showing the reduction in error rates that can be obtained with this method when compared with maximum likelihood estimation, especially under certain forms of model misspecification. Analysis of a melanoma data set is presented to illustrate the method in practice.

email: jdrice@umich.edu

## Combination of Longitudinal Biomarkers in Predicting Binary Events with Application in a Fetal Growth Study

**Danping Liu\***, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

**Paul S. Albert**, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

In disease screening, the combination of multiple biomarkers often substantially improves the diagnostic accuracy over a single marker. This is particularly true for longitudinal biomarkers where individual trajectory may improve the diagnosis. We propose a pattern mixture model (PMM) framework to predict a binary disease status from a longitudinal sequence of biomarkers. The marker distribution given the disease status is estimated from a linear mixed effects model. A likelihood ratio statistic is computed as the combination rule, which is optimal in the sense of the maximum ROC curve under the correctly specified mixed effects model. The individual disease risk score is then estimated by Bayes' theorem, and we derive the analytical form of the 95% confidence interval. We show that this PMM is an approximation to the shared random effects (SRE) model proposed by Albert (2012). Further, with extensive simulation studies, we found that the PMM is more robust than the SRE under wide classes of models. This new PPM approach for combining biomarkers is motivated by and applied to a fetal growth study, where the interest is in predicting macrosomia using longitudinal ultrasound measurements.

email: danping.liu@nih.gov

## Predictive Accuracy of Time-Dependent Markers for Survival Outcomes

**Li Chen\***, University of Kentucky

**Donglin Zeng**, University of North Carolina, Chapel Hill

**Danyu Lin**, University of North Carolina, Chapel Hill

In clinical cohort studies, potentially censored times to a certain event, such as death or disease progression, and patient characteristics at the time of diagnosis or the time of inclusion in the study (baseline) are often recorded. Serial measurements on clinical markers during follow up may also be recorded for monitoring purpose. Recently there are increasing interests in incorporating these serial measurements of markers for the prediction of future survival outcomes and assessing the predictive accuracy of these time-dependent markers. In this paper we propose a new graphical measure, the negative predictive function, to quantify the predictive accuracy of time-dependent markers for survival outcomes. This new measure has direct relevance to patient survival probabilities and thus has direct clinical utility. We construct a nonparametric estimator for the proposed function, allowing censoring to depend on markers, and adopt the bootstrap method to obtain the asymptotic variances. Simulation studies demonstrate that the proposed method performs well in practical situations. A medical study is presented.

email: lichenuky@uky.edu

## An Investigation of the Assumptions of the Current Status Model

**Jian-Lun Xu\***, National Cancer Institute, National Institutes of Health

Current status data usually arise in studies where the target measurement  $X$  is the time of occurrence of some event of interest, but observation is limited to indicator  $\delta = I(X \leq T)$  of whether or not the event has occurred at the examination time  $T$ . Although estimation of the distribution of  $X$  using the current status data  $(T_i, \delta_i)_{i=1}^n$  is well understood, techniques for checking the independence assumption between  $X$  and  $T$  or constant sum condition  $P(X \leq t, T = t) = P(X \leq t)$  are not well developed. In this paper we show that  $Q(t) = P(\delta = 1, T > t)$  is nondecreasing in  $t$  when  $X$  and  $T$  satisfy the constant sum condition. We also prove that the non-decreasing property of  $Q(t)$  is equivalent to the conditional odds ratio or the conditional cross-product ratio  $OR(t) = \frac{P_{11}(t)P_{00}(t)}{P_{10}(t)P_{01}(t)} \geq 1$ , where  $P_{ij}(t) = P(I(T > t) = i, \delta = j, T \geq t)$  for  $i, j = 0, 1$ . We use this property to develop a method of testing to determine when the constant sum condition does not hold for a data set. The result is also illustrated by examples.

