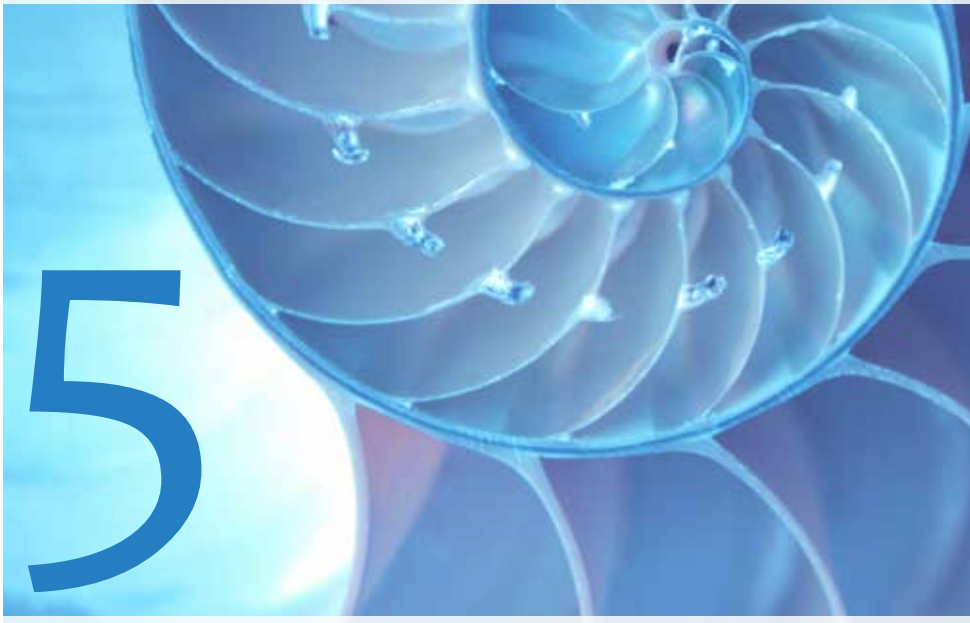
ENAR

ENAR 2015

SPRING MEETING

With IMS & Sections of ASA



MARCH 15–18

Hyatt Regency Miami | Miami, FL

ABSTRACTS

ENAR 2015

Abstracts & Poster Presentations

1. POSTERS: Latent Variable and Mixture Models

1a. ASSESSMENT OF DIMENSIONALITY CAN BE DISTORTED BY TOO MANY ZEROES: AN EXAMPLE FROM PSYCHIATRY AND A SOLUTION USING MIXTURE MODELS

Melanie M. Wall*, Columbia University

Irini Moustaki, London School of Economics

Common methods for determining the number of latent dimensions underlying an item set include eigenvalue analysis and examination of fit statistics for factor analysis models with varying number of factors. Given a set of dichotomous items, we will demonstrate that these empirical assessments of dimensionality are likely to underestimate the number of dimensions when there is a preponderance of individuals in the sample with all zeros as their responses, i.e. all incorrect answers. A simulated data experiment is conducted to demonstrate this phenomena. An example is shown from psychiatry assessing the dimensionality of a social anxiety disorder battery where only one latent dimension is found if the full sample is used, while three latent dimensions are found if the excess zeroes are accounted for correctly. A mixture model, i.e. hybrid latent class latent factor model, is used to assess the dimensionality of the underlying subgroup corresponding to those who come from the part of the population with some measurable trait. Implications of the findings are discussed, in particular regarding the potential for different findings in community versus patient populations.

email: **mmw2177@columbia.ed**

1b. LOCAL INFLUENCE DIAGNOSTICS FOR HIERARCHICAL COUNT DATA MODELS WITH OVERDISPERSION AND EXCESS ZEROS

Trias Wahyuni Rakhmawati*, Universiteit Hasselt

Geert Molenberghs, Universiteit Hasselt and Katholieke Universiteit Leuven

Geert Verbeke, Katholieke Universiteit Leuven and Universiteit Hasselt

Christel Faes, Universiteit Hasselt and Katholieke Universiteit Leuven

We consider models for hierarchically observed and possibly overdispersed count data, that in addition allow for excess zeros. The model extends the Poisson-normal generalized linear mixed model by including gamma random effects to accommodate overdispersion. Excess zeros are handled using either a zero-inflation or a hurdle component. These models were studied by Kassahun et al. (2014). While flexible, the model is

quite elaborate in parametric specification and therefore model assessment is imperative. We derive local influence measures to detect influential subjects, i.e., subjects who have undue influence on either the fit of the model as a whole, or on specific important sub-vectors. The latter include the fixed effects for the Poisson component and for the excess zeros component, the variance components for the normal random effects, and the parameters describing the gamma random effects. Interpretable influence components are derived. The methods are illustrated using data from a longitudinal clinical trial in patients with dermatophyte onychomycosis.

email: **triaswahyuni.rakhmawati@uhass**

1c. FINITE MULTIVARIATE MIXTURES OF SKEW-T DISTRIBUTIONS WITH COLLAPSE CLUSTERS WITH APPLICATION IN FORESTRY

Josef Hoefler*, Technical University Munich

Donna Pauler Ankerst, Technical University Munich

Finite mixtures of skew-t distributions offer a flexible framework for modeling non-Normal data, in particular data that possess skewness, multiple clusters and/or outliers. Such models have been extended to multivariate data, with detailed Expectation-Maximization (EM) algorithms for fitting. A practical problem that arises with these models is the existence of collapsed clusters, which reside on smaller-dimensional planes than the remaining clusters. This occurs for certain competition indices measured in trees – all trees in a particular plot or forest may experience zero lateral competition, for example. To handle this problem we have developed an R package [fitmixst4] that accommodates collapse clusters by constraining the variance of the cluster to be below a fixed upper bound. We apply the R package to retrospectively describe clusters of trees that died and remained alive over series of five-year follow-up periods from 9,292 beech trees in a Bavarian long-term forest research plot network. Heat maps of the mixture density ratios corresponding to dead versus alive trees are assessed as a potential prediction tool for forest mortality for future forest management.

email: **josef.hoefler@tum.de**

1d. WEIBULL MIXTURE REGRESSION FOR ZERO-HEAVY CONTINUOUS SUBSTANCE USE OUTCOMES

Mulugeta Gebregziabher, Medical University of South Carolina

Delia Voronca*, Medical University of South Carolina

Abeba Teklehaimanot, Medical University of South Carolina

Elizabeth J. Santa Ana, Ralph H. Johnson Department of Veterans Affairs Medical Center

Outcomes with preponderance of zero values are ubiquitous in data that arise from studies of addictive disorders. This is known to lead to violation of standard assumptions in parametric inference and enhances the risk of misleading conclusions unless managed properly. Two of the most popular models used to handle this issue for count outcomes are hurdle and zero-inflated models. Both models can be expressed as two-component mixtures. In a hurdle, the second component follows a zero-truncated distribution, while in a zero-inflated it follows a count distribution with positive probability of generating zeroes. However models that deal with this problem are not well developed for zero-heavy continuous outcomes. Thus, in this paper, we propose and evaluate a two-component Weibull mixture model for effectively dealing with the problem. We use simulated and real data from a randomized-controlled-trial (RCT) to demonstrate its application and make comparisons with other methods via statistical information and mean squared error criteria. Our results show that the two component Weibull mixture model is superior for modeling zero-heavy continuous data.

email: **gebregz@musc.edu**

1e. MODEL-FREE ESTIMATION OF TIME-VARYING CORRELATION COEFFICIENTS AND THEIR CONFIDENCE INTERVALS WITH AN APPLICATION TO fMRI DATA

Maria A. Kudela*, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis

Jaroslaw Harezlak, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis

Martin Lindquist, Johns Hopkins Bloomberg School of Public Health

One of main interests in fMRI (functional magnetic resonance imaging) research is the study of associations between time series from different brain regions, so called functional connectivity (FC). Recently, it has become increasingly important to assess dynamic changes

in FC, both during resting state and task-based fMRI experiments, as this is thought to provide the information needed to better understand the brain's inner workings. Currently, the most common approach to estimate these dynamic changes is by computing the correlation coefficient between time series within a sliding-window. However, one of the disadvantages of this method is that it tends to overestimate the association between the time series obtained from different brain regions (Lindquist et al. 2014). Here we propose a new approach for estimating time-varying FC using the correlation between two time series and provide valid confidence bands for this estimator. We propose an algorithm based on the sliding-window approach which utilizes the multivariate linear process bootstrap. Both numerical results and an application to fMRI study of alcoholism risk factors will be presented.

email: **maria.kudela@gmail.com**

1f. ZERO-AND-ONE INFLATED BETA REGRESSION WITH MIXED EFFECTS FOR MODELING RELATIVE FREQUENCY OF CONDOM USE IN MEN WHO HAVE SEX WITH MEN (MSM) IN GHANA

Nanhua Zhang*, Cincinnati Children's Hospital Medical Center

Yue Zhang, University of Cincinnati

LaRon E. Nelson, University of Rochester

Zero-and-one inflated proportion data are common in behavioral studies. Motivated by a study of the effect of psycho-social variables on the relative frequencies of condom use among men who have sex with men (MSM) in 22 social networks in Ghana, we consider a zero and one inflated beta regression with mixed effects. The proposed model addresses the issue of abundance of zeros and ones in the relative frequency of condom uses, and also the dependence of the relative frequencies of MSM from the same social network. We achieve this by linking the mixed-effects regression to the beta distribution mean through a logit link function while keeping the other positive parameter constant. We also discuss extension of the proposed model by also relating the positive parameter to the mixed-effects through a log link.

email: **nanhua.zhang@cchmc.org**

1g. INFERENCE FOR THE NUMBER OF TOPICS IN THE LATENT DIRICHLET ALLOCATION MODEL VIA A PSEUDO-MARGINAL METROPOLIS-HASTINGS ALGORITHM

Zhe Chen*, University of Florida

Hani Doss, University of Florida

Latent Dirichlet Allocation (LDA) is a model that is used for automatically organizing, understanding, searching, and summarizing a corpus of documents. Let V be the set of all words that appear at least once in at least one document. By definition, a topic is a distribution on V. LDA posits that for every word in every document, there is a latent topic from which that word is drawn. In standard LDA, the number of topics, T, must be specified in advance. A prior distribution is placed on the T topics, and also on the latent word-specific topic-indicator variables that are associated with each word in the corpus. The need to specify T in advance is problematic. If we put a prior on T, then the distribution on the latent variables is a mixture of distributions on spaces of different dimensions, and estimating this mixture distribution by MCMC is very challenging. We present a variant of the Metropolis-Hastings algorithm that can be used to effectively estimate this mixture distribution and in particular the posterior distribution on the number of topics. We also provide theory to justify that our algorithm can correctly estimate the true posterior distribution of T given the words. We evaluate the performance of our algorithm on synthetic data, with a comparison with the existing method. We also give an illustration on a collection of articles from Wikipedia.

email: **zhe.chen@ufl.edu**

1h. APPLYING A STOCHASTIC VOLATILITY MODEL TO US STOCK MARKETS WITH A UMM UNDERGRADUATE STUDENT

Jong-Min Kim*, University of Minnesota, Morris

Li Qin, University of Minnesota, Morris

This research is for University of Minnesota-Morris statistics undergraduate senior project. We use the log-normal & stochastic volatility model as a basis for investigating the absolute returns and the squared returns as two measures of latent volatility in financial markets. We also use linear regression with time varying parameters for investigating the relationships in financial markets. Furthermore, we show the log volatility forecasts by using the log-normal & stochastic volatility model.

email: **jongmink@morris.umn.edu**

1i. A MIXTURE MODEL OF HETEROGENEITY IN TREATMENT RESPONSE

Hongbo Lin*, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health, Indianapolis

Changyu Shen, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health, Indianapolis

In clinical trials, randomized studies are designed to estimate the average treatment effect. However, it is widely accepted that heterogeneity may exist in treatment effects; some patients may benefit from a medical intervention, while others may not. We propose a mixture-model based approach to study the heterogeneity of the treatment effect. We consider a latent binary variable that indicates whether or not a subject will benefit from an intervention. Our mixture model combines a logistic formulation for the probability of a patient benefitting from an intervention with proportional hazards models conditional on the status of the latent variable. EM algorithm is used to estimate the parameters in the model. Standard errors are calculated using Louis's method. Simulations are performed to study the properties of the estimators. The method is also applied 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@iu.edu**

1j. BAYESIAN RANDOM GRAPH MIXTURE MODEL FOR COMMUNITY DETECTION IN WEIGHTED NETWORKS

Christopher Bryant*, University of North Carolina, Chapel Hill

Mihye Ahn, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

Joseph Ibrahim, University of North Carolina, Chapel Hill

The network paradigm has become a popular approach for modeling complex systems, with applications ranging from social sciences to genetics to neuroscience and beyond. Often the individual connections between network nodes are of less interest than network characteristics such as its community structure - the tendency in many real-data networks for nodes to be naturally organized in groups with dense connections between nodes in the same (unobserved) group but sparse connections between nodes in different groups. Most community detection algorithms involve optimization of various connectedness measures in order to achieve this structure, rather than an explicit probabilistic framework. Random graph mixture models utilize such a framework and can accurately capture latent communities in either binary or weighted networks. We fit a Bayesian hierarchical model to Gaussian-weighted networks via Gibbs sampling, which allows for community detection across multiple subjects and even for small graphs or sub-graphs. We show results from simulated networks and apply the method to estimate the community structures in the functional resting brain networks of 185 subjects from the ADHD-200 sample.

email: **cmbrya@unc.edu**

1k. TIME SERIES FORECASTING USING MODEL-BASED CLUSTERING AND MODEL AVERAGING

Fan Tang*, University of Iowa

Joseph Cavanaugh, University of Iowa

Time series forecasting is an important practical problem. By incorporating information from series that exhibit similar long-term and transitory behaviors, we can potentially improve the forecasting accuracy for a particular series of interest. However, identifying and appropriately utilizing a cluster of related series from a larger pool presents a daunting challenge. This paper introduces a time-series forecasting procedure that relies on model-based clustering and model averaging. The clustering algorithm employs a state-space model comprised of three latent structures: a long-term trend component; a seasonal component, to capture recurring global patterns; and an anomaly component, to reflect local perturbations. A two-step clustering algorithm is applied to identify series that are both globally and locally correlated, based on corresponding smoothed latent structures. For each series in a particular cluster, a set of forecasting models are fit, using covariate series from the same

cluster. To fully utilize the cluster information and to improve forecasting for a series of interest, multi-model averaging is employed. The proposed technique is applied to a collection of monthly disease incidence series. Our approach yields both clinically and statistically meaningful clusters, and through model averaging, produces more accurate forecasts than any individual model.

email: **fan-tang@uiowa.edu**

1l. MULTILEVEL FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS OF SURFACES WITH APPLICATION TO CT IMAGE DATA OF PEDIATRIC THORACIC SHAPE

Lucy F. Robinson*, Drexel University

Jonathan Harris, Drexel University

Sriram Balasubramanian, Drexel University

We propose a multilevel functional principal components model for multivariate spatial functional data with one and two-dimensional arguments. As a motivating example, we consider 3D CT image-based reconstructions of rib cages from normative pediatric subjects and those with skeletal deformities. Variation and symmetry in the shape of individual ribs and of the exterior surface of each side of the rib cage are of interest. Existing analysis methods for thoracic shape typically focus on a priori defined simple geometric summaries. The data have a multilevel structure with multiple functional objects (ribs, sides) observed within each subject's image, and variation is modeled at the levels of subject, rib, and side. We propose a multilevel functional principal components model for shape of ribs and the ribcage surface, using a spherical harmonic basis to represent the functional objects. Derived components are used to assess lateral symmetry within each subject, to model the relationship between shape and covariates of interest, and to identify latent clinical subpopulations in patients with thoracic deformities.

email: **lucy.f.robinson@gmail.com**

1m. A NEW APPROACH FOR TREATMENT NONCOMPLIANCE WITH STRUCTURAL ZERO DATA

Pan Wu*, Christiana Care Health System

In randomized controlled trails, the confounding of non-compliance after initial treatment assignment is a serious problem that could lead to biased estimation of treatment effect and cause-plausible interpretation for study results. Further, it is usually inappropriate to assume the variable measuring post-treatment non-compliance follows single-mode distributions such as normal or Poisson, especially for mental health studies, since such a non-compliance variable, i.e., the amount of treatment participation, reflects the attitude of acceptance of such treatment by patients, which can be quite heterogeneous across patients. Existing approaches are unable to address such non-compliance patterns that are described by models for structural zero data. In the talk, we would like to propose a new framework of Structural Equation Models with robust inference to estimate the causal effect between two active treatment arms with non-compliance in each group. The proposed models are able to address the patient heterogeneity in acceptance of treatment. Instead of using likelihood based inference, the proposed methods require no assumption of parametric distribution and offer consistent estimation of model parameters with asymptotically normal distributions under mild regularity conditions.

email: **pan_wu@urmc.rochester.edu**

2. POSTERS: Imaging Methods and Applications

2a. DETERMINING MULTIMODAL NEUROIMAGING MARKERS OF PARKINSON'S DISEASE

DuBois Bowman*, Columbia University

Weingiong Xue, Boehringer Ingelheim

Daniel Drake, Columbia University

Advances in biomedical imaging technology have led to an increase in health research studies that collect large-scale, multimodal data sets, frequently in conjunction with genomic data, and biologic and clinical measures. Such studies provide an unprecedented opportunity for cross-cutting investigations that stand to gain a deeper understanding of the pathophysiology associated with major diseases. We develop a Bayesian hierarchical model for the analysis of multimodal neuroimaging data to investigate neural markers of Parkinson's disease (PD). Our model incorporates imaging data, reflecting both functional and structural characteristics of the brain, incorporates spatial correlations between distinct brain regions, and yields classifications of subjects as either PD patients or healthy controls (HCs). Applying the model to multimodal magnetic resonance-based images, we demonstrate the ability to isolate neural characteristics that reflect accurate signatures of PD and that hold promise for serving as useful early stage PD biomarkers.

email: **dubois.bowman@columbia.edu**

2b. SEGMENTATION OF INTRACEREBRAL HEMORRHAGE IN CT SCANS USING LOGISTIC REGRESSION

John Muschelli*, Johns Hopkins Bloomberg School of Public Health

Natalie Ullman, Johns Hopkins School of Medicine

Daniel Hanley, Johns Hopkins School of Medicine

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

Intracranial hemorrhage (ICH) is a neurological condition that results from a blood vessel rupturing into tissues and possibly the ventricles of the brain and is often fatal. X-ray computed tomography (CT) scans is the most commonly used diagnostic tool in patients with ICH and allows quantitative description of ICH in Hounsfield units (HU). The size of the ICH is highly predictive of good functional outcome in patients. The gold standard measurement of ICH is manual segmentation of CT scans, which is time-consuming and subject to intra- and inter-observer variability. We present a regression modeling framework for estimating the probability of ICH in a voxel. We estimated this model from 10 patient scans, out of 112 scans, and estimated model performance on the left-out 92 scans. The area under the curve (AUC) for a receiver operating characteristic (ROC) curve, partial AUC (pAUC), and total accuracy (% correctly classified) were estimated to determine which segmentation methods performed well. We achieved greater than 90% accuracy in all left-out scans. This model represents the first automated segmentation procedure for ICH using regression methods. As such, we can infer the predictive power for each explanatory variable.

email: **jmuschel@jhsph.edu**

2c. RELATING MULTI-SEQUENCE LONGITUDINAL DATA FROM MS LESIONS ON STRUCTURAL MRI TO CLINICAL COVARIATES AND OUTCOMES

Elizabeth Sweeney*, Johns Hopkins Bloomberg School of Public Health

Blake Dewey, National Institute of Neurological Disease and Stroke, National Institutes of Health

Daniel Reich, National Institute of Neurological Disease and Stroke, National Institutes of Health

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

Russell Shinohara, University of Pennsylvania

Ani Eloyan, Johns Hopkins Bloomberg School of Public Health

Structural magnetic resonance imaging (MRI) can be used to detect lesions in the brains of multiple sclerosis (MS) patients. The formation of these lesions is a complex sequence of inflammation, degeneration, and repair that MRI has been shown to be sensitive to. We characterize the lesion formation process with multi-sequence structural MRI. We have longitudinal MRI from 60 MS patients, each with between 10 and 40 studies consisting of a T1-weighted, T2-weighted, fluid attenuated inversion recovery

(FLAIR) and proton density (PD) volume. We extract the multi- sequence longitudinal voxel intensities from the four volumes using SuBLIME, a method for detection of incident and enlarging MS lesions voxels. Next we spatially and temporally smooth the volumes and use functional principal component analysis to identify voxels that contain permanent damage and repair. We then investigate this repair and permanent damage in relation to clinical covariates such as disease duration, MS subtype, Expanded Stability Status Score (EDSS), and treatment.

email: e**msweene@jhsph.edu**

2d. USING MULTIPLE IMPUTATION TO EFFICIENTLY CORRECT MAGNETIC RESONANCE IMAGING DATA IN MULTIPLE SCLEROSIS

Alicia S. Chua*, Brigham and Women's Hospital, Boston

Svetlana Egorova, Brigham and Women's Hospital, Boston

Mark C. Anderson, Brigham and Women's Hospital, Boston

Mariann Polgar-Turcsanyi, Brigham and Women's Hospital, Boston

Tanuja Chitnis, Brigham and Women's Hospital, Boston

Howard L. Weiner, Brigham and Women's Hospital, Boston

Charles R. Guttmann, Brigham and Women's Hospital, Boston

Rohit Bakshi, Brigham and Women's Hospital, Boston

Brian C. Healy, Brigham and Women's Hospital, Boston

Current technologies have enabled fully/semi-automated segmentation of MRI scans for the assessment of multiple sclerosis (MS). However, manual correction of these images by an expert reader remains desirable. Since automated segmentation data awaiting manual correction is "missing", we proposed to use multiple imputation (MI) to fill-in the missing manually-corrected MRI data for measures of brain atrophy and lesion burden. Scans from 1370 patients enrolled in the Comprehensive Longitudinal Investigation of Multiple Sclerosis at the Brigham and Women's Hospital (CLIMB) study were identified. Simulation studies were conducted to assess the performance of MI with missing data both missing completely at random and missing at random. An imputation model including the semi-automated data explained a very high proportion of the variance in the manually corrected data for both outcome measures (R2>.90 for each), demonstrating the potential to accurately impute the missing data. Further, our results demonstrate that MI allows for the accurate estimation of group differences with little to no bias and with similar precision compared to an analysis with no missing data. We believe that our findings provide important insights for efficient correction of automated or semi-automated MRI measures to reduce the burden of manual of correction.