email: xujia@mail.nih.gov

## 130. NEW METHODS FOR GWAS

### Lassot: A Hybrid of LASSO and T-Regularization for Penalized Regression and Applications to Genomic Selection

**Long Qu\***, Wright State University

High-dimensional regression is often applied in the genomic selection problem where the genotypes of a large number of genetic markers are used to predict the phenotypes of offspring and to identify causal genomic regions for further investigation. The so-called BayesA method places scaled-t priors on regression coefficients to achieve relatively good predictive performance, but it lacks sparsity in regression parameter estimates and hinders biological interpretation. The lasso method places double-exponential priors on regression coefficients to simultaneously shrink estimates and select interesting markers, but it introduces non-diminishing bias that is unfavorable in the presence of strong marker effects. In this study, we introduce a hybrid method, lassot, that shares the strengths of both the lasso and the BayesA by imposing the lasso penalty for small coefficients to introduce sparsity and switching to the BayesA penalty for large coefficients to reduce estimation bias. Compared to existing similar attempts, lassot enjoys the interpretation of placing a proper prior that has a double-exponential center and scaled-t tails. An efficient coordinate descent algorithm is developed for estimation. Application of the lassot method is illustrated through the analysis of a real dataset.

email: long.qu@wright.edu

### SHAVE: Shrinkage Estimator Measured for Multiple Visits Increases Power in GWAS of Quantitative Traits

**Osorio D. Meirelles\***, National Institute on Aging, National Institutes of Health

**Jun Ding**, National Institute on Aging, National Institutes of Health

**Toshiko Tanaka**, National Institute on Aging, National Institutes of Health

**Serena Sanna**, Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy

**Hsih-Te Yang**, Taiwan Food and Drug Administration

**Dawood B. Dudekula**, National Institute on Aging, National Institutes of Health

**Francesco Cucca**, Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy

**Luigi Ferrucci**, National Institute on Aging, National Institutes of Health

**Goncalo Abecasis**, University of Michigan

**David Schlessinger**, National Institute on Aging, National Institutes of Health

Measurement error and biological variability generate distortions in quantitative phenotypic data. In longitudinal studies with repeated measurements, the multiple measurements provide a route to reduce noise and correspondingly increase the strength of signals in genome-wide association studies (GWAS). To optimize noise correction, we have developed Shrunken Average (SHAVE), an approach using a Bayesian Shrinkage estimator. This estimator uses regression toward the mean for every individual as a function of (1) their average across visits; (2) their number of visits; and (3) the correlation between visits. Computer simulations support an increase in power, with results very similar to those expected by the assumptions of the model. The method was applied to a real data set for 14 anthropometric traits in <6000 individuals enrolled in the SardiNIA project, with up to three visits (measurements) for each participant. Results show that additional measurements have a large impact on the strength of GWAS signals, especially when participants have different number of visits, with SHAVE showing a clear increase in power relative to single visits. In addition, we have derived a relation to assess the improvement in power as a function of number of visits and correlation between visits.

email: osoriomeirelles@gmail.com

### Secondary Trait Analysis for Case-Control Association Studies in the Presence of Covariates

**Godwin Yung\***, Harvard University

**Xihong Lin**, Harvard University

Case-control genome-wide association studies (GWAS) often collect from their subjects extensive information on secondary phenotypes. Reusing the data and studying the association between genes and secondary phenotypes provide an attractive and cost effective approach that can lead to discovery of new genetic associations and provide further clues to causal pathways. For that purpose, a number of approaches have been considered, including naive analyses that ignore ascertainment or stratify on case-control status, and more complex methods based on weighted or semi-parametric likelihoods. However, justification for many of these methods relies on the assumption of no covariates or a logistic penetrance model for the primary disease. Of course, neither are necessarily true in practice. In this paper we carry out an extensive investigation of robustness for naive methods in the presence of covariates and model misspecification. We show that, contrary to popular belief, naive analyses can still have improper type I error rates when the disease is rare or when the association between the primary and secondary trait does not depend on the SNP genotype. Our principles are justified theoretically and via simulations, and illustrated by a lung cancer case-control GWAS.

email: ghy635@mail.harvard.edu

### Penalized Multi-Marker Versus Single-Marker Regression Methods for Genome-Wide Association Studies of Quantitative Traits

**Hui Yi\***, Virginia Tech

**Netsanet Imam**, Virginia Tech

**Ina Hoeschele**, Virginia Tech

The goal of genome-wide association studies (GWAS) is to select a subset of DNA markers, typically single nucleotide polymorphisms (SNPs), which affect a biomedical trait of interest. GWAS, from a statistical point of view, is a large-scale variable selection problem with millions of common SNPs available in current human studies. Practitioners of GWAS predominantly employ single marker analysis methods while a few penalized regression approaches have been considered. Single marker analysis employs an incorrect statistical model and is affected both by a loss of power and an unnecessarily high rate of false positives. Here we present a comprehensive comparison of penalized regression methods with single marker regression on realistically simulated GWAS data on single and multiple chromosomes, and for a single, continuous response. The selection of tuning parameter values is based on control of the False Discovery Rate. We investigate the value of incorporating Linkage Disequilibrium into penalized regression. We provide guidelines for the use of penalized regression in GWAS. While we focus on a single response here, our results have relevance for the more challenging problem of performing GWAS on high-dimensional, correlated responses.