email: **aschua@partners.org**

2e. BACKGROUND ADJUSTMENT AND VOXELWISE INFERENCE FOR TEMPLATE-BASED GAUSSIAN MIXTURE MODELS

Meng Li*, North Carolina State University

Armin Schwartzman, North Carolina State University

In brain oncology, it is routine to analyze the progression or remission of the disease based on the differences between a pre-treatment and a post-treatment Positron Emission Tomography (PET) scan. The analysis is challenging because differences between two scans are expected even in the regions that are not affected by the disease. To overcome this problem, it has been previously proposed to segment the images using a template-based Gaussian mixture model (GMM) and then adjust the background of the two scans within each class, making the differences in the disease regions stand out. However, in spite of the anatomical guidance provided by the spatial template, the voxelwise mixture probabilities are typically not accurately estimated, making the background adjustment and inference difficult. In this paper, we propose a statistical testing procedure to detect localized differences between the images using a template-based GMM approach. We show that the voxelwise test statistic produced by background adjustment is very close to the standard Gaussian in a wide range of scenarios, making it suitable for statistical inference. In particular, the standard Gaussian approximation is stable even when the mixture probabilities are not accurately estimated, and it tends to be conservative at the tails, assuring the test's validity. We confirm the good performance

of the proposed approach by simulations and phantom experiments. The proposed approach can be applied directly by practitioners since the resulting p-value can provide immediate reference when making conclusions and decisions with respect to the change of the disease status.

email: **mli9@ncsu.edu**

2f. FAST, FULLY BAYESIAN SPATIO-TEMPORAL INFERENCE FOR fMRI

Donald R. Musgrove*, University of Minnesota

John Hughes, University of Minnesota

Lynn E. Eberly, University of Minnesota

We propose a sparse spatial Bayesian variable selection method for functional magnetic resonance imaging (fMRI). Typical fMRI experiments generate huge datasets with complex spatiotemporal dependence structures. To ease the computational burden, we separate the brain into three-dimensional parcels whereby inference occurs parcel-wise in parallel. Volume element (voxel) activation within parcels is modeled as a series of autocorrelated regressions on a lattice. Regressors represent change in blood oxygenation in response to stimuli while indicator variables capture the nonzero change. Via a reparameterized Gaussian Markov random field prior, a sparse spatial generalized linear mixed model (SSGLMM) is used to model spatial dependence among indicator variables within a parcel for a given stimulus. In particular, the SSGLMM accounts for both large-scale and small-scale spatial variation. Simulations show that the parcellation performs well under varying assumptions. Indicators on parcel boundaries do not suffer edge effects and maintain a low false discovery rate. With an event related experiment, we show that the model is easy to implement and offers certain advantages over whole brain modeling.

email: **musgr007@umn.edu**

2g. BAYESIAN SPATIAL VARIABLE SELECTION FOR ULTRA-HIGH DIMENSIONAL NEUROIMAGING DATA: A MULTIRESOLUTION APPROACH

Yize Zhao*, Statistical and Applied Mathematical Sciences Institute

Jian Kang, Emory University

Qi Long, Emory University

Ultra-high dimensional variable selection has become increasingly important in analysis of neuroimaging data. For example, in the Autism Brain Imaging Data Exchange (ABIDE) study, neuroscientists are interested in identifying important biomarkers for early detection of the autism spectrum disorder (ASD) using high resolution brain images that include hundreds of thousands voxels. However, most existing methods are not feasible for solving this problem due to their extensive computational costs. In this work, we propose a novel multiresolution variable selection procedure under a Bayesian probit regression framework. It recursively uses posterior samples for coarser-scale variable selection to guide the posterior inference on finer-scale variable selection, leading to very efficient Markov chain Monte Carlo (MCMC) algorithms. The proposed algorithms are computationally feasible for ultra-high dimensional data. Also, our model incorporates two levels of structural information into variable selection using Ising priors: the spatial dependence between voxels and the functional connectivity between anatomical brain regions. Applied to the resting state functional magnetic resonance imaging (R-fMRI) data in the ABIDE study, our methods identify voxel-level imaging biomarkers highly predictive of the ASD, which are biologically meaningful and interpretable. Extensive simulations also show that our methods achieve better performance in variable selection compared to existing methods.

email: **yzhao@samsi.info**

2h. ANALYSIS OF HIGH DIMENSIONAL BRAIN SIGNALS IN DESIGNED EXPERIMENTS USING PENALIZED THRESHOLD VECTOR AUTOREGRESSION

Lechuan Hu*, University of California, Irvine

Hernando Ombao, University of California, Irvine

One way to measure cortical activity is by electroencephalograms (EEG) which recorded across many channels on the entire surface of the scalp. Using EEG, one can infer the nature of neuronal activity in the cortex and the cross-regional interactions. In designed experiments there are many high-dimensional EEG

traces recorded across many trials. Our goal is to infer the brain dynamics using all the high dimensional time series data using threshold vector autoregressive models (T-VAR). Due to the sheer size and dimension of the data, it is difficult to estimate the parameters in the VAR model. Here we will develop a complexity-penalized TVAR to infer connectivity between brain regions. We will use the proposed model to study connectivity in resting-state using 1-second multichannel- EEG traces recorded for over 3 minutes. This work has been in collaboration with the Space-Time Modeling Group at UC Irvine.

email: l**echuanh@uci.edu**

2i. SPATIALLY WEIGHTED REDUCED-RANK FRAMEWORK FOR NEUROIMAGING DATA WITH APPLICATION TO ALZHEIMER'S DISEASE

Mihye Ahn*, University of Nevada, Reno

Haipeng Shen, University of North Carolina, Chapel Hill

Chao Huang, University of North Carolina, Chapel Hill

Yong Fan, University of Pennsylvania

Hongtu Zhu, University of North Carolina, Chapel Hill

In neuroimaging studies, it is challenging to incorporate multiple subjects for group inference due to spatial-temporal functional variation. In this paper, we propose a new modelling framework for analyzing functional connectivity pattern across subjects by considering spatial and temporal similarity on whole brain images. To reduce noise sensitivity, we conduct the analysis in the frequency domain, and also impose sparsity on the frequency basis function for better interpretation. From the framework, we also extract the subject-specific spatial factors which enable us group comparison. We discuss optimization strategies to avoid lack of memory in practice. Numerical results show that the spatially weighted framework has lower variability regardless of poor alignment and high inter-subject variability. Finally, we apply the proposed method to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data.

email: **ahnm@email.unc.edu**

2j. HIGHLY ADAPTIVE TEST FOR GROUP DIFFERENCES IN BRAIN FUNCTIONAL CONNECTIVITY

Junghi Kim*, University of Minnesota

Wei Pan, University of Minnesota

Resting-state functional magnetic resonance imaging (rs-fMRI) and other technologies have been offering increasing evidence showing that altered brain functional networks are associated with neurological illnesses such as Alzheimer's disease. However, group-level network analysis is both challenging and necessary due to the high dimensionality of network models and high noise levels in neuroimaging data. Varoquaux and Craddock (2013) highlighted that "there is currently no unique solution, but a spectrum of related methods and analytical strategies" to learn and compare brain connectivity. An important issue is how to choose several critical parameters in estimating a network, such as what association measure to use and what is the sparsity of the estimated network. In particular, an optimal choice of a parameter for network estimation may not be optimal for testing. On the other hand, mis-specified values of these parameters may lead to extremely low-powered tests. Here we present highly adaptive tests for group differences in brain connectivity, which automatically combine statistical evidence against a null hypothesis from multiple sources across a wide range of the plausible parameter values. These highly adaptive tests are not only easy to use, but also high-powered robustly across various scenarios. The advantages of these novel tests are demonstrated on realistically simulated data and an Alzheimer's disease dataset.

email: **kimx2859@umn.edu**

2k. PRE-SURGICAL fMRI DATA ANALYSIS USING A SPATIALLY ADAPTIVE CONDITIONALLY AUTOREGRESSIVE MODEL

Zhuqing Liu*, University of Michigan

Veronica J. Berrocal, University of Michigan

Andreas J. Bartsch, University of Heidelberg

Timothy D. Johnson, University of Michigan

Spatial smoothing is an essential step in the analysis of functional magnetic resonance imaging (fMRI) data. One standard smoothing method is to convolve the image data with a three-dimensional

Gaussian kernel that 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 can blur the boundaries between activated and deactivated regions of the brain. Moreover, while in a standard fMRI analysis strict false positive control is desired, for pre-surgical planning false negatives are of greater concern. To this end, we propose a novel spatially adaptive conditionally autoregressive model with smoothing variances that are proportional to error variances, allowing the degree of smoothing to vary across the brain and present a new loss function that allows for the asymmetric treatment of false positives and false negatives. We compare our proposed model with two existing spatially adaptive smoothing models. Simulation studies show that our model outperforms these other models; as a real model application, we apply the proposed model to the pre-surgical fMRI data of a patient to assess peri- and intra-tumoral brain activity.

email: **zhuqingl@umich.edu**

2l. SEMIPARAMETRIC BAYESIAN MODELS FOR LONGITUDINAL MR IMAGING DATA WITH MULTIPLE CONTINUOUS OUTCOMES

Xiao Wu*, University of Florida

Michael J. Daniels, University of Texas, Austin

This research is motivated by data from a Duchenne Muscular Dystrophy study on changes in muscle imaging data to capture disease progression over time. We develop a semiparametric Bayesian modeling approach for MR imaging analysis involves combining multiple measures of multiple muscles over time. Dependence among different outcomes is induced through latent variables and nonparametric priors are used for the random effects distribution. A Markov chain Monte Carlo algorithm is proposed for estimating the posterior distributions of the parameters and latent variables.

email: **xiaowu@stat.ufl.edu**

2m. IMPROVING RELIABILITY OF SUBJECT-LEVEL RESTING-STATE BRAIN PARCELLATION WITH EMPIRICAL BAYES SHRINKAGE

Amanda F. Mejia*, Johns Hopkins University

Mary Beth Nebel, Johns Hopkins University

Haochang Shou, Johns Hopkins University

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

James J. Pekar, Johns Hopkins University School of Medicine

Stewart Mostofsky, Johns Hopkins University

Brian Caffo, Johns Hopkins University

Martin Lindquist, Johns Hopkins University

A recent interest in resting state functional magnetic resonance imaging (rsfMRI) lies in subdividing the human brain into functionally distinct regions of interest. One common parcellation technique is

clustering, which begins with a measure of similarity between voxels. The goal of this work is to improve the reproducibility of single-subject parcellation using shrinkage-based estimators of such measures, allowing the noisy subject-specific estimator to borrow strength from a larger population of subjects. We present several shrinkage estimators and outline methods for estimating the within-subject variance when multiple scans are not available for each subject. We perform shrinkage on raw inter-voxel correlation estimates and use both raw and shrinkage estimates to produce parcellations by performing clustering on the voxels. Using two datasets - a simulated dataset where the true parcellation is known, and a test-retest dataset consisting of two 7-minute resting-state fMRI scans from 20 subjects - we show that parcellations produced from shrinkage correlation estimates have higher reliability and validity than those produced from raw correlation estimates. Application to test-retest rsfMRI data shows that using shrinkage estimators increases the reproducibility of subject-specific parcellations of the motor cortex by up to 30 percent.

email: **amejia@jhsph.edu**

3. POSTERS: Clinical Trials, Adaptive Designs and Applications

3a. THE ROLE OF STATISTICIANS IN REGULATORY DRUG SAFETY EVALUATION

Clara Kim*, U.S. Food and Drug Administration

Mark Levenson, U.S. Food and Drug Administration

Food and Drug Administration Amendments Act of 2007 granted the FDA the authority to require post-marketing safety studies, and labeling change to include new safety information. Since then, FDA has substantially strengthened its safety program for marketed drugs. Major actions to advance drug safety monitoring include enhanced capabilities of statistical analysis. The Division of Biometrics 7 (DB7) of the Office of Biostatistics in the Center for Drug Evaluation and Research of FDA is dedicated to full-cycle drug safety evaluation. This division is responsible for meta-analyses, evaluating clinical trials designed primarily to study safety outcomes, and observational studies submitted as a post-marketing requirement. Additionally, DB7 has expertise in the design and statistical methods used in studies that utilize surveillance systems, and registry or health care databases, such as Sentinel and FDA initiated pharmacoepidemiological studies. This poster will describe these activities with examples that exemplify DB7 contributions that resulted in regulatory actions.

email: **Clara.kim@fda.hhs.gov**

3b. ANALYZING MULTIPLE ENDPOINTS IN A CONFIRMATORY RANDOMIZED CLINICAL TRIAL: AN APPROACH THAT ADDRESSES STRATIFICATION, MISSING VALUES, BASELINE IMBALANCE AND MULTIPLICITY FOR STRICTLY ORDINAL OUTCOMES

Hengrui Sun*, University of North Carolina, Chapel Hill

Atsushi Kawaguchi, Kyoto University, Japan

Gary Koch, University of North Carolina, Chapel Hill

Background: Many confirmatory randomized clinical trials that compare two treatments have strictly ordinal response outcomes with stratified design. Multiple endpoints are often collected when one single endpoint does not represent the overall efficacy of the treatment. Baseline imbalance and missing values add another layer of difficulty in the analysis plan. Therefore, the development of an approach that provides a consolidated solution is essential. Methods: Multivariate Mann-Whitney estimators with stratification adjustment were used to handle the strictly ordinal responses. Randomization based nonparametric analysis of covariance was applied to account for the possible baseline imbalances. Several approaches that handle missing values were compared. A global test followed by closed testing procedure was conducted so that family wise error rate (FWER) was controlled in the strong sense in the analysis of multiple endpoints. Results: The data analyzed are from a clinical trial that compares a test treatment and a control for the pain management for patients with osteoarthritis. Four outcomes indicating joint pain, stiffness and functional status were analyzed collectively and individually through the procedures. Treatment efficacy was observed in the combined endpoint as well as in the individual endpoints. Conclusions: The proposed approach is effective in solving the aforementioned problems simultaneously.

email: **hrsun@email.unc.edu**

3c. COMPARING THE STATISTICAL POWER OF ANALYSIS OF COVARIANCE AFTER MULTIPLE IMPUTATION AND THE MIXED MODEL IN TESTING THE TREATMENT EFFECT FOR PRE-POST STUDIES WITH LOSS TO FOLLOW-UP

Wenna Xi*, The Ohio State University

Michael L. Pennell, The Ohio State University

Rebecca R. Andridge, The Ohio State University

Electra D. Paskett, The Ohio State University

In pre-post studies with complete follow-up, previous studies have shown that analysis of covariance (ANCOVA) is more powerful than the change-score analysis in comparing the intervention

group to control. However, there have been no comparisons of power under missing post-test values. The goal of this study was to compare the power of two methods: ANCOVA after multiple imputation (MI) and the mixed model, in testing the treatment effect when post-test values are missing. To do so, we analyzed the BePHIT data and performed simulation studies. Four methods were used and compared: ANCOVA after MI, complete-case ANCOVA, the all-available data mixed model, and the complete-case mixed model. Simulation studies were conducted under various sample sizes, missingness rates, and missingness scenarios. In the analysis of the BePHIT study data, ANCOVA after MI had the smallest p-value. The simulation results demonstrated that ANCOVA after MI was usually more powerful than the all-available data mixed model when the missingness percentage was moderate (20% and 30%). However, the power of ANCOVA after MI dropped the fastest as the missingness rate increased and, in most simulated scenarios, was the least powerful method when 50% of the post-test outcomes were missing.

email: **xi.34@osu.edu**

3d. EXTENDING LOGISTIC REGRESSION LIKELIHOOD RATIO TEST ANALYSIS TO DETECT SIGNALS OF VACCINE-VACCINE INTERACTIONS IN VACCINE SAFETY SURVEILLANCE

Kijoeng Nam*, U.S. Food and Drug Administration

Nicholas C. Henderson, University of Wisconsin, Madison

Patricia Rohan, U.S. Food and Drug Administration

Emily Jane Woo, U.S. Food and Drug Administration

Estelle Russek-Cohen, U.S. Food and Drug Administration

Adverse vaccine effects (AVEs) might arise from vaccine interactions in addition to AVEs from individual vaccines and may not be detected until the postmarket stage. The Vaccine Adverse Event Reporting System (VAERS) is a national vaccine safety surveillance program co-sponsored by the Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA). The VAERS database contains reports of adverse events associated with immunization and disproportionality analysis can be used to explore vaccine interaction adverse effects (VIAEs). In this paper, we develop a logistic regression based likelihood ratio test (LR-LRT) for detecting interactions between vaccines that may signal potential safety concerns. We evaluate our procedure with several numerical simulations, and we compare our results with known safety profiles, to validate the ability of our method to detect potential VIAEs.

email: **Kijoeng.Nam@fda.hhs.gov**

3e. DOSE-FINDING APPROACH BASED ON EFFICACY AND TOXICITY OUTCOMES IN PHASE I ONCOLOGY TRIALS FOR MOLECULARLY TARGETED AGENTS

Hiroyuki Sato*, Pharmaceuticals and Medical Devices Agency

Akihiro Hirakawa, Nagoya University Graduate School of Medicine

Chikuma Hamada, Tokyo University of Science

The paradigm of oncology drug development is expanding from cytotoxic agents to biological or molecularly targeted agents. It is common for cytotoxic agents that the efficacy and toxicity monotonically increase with dose escalation. However, for some molecularly targeted agents, the efficacy may exhibit non-monotonic patterns in their dose-response relationships. Many existing dose-finding approaches form non-monotonic patterns in dose-efficacy curve by using specific models, such as Quadratic model. In this study, we propose a novel Bayesian adaptive dose-finding approach based on binary efficacy and toxicity outcomes. We develop a dose-efficacy model whose parameters are allowed to change before and after the change point of dose in order to take into consideration the non-monotonic pattern of the dose-efficacy relationship. The change point is obtained as the study dose maximizing log likelihood given the model parameter estimates. These model parameters are estimated by using Markov chain Monte

Carlo methods under the assumption that each study dose is the change point. During the trial, we continuously estimate the posterior probabilities of efficacy and toxicity and assign patients to the most appropriate dose based on the decision rules we defined. We evaluate the operating characteristics of the proposed approach through simulation studies under various scenarios.

email: **sato-hiroyuki@pmda.go.jp**

3f. EFFECT SIZE MEASURES AND META-ANALYSIS FOR ALTERNATING TREATMENT SINGLE CASE DESIGN DATA

D Leann Long*, West Virginia University

Mathew Bruckner, West Virginia University

Regina A. Carroll, West Virginia University

George A. Kelley, West Virginia University

Single case designs (SCD) are employed in several fields of research where the treatment and outcome measures of interest require a high degree of tailoring to individual cases, as well as when the study conditions are in short supply. Alternating treatment SCD are characterized by the swift alternation between different treatments or conditions within a case, each associated with a distinct stimulus. Unique statistical challenges arise for SCD, particularly due to smaller sample sizes than desired for traditional statistical theory and the repeated nature of the treatments within a subject. The statistical methods generally conducted in these scenarios focus on treatment effects within subjects rather than the examination of treatment effects across several cases. For alternating treatment SCD data, we investigate various measures of effect size within each case and apply well-established meta-analytic methods for examination of treatment effects across cases. Our motivating example arises from the behavior analysis literature, where researchers are interested in assessing educational interventions on children with autism spectrum disorders.

email: **dllong@hsc.wvu.edu**

3g. CLINICAL TRIALS WITH EXCLUSIONS BASED ON RACE, ETHNICITY, AND ENGLISH FLUENCY

Brian L. Egleston*, Fox Chase Cancer Center, Temple University

Omar Pedraza, Fox Chase Cancer Center, Temple University

Yu-Ning Wong, Fox Chase Cancer Center, Temple University

Roland L. Dunbrack Jr., Fox Chase Cancer Center, Temple University

Eric A. Ross, Fox Chase Cancer Center, Temple University

J. Robert Beck, Fox Chase Cancer Center, Temple University

Recruiting diverse populations to clinical trials helps ensure that study results are generalizable to the population at large. We are currently examining the characteristics of clinical trials that have explicit inclusion criteria related to race, ethnicity, or English fluency. While explicit exclusion of African Americans from clinical trials has received attention from advocacy organizations, there has not been much research documenting the degree to which such exclusions continue to be widespread. We hypothesize that clinical trials with sponsoring institutions located in more racially and ethnically diverse areas are more likely to have racial, ethnic, or English fluency-related eligibility criteria. We are using data from the ClinicalTrials.gov database linked to United States Census and American Community Survey data. Our preliminary findings with respect to English language exclusions suggest that trials located at institutions in ZIP codes with more residents self-identifying as Black/African American or Asian are more likely to require that participants be fluent in English. Conversely, clinical trials located in areas with more residents self-identifying as Hispanic are less likely to have English fluency requirements. Clinical trial statisticians may have an opportunity to address inclusion concerns when designing trials.