email: yihuiviv@gmail.com

### **Association Studies with Imputed SNPS Using Expectation-Maximization-Likelihood-Ratio Test**

**Kuan-Chieh Huang\***, University of North Carolina, Chapel Hill

**Yun Li**, University of North Carolina, Chapel Hill

Genotype imputation has become standard practice in modern genetic studies. As sequencing-based reference panel continues to grow, we have increasingly more well imputed markers but at the same time also more markers with relatively low imputation quality. Here, we propose new approaches that attempt to more elegantly incorporate uncertainty when analyzing imputed genotypes. We consider two scenarios: 1) when posterior probabilities of genotypes are estimated or 2) when only imputed dosages are available. When posterior probabilities are estimated, we developed an expectation-maximization (EM) likelihood-ratio test (LRT) for association studies. When only dosages are observed, instead of modeling dosages directly, we first sample the probabilities of all three possible genotypes and then apply the EM-LRT on the sampled probabilities. Extensive simulations have shown that type I error rates of the EM-LRT approaches under both scenarios are protected. Regarding power, EM-LRT-Prob offers enhanced statistical power across the whole spectrum of imputation quality and EM-LRT-Dose has similar power performance as EM-LRT-Prob and, more importantly, is better than standard method, that models dosages directly, especially for markers with relatively low imputation quality.

email: kchuang@live.unc.edu

### **Statistical Calibration of qRT-PCR, Microarray and RNA-Seq Gene Expression Data with Measurement Error Models**

**Zhaonan Sun\***, Purdue University

**Thomas Kuczek**, Purdue University

**Yu Zhu**, Purdue University

The accurate quantification of gene expression levels is crucial for transcriptome study. Microarray and RNA-Seq are commonly used high-throughput technologies for transcriptome profiling. The gene expression measurements obtained by microarray and RNA-Seq are however subject to various measurement errors. A third platform called qRT-PCR is acknowledged to provide more accurate quantification of gene expression levels than microarray and RNA-Seq, but it has limited throughput capacity. We propose to use a system of functional measurement error models to model gene expression measurements and calibrate the microarray and RNA-Seq platforms with qRT-PCR. Based on the system, a two-step approach was developed to estimate the biases and error variance components of the three platforms and calculate calibrated estimates of gene expression levels. The calibrated estimates provide a more accurate and consistent quantification of gene expression levels. The system was applied to analyze two gene expression data sets from the Microarray Quality Control (MAQC) and Sequencing Quality Control (SEQC) projects.

email: sunz@purdue.edu

### **IUTA: A Statistical Method to Detect Differential Isoform Usage from mRNA-Seq Data**

**Liang Niu\***, National Institute of Environmental Health Sciences, National Institutes of Health

**Weichun Huang**, National Institute of Environmental Health Sciences, National Institutes of Health

**David M. Umbach**, National Institute of Environmental Health Sciences, National Institutes of Health

**Leping Li**, National Institute of Environmental Health Sciences, National Institutes of Health

So far, most computational and statistical methods for RNA-Seq data have focused on detecting differences between experimental conditions in the absolute expression (the abundance of transcript) at the gene or isoform level. Few methods focus on detecting differences in isoform usage, the relative expression of different isoforms, within a single gene. However, growing evidence suggests that differential isoform usage can be associated with diseases such as cancers. We regard isoform usage for a given gene as a compositional vector, that is, the relative expression of each isoform sums to one. We propose the Isoform Usage Two-Step Analysis (IUTA) to detect differential isoform usage between two groups when multiple samples are available from each group. In the first step, we estimate the isoform usage for the specified gene separately in each sample from each group by fitting a statistical model to the RNA-seq data. In the second step, we use the estimates from the first step to test the null hypothesis that the mean isoform usage for the gene is the same in each group, employing several previously proposed tests. We used real mRNA-seq data sets as well as simulated data sets to compare our approach with the popular methods Cuffdiff. We showed that our approach works for much more genes than Cuffdiff does and our approach achieves high sensitivity and specificity.

email: niul@niehs.nih.gov