email: **brian.egleston@fccc.edu**

3h. COMPARING FOUR METHODS FOR ESTIMATING OPTIMAL TREE-BASED TREATMENT REGIMES

Aniek Sies*, Katholieke Universiteit Leuven

Iven Van Mechelen, Katholieke Universiteit Leuven

When multiple treatment alternatives are available for a certain disease, an important challenge is to find an optimal treatment regime, which specifies for each patient the preferred treatment

alternative given his or her pretreatment characteristics. An interesting class of treatment regimes is that of the tree-based ones, because they provide a straightforward and most insightful representation of the decision structure. Recently, several methods for the construction of tree-based regimes have been proposed. Up to now however, only partial information is available concerning their absolute and relative performance. Our paper addresses this issue by reporting the results of an extensive simulation study to evaluate four tree-based methods, namely "Interaction Trees", "Model-based recursive partitioning", "Quint", and an approach developed by Zhang et al. (2012). The main evaluation criterion is the expected potential outcome if the entire population would be subjected to the optimal treatment regime resulting from each method under study.

email: **aniek.sies@ppw.kuleuven.be**

3i. COMPARING METHODS OF ADJUSTING FOR CENTER EFFECTS USING PEDIATRIC ICU GLYCEMIC CONTROL DATA

Samantha Shepler*, Emory University

Scott Gillespie, Emory University

Traci Leong, Emory University

In multi-site randomized clinical trials, randomization is balanced by institution to minimize any confounding center effects. However, in a multi-site, non-randomized clinical trial with a single intervention, this balancing is not possible. In this presentation, we test for center effects in the estimation of length of stay by known prognostic factors. We propose three methods to adjust for these effects: (1) Clustering, (2) Hierarchical Bayesian, (3) and Latent Variable and will explain the practical interpretation. Through simulation, bias of each method will be calculated. These approaches will be applied to the analysis of glycemic control data in six pediatric Intensive Care Units (ICU) from around the country, utilizing ICU length of stay as our primary endpoint.

email: **ssheple@emory.edu**

3j. BAYESIAN DOSE FINDING PROCEDURE BASED ON INFORMATION CRITERION

Lei Gao*, Sanofi

William F. Rosenberger, George Mason University

Zorayr Manukyan, Pfizer Inc.

In dose-finding studies with toxicity-efficacy responses, penalty functions and Bayesian procedures are used to find a single optimal dose with ethical toxicity-efficacy trade-offs. It has been widely seen that such designs can select the wrong dose when the working prior is wrong, largely due to a "stickiness" property of miring at a single dose. As one possible remedy, we present a family of compound optimal designs that involve both efficiency of estimation for updating prior parameters and the ethical criteria that minimize highly toxic or ineffective does. It is shown that most Bayesian sequential designs for dose finding in the literature can be thought of as a special case of this family of designs. We conduct simulations using Markov chain Monte Carlo (MCMC) algorithms to examine the convergence of Bayesian dose finding designs and investigate their operating characteristics.

email: **leilei.gao@hotmail.com**

3k. THE RELATIONSHIP AMONG TOXICITY, RESPONSE, AND SURVIVAL PROFILES ULTIMATELY INFLUENCE CALLING A BENEFICIAL EXPERIMENTAL DRUG FAVORABLE UNDER STANDARD PHASE I, II, AND III CLINICAL TRIAL DESIGNS

Amy S. Ruppert*, The Ohio State University

Abigail B. Shoben, The Ohio State University

Background: The success rate for investigational drugs from Phase I through III is less than 15% (Hay, 2014); potential reasons include suboptimal Phase I dose levels declared for further study, inadequate surrogate end points, lack of randomization, and other Phase II and III design decisions. Trial design efficiency within stages has been assessed, but few have evaluated the process as a continuum. We aimed to characterize the ability to recognize a clinically beneficial drug across stages using standard designs. Methods: Standard Phase I (3+3), II (Simon's optimal 2-stage), and III (1:1 randomized group sequential) designs were implemented. Dose limiting toxicity and response data were assumed binomially distributed, and survival data exponentially distributed. Two toxicity (constant and step), three response and three survival patterns (constant

and two with increasing efficacy) were evaluated using simulation. Results: With low constant toxicity, standard designs performed well regardless of response and survival patterns (experimental agent favorable in 87.6% of simulations). Under stepped toxicity and increasing response/survival profiles that were strongly dose-dependent, favorable decisions occurred less frequently (64.0% of simulations). Understanding the performance of standard design methods under a variety of toxicity, response and survival profiles will be crucial when comparing performance with non-standard designs.

email: **Amy.sr.stark@gmail.com**

3l. DOSE-FINDING USING HIERARCHICAL MODELING FOR MULTIPLE SUBGROUPS

Kristen May Cunanan*, University of Minnesota

Joseph S. Koopmeiners, University of Minnesota

Primarily, phase I clinical trials determine a new treatment's highest dosage with an acceptable toxicity rate, defined as the maximum tolerated dose (MTD), via a dose-finding study. In clinical trials, we have seen hierarchical modeling (HM) used in many applications to improve estimation and power, such as adaptive drug screening trials. Given the success of HM and the success of dose-finding methods such as the continual reassessment method, we consider a design combining the two methods to better motivate dose finding for a heterogeneous disease population, i.e. subpopulations. In oncology clinical trials, subpopulations can be defined by different standards of care or histologies. Current designs are inefficient or potentially deceiving. We propose a phase I Bayesian design that shares dose-response information across subgroups to improve and quicken dose finding within a subgroup, while allowing the flexibility to drop subgroups that find all doses overly toxic. Traditionally, patients are enrolled in cohorts and treated at the updated MTD. However to account for staggered enrollment between subgroups, we propose multiple guidelines for dose escalation within each subgroup. In a simulation study, we investigate three dose-response hierarchical models. For comparison, we investigate three corresponding saturated dose-response models that model each subgroup & dose-response independently.

email: **cunan001@umn.edu**

3m. DETECTING OUTLYING TRIALS IN NETWORK META-ANALYSIS

Jing Zhang*, University of Maryland

Haoda Fu, Eli Lilly and Company

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. However, some trials may appear to deviate markedly from the others, and thus be inappropriate to be synthesized in the NMA. In addition, the inclusion of these trials in evidence synthesis may lead to bias in estimation. We call such trials trial-level outliers. To the best of our knowledge, while heterogeneity and inconsistency in NMA have been extensively discussed and well addressed, few previous papers have considered the proper detection and handling of trial-level outliers. In this paper (poster), we propose several Bayesian outlier detection measures, which are then applied to a diabetes data set, and whose performance is evaluated through simulation studies.

email: **jingzhang2773691@gmail.com**

3n. SUBGROUP ANALYSIS IN CONFIRMATORY CLINICAL TRIALS

Brian Millen*, Eli Lilly and Company

The advent of personalized medicine has brought increased attention to the study of subpopulations in confirmatory clinical trials. Increasingly, exploratory subgroup analyses have the potential to influence regulatory recommendations on appropriate populations for treatment with medicines in review. The recent EMA draft guideline on Subgroup Analyses in Confirmatory Trials focuses on this issue. In addition, confirmatory subgroup analysis approaches are increasingly employed by sponsors interested in developing tailored therapies or personalized medicines. In this talk, we will discuss statistical and inferential considerations in these settings, along with thoughts on implications for drug development in the future

email: **millen_brian_a@lilly.com**

4. POSTERS: Survival Analyses

4a. TIME DEPENDENT COVARIATES IN THE PRESENCE OF LEFT TRUNCATION

Rebecca A. Betensky*, Harvard School of Public Health

A time varying marker process that is measured at study entry is problematic in the presence of left truncation. In this talk, we describe possible approaches to this problem and explain their drawbacks. We propose methods to appropriately handle this problem based on residuals of the marker process and alternative modeling of it. We present simulations and apply the methods to a longitudinal Alzheimer's disease study.

email: **betensky@hsph.harvard.edu**

4b. 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 provides simple but reliable procedures that yield asymptotically normal estimators whose covariance matrices can be easily estimated. 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**

4c. A MARTINGALE APPROACH TO ESTIMATING CONFIDENCE BAND WITH CENSORED DATA

Eun-Joo Lee*, Millikin University

Some non-parametric simultaneous confidence bands for survival function are developed when data are randomly censored on the right. To construct the confidence bands, a computer-assisted method is utilized and this approach requires no distributional assumptions, so the confidence bands can be easily estimated. To improve the estimation procedures for the finite sample sizes, the log-minus-log transformation is employed.

email: **elee@millikin.edu**

4d. NOVEL IMAGE MARKERS FOR NON-SMALL CELL LUNG CANCER CLASSIFICATION AND SURVIVAL PREDICTION

Hongyuan Wang*, University of Kentucky

Fuyong Xing, University of Florida

Hai Su, University of Florida

Arnold Stromberg, University of Kentucky

Lin Yang, University of Florida

Non-small cell lung cancer (NSCLC), the most common type of lung cancer, is one of serious diseases causing death for both men and women. Computer-aided diagnosis and survival prediction of NSCLC is of great importance in providing assistance to diagnosis and personalize therapy planning for lung cancer patients. In this presentation we would propose an integrated framework for NSCLC computer-aided diagnosis and survival analysis using novel image markers. The entire biomedical imaging informatics framework consists of cell detection, segmentation, classification, discovery of image markers, and survival analysis. After the extraction of a set of extensive cellular morphological features using efficient feature descriptors, eight different classification techniques that can handle high-dimensional data have been evaluated and then compared for computer-aided diagnosis. Moreover, a Cox proportional hazards model is fitted by component-wise likelihood based boosting. Significant image markers have been discovered using the bootstrap analysis and the survival prediction performance of the model is also evaluated. The proposed framework has been applied to a lung cancer dataset that contains 122 cases with complete clinical information. The classification performance exhibits high correlations between the discovered image markers and NSCLC subtypes. The survival analysis demonstrates strong prediction power of the discovered image markers.

email: **hongyuan.wang@uky.edu**

4e. GENERALIZED ESTIMATING EQUATIONS FOR MODELING RESTRICTED MEAN SURVIVAL TIME UNDER GENERAL CENSORING MECHANISMS

Xin Wang*, University of Michigan

Douglas E. Schaubel, University of Michigan

Restricted mean lifetime is often of great clinical interest in practice. Several existing methods involve explicitly projecting out patient-specific survival curves using parameters estimated through Cox regression. However, it would often be preferable to directly model the restricted mean, to yield more clinically meaningful treatment and covariate effects. We propose generalized estimating equation methods to relate restricted mean lifetime to baseline covariates. The proposed methods avoid potentially problematic distributional assumptions pertaining a restricted survival time, and allow for censoring to depend on time-dependent factors. Our methods are motivated by the desire to quantify the impact on pre-transplant survival of characteristics of end-stage liver disease (ESLD) patients wait listed for liver transplantation. This analysis requires accommodation for dependent censoring since pre-transplant survival is dependently censored by time-dependent factors due to the nature of the liver allocation system. Large sample properties of the proposed estimators are derived and simulation studies are conducted to assess their finite sample performance. We apply the proposed methods to model pre-transplant mortality among end-stage liver disease patients using national registry data.

email: **wangxinnju@gmail.com**

4f. GENERALIZED ACCELERATED FAILURE TIME SPATIAL FRAILTY MODEL

Haiming Zhou*, University of South Carolina

Timothy Hanson, University of South Carolina

Jiajia Zhang, University of South Carolina

Flexible incorporation of both geographical patterning and risk effects in cancer survival models is becoming increasingly important, due in part to the recent availability of large cancer registries. Most spatial survival models stochastically order survival curves from different subpopulations. However, it is common for survival curves from two subpopulations to cross in epidemiological cancer studies and thus interpretable standard survival models cannot be used without some modification. Common fixes are the inclusion of time-varying regression effects in the proportional hazards model or fully nonparametric modeling, either of which destroys any easy interpretability from the fitted model. To address this issue, we develop a generalized accelerated failure time model which is interpretable in terms of median regression and able to capture crossing survival curves in the presence of spatial correlation. An efficient Markov chain Monte Carlo algorithm is presented for posterior computation and an R package is developed to fit the model using complied C++. We apply our approach to a subset of the prostate cancer data gathered for Louisiana by the Surveillance, Epidemiology, and End Results program of the National Cancer Institute.

email: **zhouh@email.sc.edu**

4g. PENALIZED VARIABLE SELECTION IN COMPETING RISKS REGRESSION

Zhixuan Fu*, Yale University

Chirag R. Parikh, Yale University School of Medicine

Bingqing Zhou, Yale University

The penalized variable selection methods have been extensively studied for standard time-to-event data. Such methods, cannot be directly applied when subjects are at risk of several mutually exclusive events, known as competing risks. The proportional subdistribution hazard (PSH) model proposed by Fine and Gray has become a popular semi-parametric model for time-to-event data with competing risks. It allows for direct assessment of covariate effects on the cumulative incidence function. In this paper, we propose a general penalized variable selection strategy that simultaneously handles variable selection and parameter estimation in the PSH model. We rigorously establish general asymptotic properties for the proposed penalized estimators and present a numerical algorithm for implementing the variable selection procedure. Simulation studies are conducted to demonstrate the good performance of the proposed method. Diseased donor kidney transplant data from the United Network of Organ Sharing illustrate the utility of the proposed method.

email: **zhixuan.fu@yale.edu**

4h. STATISTICAL MODELING OF GAP TIMES IN PRESENCE OF PANEL COUNT DATA WITH INTERMITTENT EXAMINATION TIMES: AN APPLICATION TO SPONTANEOUS LABOR IN WOMEN

Ling Ma*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

In longitudinal studies of serial events where each subject may be observed only at several distinct and random observation times, only the numbers of occurrences of the events are known at the observation times. Data of this type are commonly referred to as panel count data. Most of the existing methods for panel count data focus on statistical inference for the point process while it is also of interest to make inference for the gap times between events. The application of interest in this project is to provide a framework for modeling gap-time distributions between cervical dilations in the first stage labor process (e.g. 3 cm dilation to 4 cm dilation). One well-known problem in obstetrics is that the start of labor is not clearly defined, and thus the benchmark reference time is often chosen as the end of the process (full dilation at 10 cm) and time is run backwards. We propose a parametric model for the gap times and use random effects to capture the correlation among gap times of the same subject and present a maximum likelihood approach for parameter estimation. We will investigate our proposed approach through simulations and show analysis of Collaborative Perinatal Project study data.

email: **mlbegood@gmail.com**

4i. COMPETING RISKS MODEL OF SCREENING AND SYMPTOMS DIAGNOSIS FOR PROSTATE CANCER

Sheng Qiu*, University of Michigan

Alexander Tsodikov, University of Michigan

Introduction of screening for prostate cancer using the prostate-specific antigen (PSA) marker of the disease around 1989 led to remarkable dynamics of the incidence of the disease observed in European countries. A competing risks model for cancer screening diagnosis and diagnosis due to symptoms is developed. The risks are driven by a latent process modeling tumor onset. Intensity of screening and hazard driving prostate cancer diagnosis in the absence of screening are estimated jointly and semi-parametrically using estimating equations and the NPMLE method. Examples using data from European cancer registries (EUREG) are illustrated.

email: **shqiu@umich.edu**

4j. JOINT MODELING OF RECURRENT EVENT PROCESSES AND INTERMITTENTLY OBSERVED TIME-VARYING BINARY COVARIATE PROCESSES

Shanshan Li*, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis

When conducting recurrent event data analysis, it is common to assume that the covariate processes are observed throughout the follow-up period. In most applications, however, the values of time-varying covariates are only observed periodically rather than continuously. A popular ad-hoc approach is to carry forward the last observed covariate value until it is measured again. This simple approach, however, usually leads to biased estimation. To tackle this problem, we propose to model the covariate effect on the risk of the recurrent events through jointly modeling the recurrent event process and the longitudinal measures. Despite its popularity, estimation of the joint model with binary longitudinal measurements remains a challenge, because the standard linear mixed effects model approach is not appropriate for binary measures. In this work, we postulate a Markov model for the binary covariate process and a random-effect proportional intensity model for the recurrent event process. We use a Markov chain Monte Carlo algorithm to estimate all the unknown parameters. The performance of the proposed estimator is evaluated via simulations. The methodology is applied to an observational study designed to evaluate the effect of Group A streptococcus (GAS) on pharyngitis among school children in India.

email: **sl50@iu.edu**

4k. COMPOSITE OUTCOMES VERSUS COMPETING RISKS

Paul Kolm*, Christiana Care Health Systems

Many randomized trials as well as observational comparative effectiveness studies analyze a composite outcome that includes several singular outcomes of interest. An example of a composite outcome often used in cardiovascular research includes death due to cardiovascular causes / myocardial infarction / stroke / rehospitalization for revascularization. The composite outcome is coded "yes" if any one of the outcomes occurs for a given patient. Although the study is usually powered on the composite outcome, separate analyses of the single outcomes are often made with the intent of determining the one that exerts the major influence on the composite outcome. This study compares analysis of a composite outcome with an analysis of the outcomes from a competing risks perspective with respect to regression coefficient estimates, standard errors, power and conclusions.

email: **pkolm@christianacare.org**

4l. QUANTILE REGRESSION MODELS FOR INTERVAL-CENSORED FAILURE TIME 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

Interval-censored case 2 failure time data arise frequently in longitudinal studies where the exact failure time cannot be determined but is known only to have occurred between two random observation times. We propose a quantile regression model to analyze interval-censored data since it relaxes the requirements on the error term and the coefficients are interpretable as direct regression effects on the failure time. It is assumed that the conditional quantile of failure time is a linear function of covariates and the failure time and observation time are conditional independent. 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 Atherosclerosis Risk in Communities Study.

email: **fou@unc.edu**

4m. EMPIRICAL LIKELIHOOD CONFIDENCE BANDS FOR THE DIFFERENCE OF SURVIVAL FUNCTIONS UNDER THE PROPORTIONAL HAZARDS MODEL

Mai Zhou, University of Kentucky

Shihong Zhu*, University of Kentucky

When comparing two treatments giving rise to censored time-to-event outcomes, the difference of two predicted individualized survival functions provides valuable information at the individual level about how the treatment effect evolves over time. In this manuscript, we propose a method to construct the simultaneous confidence band associated with the predicted difference by converting an Empirical Likelihood ratio test statistic. Simulation studies are conducted to demonstrate the superior coverage accuracy of the proposed confidence band over its existing competitor.

email: **szh224@uky.edu**

5. POSTERS: Causal Inference

5a. A CAUSAL FRAMEWORK FOR META ANALYSES

Michael E. Sobel*, Columbia University

David Madigan, Columbia University

Wei Wang*, Columbia University

We construct a framework for meta-analysis that helps to clarify and empirically examine the sources of between study heterogeneity in treatment effects. The key idea is to consider, for each of the treatments under investigation, the subject's potential outcome in each study were he to receive that treatment. We consider four sources of heterogeneity: 1) response inconsistency, whereby a subject's response to a given treatment varies across different studies, 2) the grouping of non-equivalent treatments, where two or more treatments are grouped and treated as a single treatment under the incorrect assumption that a subject's responses to the different treatments would be identical,

3) non-ignorable treatment assignment, and 4) response related variability in the composition of subjects in different studies. We then examine the implications of these assumptions for heterogeneity/homogeneity of conditional and unconditional treatment effects. To illustrate the utility of our approach, we re-analyze individual patient data from 29 randomized placebo controlled studies of Vioxx on the cardio-vascular risk of Vioxx, a Cox-2 selective non- steroidal anti-inflammatory drug approved by the FDA in 1999 for the management of pain and withdrawn from the market in 2004.

email: **mes105@columbia.edu**

5b. THE PRINCIPAL DIRECTION OF MEDIATION

Oliver Chen*, Johns Hopkins Bloomberg School of Public Health

Elizabeth Ogburn, Johns Hopkins Bloomberg School of Public Health

Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Brian Caffo, Johns Hopkins Bloomberg School of Public Health

Martin Lindquist, Johns Hopkins Bloomberg School of Public Health

Mediation analysis is often used in the behavioral sciences to investigate the role of intermediate variables that lie in the causal path between a randomized treatment and an outcome variable. However, little is known about mediation analysis when the intermediate variable (mediator) is a high dimensional vector. For example, in a functional magnetic resonance imaging (fMRI) study of thermal pain we are interested in determining whether brain measurements (over hundreds of thousands of voxels) mediate the relationship between the application of thermal pain and reported amount of perceived pain. To address the problem of high dimensional mediators, we propose a framework called the principal direction of mediation (PDM). This framework is philosophically similar to principal component analysis (PCA), but addresses a fundamentally different problem: the first principal direction of mediation is the linear combination of high dimensional potential mediators that is simultaneously most strongly predicted by the treatment and predictive of the outcome. We study this method using simulation and an application to data from an fMRI study of thermal pain.

email: **ychen219@jhu.edu**

5c. DYNAMIC MARGINAL STRUCTURAL MODELS TO TEST THE BENEFIT OF LUNG TRANSPLANTATION TREATMENT REGIMES

Jeffrey A. Boatman*, University of Minnesota

David M. Vock, University of Minnesota

Patients awaiting lung transplantation may confront a difficult decision if offered a low-quality organ: accept the organ or remain on the waiting list with the hope of receiving a better organ. Patients may have multiple opportunities to accept or decline a transplant, and organ assignment is not independent across subjects, but previous statistical methods to infer optimal treatment strategies do not fully account for these problems. To overcome these issues, we extend dynamic marginal structural models, estimated by inverse-of-probability-of-compliance weighted estimators, to estimate the survival benefit if a patient were to follow a certain rule on whether or not to accept an offered organ. Specifically we are interested in testing the survival benefit of declining organs below a certain threshold of donor quality. We developed an organ quality score based on a Cox regression model of post-transplant survival using donor characteristics as predictors, and we implement our proposed method using data from the United Network for Organ Sharing (UNOS) national registry of lung transplants. Our work may be easily extended to allow for time-varying strategies that accommodate patient condition and the prevalence of donor organs at a particular center.

email: **boat0036@umn.edu**

5d. A MODEL BASED APPROACH FOR PREDICTING PRINCIPAL STRATUM MEMBERSHIP IN ENVIRONMENTAL INTERVENTIONS

Katherine E. Freeland*, Johns Hopkins Bloomberg School of Public Health

Environmental interventions targeted at reducing indoor air pollution have shown promise as a method for improving respiratory health outcomes in children by reducing particulate matter (PM) in the home. However, in these interventions, it is difficult to determine the effect of reduced PM, a post-randomization variable, on the respiratory outcomes. Using principal stratification, a framework for calculating principal effects (i.e. effects within a stratum), we are able to measure the effect of reduced PM on respiratory

outcomes. These principal effects allow for the comparison of treatment effects for those who would and would not have seen a reduction in PM levels. With the PM reduction variable, we can identify principal strata membership for some individuals in the control and treatment groups. However, the observed data only allow us partial identification of strata for other individuals. We explore the use of various models to predict the partially identified individuals' strata and calculate "principal effects" based on predicted membership. The customary statistical uncertainty of these estimates was explored, along with additional variability introduced by the process of model selection and strata classification. A resampling based estimator of the principal effects was developed, accounting for the major sources of variability in this process.

email: **kfreela1@jhu.edu**

5e. PROPENSITY SCORE APPROACH TO MODELING MEDICAL COST USING OBSERVATIONAL DATA

Jiaqi Li*, University of Philadelphia

Nandita Mitra, University of Philadelphia

Elizabeth Handorf, Fox Chase Cancer Center

Justin Bekelman, University of Philadelphia

Medical cost estimation is vital to health economics evaluation and decision-making. Often it is not feasible to follow subjects for the full duration of interest so an individual's total medical costs are subject to non-independent right censoring. Moreover, medical costs are commonly right skewed. Therefore, standard regression models and survival analysis techniques are inadequate. Lin (2000) and Robins and Rotnitzky (1992) have developed linear regression and weighting techniques to model medical cost from trial data. Since medical costs are often collected in observational data, we develop propensity score (PS) methods to estimate costs that are adjusted for potential confounding inherent to observational studies. We compare common PS methods including generalized linear regression with gamma variances, stratification, inverse probability weighting (IPW) and doubly robust weighting. Specifically, for the IPW method, we develop a joint model with subject-specific random effects to account for possible correlation of PS estimates and the probability of observing complete costs. Large sample variances are derived using a general estimation equation (GEE) framework. These modeling approaches are applied to a cost analysis of two bladder cancer treatments, cystectomy versus bladder preservation therapy, using SEER-Medicare data.

email: **jiaqili0321@gmail.com**

5f. GENERALIZING EVIDENCE FROM RANDOMIZED TRIALS USING INVERSE PROBABILITY OF SELECTION WEIGHTS

Ashley L. Buchanan*, University of North Carolina, Chapel Hill

Michael G. Hudgens, University of North Carolina, Chapel Hill

Stephen R. Cole, University of North Carolina, Chapel Hill

Results obtained in randomized trials may not generalize to a target population. In a randomized trial, the treatment assignment mechanism is always known, but assuming participants are a random sample from the target population may be dubious. Lack of generalizability can arise when the distribution of treatment effect modifiers in trial participants is different from the distribution in the target population. We consider an inverse probability of selection weighted (IPSW) estimator for generalizing trial results to a target population. The IPSW estimator is shown to be consistent and asymptotically normal. Expressions for the asymptotic variance and a consistent sandwich-type estimator of the variance are derived. Simulation results comparing the IPSW estimator and a previously proposed stratified estimator show that the estimators perform similarly when the propensity score model included a binary covariate. However, with a continuous covariate in the propensity score model, the IPSW estimator is less biased and the corresponding Wald confidence intervals had better coverage. The IPSW estimator is employed to generalize results from the AIDS Clinical Trials Group to all people currently living with HIV in the U.S.

email: **abuchan@email.unc.edu edu**

5g. RACIAL DISPARITIES IN CANCER SURVIVAL: A CAUSAL INFERENCE PERSPECTIVE

Linda Valeri*, Harvard School of Public Health

Jarvis Chen, Harvard School of Public Health

Nancy Krieger, Harvard School of Public Health

Tyler J. VanderWeele, Harvard School of Public Health

Brent A. Coull, Harvard School of Public Health

The National Cancer Institute has identified the elimination of cancer health disparities as one of the most urgent goals for reducing disease burden in the US. Recent research has highlighted that disparities across racial/ethnic groups involve cancer etiology, incidence, screening, diagnosis, treatment, and survival. Quantifying the interplay of mediating factors across this continuum and informing targeted interventions is therefore a priority. In the present study we propose to estimate the disparity in cancer survival between Black and White individuals that would remain if the mediator distribution of the black population were set equal to that of the white population. We identify this causal estimand under the assumption of no unmeasured confounders of the mediator-survival relationship. We then develop sensitivity analysis techniques for violation of the unmeasured confounding assumption and for selection bias due to mediator missing not at random. The approaches are applied to SEER cancer registry data from 1992-2010. This work illustrates how a causal inference perspective aids in identifying and formalizing relevant hypotheses in health disparities research that can inform policy decisions.

email: **liv839@mail.harvard.edu**

6. POSTERS: Statistical Genetics, GWAS, and 'Omics Data

6a. A DATA-ADAPTIVE SNP-SET-BASED ASSOCIATION TEST OF LONGITUDINAL TRAITS

Yang Yang*, University of Texas Health Science Center at Houston

Peng Wei, University of Texas Health Science Center at Houston

Wei Pan, University of Minnesota

The current practice of single trait-single SNP analysis in genome-wide association studies (GWAS) is underpowered to detect the median-to-small effect sizes typically expected for common diseases. When multiple measurements of a trait at different time points are available, the longitudinal trait-multiple SNP analysis becomes a promising alternative. A longitudinal study may have greater power than a cross-sectional study, given the same or even smaller sample size. Multiple SNPs tend to reveal more information and render more robust signals than a single SNP. We extended an adaptive test, called adaptive sum of powered score (aSPU) test (Pan et al, Genetics 2014) and its variants (aSPU-weighted and aSPU-score) for cross-sectional trait analysis to longitudinal trait analysis in the framework of the generalized estimating equations (GEE). We investigated the performance of the aSPU test family in different scenarios, including different sample sizes, varying number of null SNPs and the presence of opposite directions of causal SNP effects. Through extensive simulation studies, we showed that the aSPU family was generally more powerful than several other commonly used methods, especially in the presence of many null SNPs. We demonstrated the utility and statistical efficiency gains of the proposed aSPU tests using the Atherosclerosis Risk in Communities (ARIC) data.

email: **xyy2006@msn.com**

6b. GENETIC ANALYSIS OF DATA FROM STRUCTURED POPULATIONS

Yogasudha Veturi*, University of Alabama at Birmingham

Gustavo de los Campos, University of Alabama at Birmingham

Human populations exhibit various degrees of stratification and admixture. In the analysis of genomic data, population stratification is usually treated as a nuisance. Consequently, both in Genome Wide Association Studies (GWAS) and Whole Genome Regression (WGR) a common approach has been to "correct" for population structure by adding marker-derived principal components as fixed effects. However, this approach induces a mean correction that does not consider the possibility that marker effects vary across sub-populations. In

this study, we propose ways of dealing with stratification that incorporate heterogeneity explicitly using interaction models and the bivariate Genomic Best Linear Unbiased Predictor (G-BLUP). These approaches allow for the analysis of data from two or more groups jointly, provide group-specific marker effects, estimates of variances and between-group correlations. We applied the proposed methods to study genomic differences/similarities between clusters obtained from a multi-racial human population (Multi-Ethnic Study of Atherosclerosis) for height, high-density lipoprotein (HDL) and low-density lipoprotein (LDL). Our estimates of genomic heritability varied not only across traits but also across groups, and our estimates of genomic correlations ranged from low (0.3-0.4) to moderate high (0.5-0.6) providing evidence of great extents of genetic heterogeneity.

email: **sveturi@uab.edu**

6c. MAPPING DISEASE SUSCEPTIBILITY LOCI FOR MULTIPLE COMPLEX TRAITS WITH U-STATISTICS

Ming Li*, University of Arkansas for Medical Sciences

Changshuai Wei, University of North Texas

Qing Lu, Michigan State University

Many complex diseases, particularly psychiatric and behavioral disorders, are supposed to be multi-dimensional with various aspects that are physical, behavioral and psychological. While it remains a great challenge to find a unified measurement to characterize a disease, a number of phenotypic traits are usually jointly used. Testing these phenotypic traits simultaneously is advantageous to take the disease heterogeneity into account, and improve the discovery process of identifying causal genetic variants, especially those pleiotropic variants associated with multiple traits. Furthermore, complex diseases are caused by the interplay of multiple genetic variants through complicated mechanisms. Multi-locus-based approaches, which take the possible genetic interactions into account, are highly desired in genetic association studies. The existing multi-trait-based approaches are commonly single-locus-based, and are proposed for family-based association studies. In this article, we propose a multi-locus, multi-trait approach for population-based association studies. Through simulations, we demonstrated that testing multiple traits simultaneously was more powerful than testing one single trait at a time. We also illustrated the proposed approach with an application to Nicotine Dependence. The joint analysis of three traits simultaneously identified SNPs with a significant association, which was replicable across studies.

email: **mli@uams.edu**

6d. PERMUTATION-BASED TEST STATISTICS FOR INTERMEDIATE PHENOTYPES IN GENOME-WIDE ASSOCIATION STUDIES

Wei Xue*, University of North Carolina, Chapel Hill

Eric Bair, University of North Carolina, Chapel Hill

In case-control genome-wide association studies, one may wish to identify genetic markers associated with intermediate phenotypes that are correlated with case status. In such cases, naive regression methods that ignore case-control design will produce biased estimates. This may be corrected by using methods such as inverse probability weighting (IPW), which assigns weights to the observations to correct for the fact cases are overrepresented in a case-control study. However, IPW regression coefficient estimates may be unreliable when evaluating the association between genetic markers and intermediate phenotypes that are strongly associated with case status. In a case-control study of temporomandibular disorder (TMD), we may wish to identify markers associated with the severity of orofacial pain. Nearly all controls will report no orofacial pain, which causes IPW regression to produce inaccurate results. We propose a novel permutation-based method and compared it with IPW. Simulations indicate that whereas IPW produces inflated type I error rates, our method produces correct type I error rates with no loss in power. We then apply this method to identify SNPs associated with the severity of orofacial pain using data from OPPERA study, a large-scale case-control study of TMD. We identify two novel SNPs strongly associated with pain severity.

email: **xuew@live.unc.edu**

6e. STATISTICS FOR GENETIC ASSOCIATION IN THE PRESENCE OF COVARIATES–GENOME SCANNING CONSIDERATIONS

Hui-Min Lin*, University of Pittsburgh

Eleanor Feingold, University of Pittsburgh

Yan Lin, University of Pittsburgh

A number of different statistics are available for genetic association analysis in the presence of covariates. In the context of a genome-wide association study, hundreds of thousands to millions of SNPs are tested, and whatever covariate model we specify is likely to be imperfect. In addition, the results of the study often focus on the list of SNPs ordered according to the statistics rather than on certain p-value cutoffs. Therefore, it is important to investigate the behavior of extreme values of the statistics rather than the behavior of the expected values. Gail et al. (2008) discussed this issue and proposed "detection probability" and "proportion positive" to measure the success (power) of a genomic study when ranked lists are the primary outcome. In theory, the ranked lists can be dominated by SNPs with misfit models rather than by true positive results. We are conducting a comprehensive comparative study to investigate the behavior of different association statistics that model covariates. We evaluate the statistics from the perspective of which statistics can provide robust ranked lists of "top hits." These are not necessarily the same statistics that have the highest power in a conventional single-test context.

email: **hul27@pitt.edu**

6f. POWER AND SAMPLE SIZE DETERMINATION FOR TIME COURSE MICROARRAY DIFFERENTIAL EXPRESSION STUDIES: A FALSE DISCOVERY RATE AND PERMUTATION-BASED SIMULATION METHOD

Joanne C. Beer*, University of Pittsburgh

Thuan Nguyen, Oregon Health & Science University

Kemal Sonmez, Oregon Health & Science University

Dongseok Choi, Oregon Health & Science University

Microarray experiments allow researchers to assess levels of gene expression for thousands of genes at a time. A frequent goal of microarray experiments is to identify genes which are differentially expressed across various biological conditions. Several methods have been developed for determining sample size for differential expression microarray experiments, but few methods have been extended to time course experiments in which gene expression is measured over a series of time points. We propose a flexible method for sample size and power analysis of time course microarray experiments using a positive false discovery rate type I error control. Because microarray data often deviate from the assumption of normality underlying the use of parametric t-tests and F-tests, and since it has been increasingly recognized that accounting for the correlation structure of gene expression data is important for accurately estimating error rate and sample size, the method relies on a permutation-based null distribution for the test statistics. We compare results of simulation-based sample size and power calculations to those of other published sample size methods for both static and time course microarray experiments.

email: **joannecbeer@gmail.com**

6g. FUNCTIONAL RANDOM FIELD MODELS FOR ASSOCIATION ANALYSIS OF SEQUENCING DATA

Xiaoxi Shen*, Michigan State University

Ming Li, University of Arkansas for Medical Sciences

Zihuai He, University of Michigan

Qing Lu, Michigan State University

Generalized Genetic Random Field (GGRF) model holds many nice properties for small sample-size sequencing studies, and has well-controlled type I error and high power as compared with existing methods. It, however, needs to specify a weight function to consider rare variants, and models only pairwise linkage disequilibrium (LD). To further improve the method, we propose a functional random field model (FRF) for association analysis of sequencing data. By fitting a functional curve on the genotypes of genetic region for each individual, we are able to incorporate high-order LD information into the association analysis. Moreover, because it models sequencing data on the individual level rather than the population level, it does not require a weight function for considering rare variants. We compare type I error and power of FRF with shoes of GGRF, SKAT and Burden test.

Our preliminary findings show that FRF outperform the other methods, especially when the weight function is misspecified. Additional findings also suggest FRF has the advantage over existing methods when genetic effects are bi-direction and when missing genotype data is present.

email: **shenxia4@stt.msu.edu**

6h. QUANTIFYING UNCERTAINTY IN THE IDENTIFICATION OF PROTEINS, POST-TRANSLATIONAL MODIFICATIONS (PTMs) AND PROTEOFORMS

Naomi C. Brownstein*, Florida State University National High Magnetic Field Lab

Xibei Dang, Florida State University National High Magnetic Field Lab

Eric Bair, University of North Carolina, Chapel Hill

Nicolas L. Young, Florida State University National High Magnetic Field Lab

The traditional goals of top-down proteomics are protein identification and quantitation. However, the presence of additional sources of variability, such as post-translational modifications (PTMs) and genomic variants, complicates the problem of identification. Recent interest in the proteomics community has begun to shift from the relatively narrow problem of protein identification to consideration of these additional sources of variability. Combining these factors results in a unique exhaustively defined chemical species termed "proteoform". We explore a variety of scoring metrics and estimate their uncertainty via bootstrapping. We demonstrate the method using human histone H4 and the corresponding proteoforms. Results show that related proteoforms may be statistically difficult to differentiate.

email: **nbrownstein@magnet.fsu.edu**

6i. A STATISTICAL PIPELINE FOR STUDYING CO-REGULATED GENES USING SINGLE-CELL RNA-seq DATA

Ning Leng*, Morgridge Institute for Research

Li-Fang Chu, Morgridge Institute for Research

Yuan Li, University of Wisconsin, Madison

Peng Jiang, Morgridge Institute for Research

Chris Barry, Morgridge Institute for Research

Ron Stewart, Morgridge Institute for Research

James Thomson, Morgridge Institute for Research

Christina Kendziorski, University of Wisconsin, Madison

Recent advances in single-cell RNA-seq technology enable investigators to conduct transcriptome-wide gene expression studies at the single-cell level. Such studies serve as a revolutionary tool to understand cell-to-cell variation within and among cell populations. A number of robust statistical methods are available for quality control and analysis of bulk RNA-seq data. However, features common to single-cell data (increased dropout and heterogeneity, for example) prohibit direct application to the single-cell setting. We here propose a statistical pipeline for studying co-regulated genes using single cell RNA-seq data. We applied our pipeline on expression profiles from 73 undifferentiated human embryonic stem cells (hESCs), and compared it with a naïve approach based on correlation analysis. Results demonstrate that the naïve approach is largely affected by technical noise and is unable to identify much of the biological heterogeneity that is present. On the other hand, our pipeline was able to accommodate technical noise and in so doing reveal co-regulation features of cell cycle markers in the undifferentiated hESC population.

email: **nleng@wisc.edu**

6j. OUTLIER DETECTION FOR QUALITY CONTROL IN FLOW CYTOMETRY USING COMPOSITIONAL DATA ANALYSIS

Kipper Fletez-Brant*, Johns Hopkins University

Josef Spidlen, BC Cancer Agency

Ryan Brinkman, BC Cancer Agency

Pratip Chattopadhyay, National Institutes of Health

Flow cytometry experiments collect observations for N variables on C cells, with C >> N for a single experiment. Current technology allows for hundreds of experiments to be performed per day, and each experiment can have errors in sample acquisition, measurement or machine malfuntion. This can result in inaccurate observations for some, but not all, cells in an experiment. Trying

to perform manual quality control on flow cytometry is not possible for more than a handful of experiments. We have developed a method to automate quality control that uses compositional data analysis. We model a flow cytometry experiment as a set of compositions of cell populations observed over time. Each population is defined as a cell having observations above or below some threshold for each of the N variables. We derive a summary statistic for each composition which reflects the distribution of cell populations represented in it. We use this statistic to partition the data in a flow experiment into "good" and "bad" data using changepoint analysis. This statistic allows our method to take advantage of the multivariate nature of flow cytometry data, and is reasonably fast.

email: **cafletezbrant@gmail.com**

6k. POWER ANALYSIS FOR GENOME-WIDE ASSOCIATION STUDY IN BIOMARKER DISCOVERY

Wenfei Zhang*, Sanofi

Yuefeng Lu, Sanofi

Yang Zhao, Sanofi

Vincent Thuillier, Sanofi

Jeffrey Palmer, Sanofi

Sherry Cao, Sanofi

Jike Cui, Sanofi

Stephen Madden, Sanofi

Srinivas Shankara, Sanofi

Biomarker discovery is important for disease diagnosis, prognosis, and risk prediction in drug discovery and development. Genetic markers are considered as promising biomarkers that have drawn great attentions in discovery research. With recent technological innovations, genome-wide association studies (GWAS) have become possible and revolutionized the research for genetic markers underlying human complex diseases. Power analysis plays a significant role in GWAS by optimizing both practical costs and statistical validities. However, existing methods usually fail to consider the complicated correlation patterns among genetic markers. Therefore, we propose a re-sampling based power analysis method to properly address the correlations among genetic markers. Our analysis shows the results on the statistical powers under varrious odd ratios and allele frequencies.

email: **wenfei.zhang@sanofi.com**

6l. DIFFERENTIAL DYNAMICS IN SINGLE-CELL RNA-Seq EXPERIMENTS

Keegan D. Korthauer*, University of Wisconsin, Madison

Christina Kendziorski, University of Wisconsin, Madison

Measurements of genome-wide RNA transcript abundance at the single-cell level allow us to answer scientific questions that were elusive with traditional bulk data, which only provided averages across large pools of cells. Specifically, it is now clear that transcription often occurs in a bursty manner, resulting in multi-modal distributions within gene (with individual cells that are off, on at a low level, and on at a high level, for example). Identifying such genes and using them to characterize subgroups within and across biological conditions is an important first step in many single-cell RNA-seq experiments. Toward this end, we have developed a Dirichlet process mixture model based approach. The approach facilitates the identification of multi-modal genes, uses these genes to identify subgroups, and allows for the identification of differential dynamics (differential expression, differential dropout, differential proportions within modal groups) across multiple biological conditions. Advantages are demonstrated via simulation and case studies.

email: **kdkorthauer@wisc.edu**

6m. EXPERIMENTAL DESIGN FOR BULK SINGLE-CELL RNA-Seq STUDIES

Rhonda L. Bacher*, University of Wisconsin, Madison

Christina Kendziorski, University of Wisconsin, Madison

Studies of isoform expression are critical to understanding phenotypic complexity as they potentially reveal information not detectable using gene level estimates alone. With sequencing costs continually decreasing, utilizing this information has become popular in bulk RNA-Seq experiments and we expect will become popular in single-cell RNA-seq experiments as well. A few studies have investigated the trade-off between sequencing depth and sample size both for bulk and single-cell RNA-seq experiments probing gene level expression, but no guidelines are available when isoform

expression is of interest. To address this, we have developed an approach for simulating bulk and single-cell RNA-seq data at the gene and isoform level. The approach, called ReadSim, is used to assess the question of depth vs. sample size for a number of RNA-seq experimental designs. General guidelines are proposed.

email: **Rbacher@wisc.edu**

6n. A HIERARCHICAL MIXTURE MODEL FOR JOINT PRIORITIZATION OF GWAS RESULTS FROM MULTIPLE RELATED PHENOTYPES

Cong Li*, Yale University

Can Yang, Hong Kong Baptist University

Hongyu Zhao, Yale School of Public Health

The past ten years have witnessed a grand wave of endeavors hunting single nucleotide polymorphisms (SNPs) affecting various human complex traits through genome-wide association studies (GWAS). But disappointedly, the significant SNPs identified through GWAS can only explain a small fraction of the genetic contributions to complex traits. Many lines of evidence suggest the major reason being the existence of numerous weak-effect SNPs that are difficult to identify under the current GWAS sample sizes. In our previous work, a statistical method called "GPA" was developed to improve our power to detect these weak-effect SNPs by borrowing information from GWAS results of genetically related traits and genomic functional annotations. Despite its success, it is challenging for GPA to handle more than three phenotypes simultaneously in its current form, limiting its applications in practice. To address this limitation, we have reformulated the GPA model by adding a hierarchical prior on the association status matrix. A low-rank structure is imposed on the logit transformation of the prior matrix to encourage the correlation across multiple phenotypes. Through both simulations and real data applications, we have shown that our method can effective integrate multiple related phenotypes and boost the power of detecting associated SNPs.

email**: licong.jason@gmail.com**

6o. NONPARAMETRIC TESTS FOR DIFFERENTIAL ENRICHMENT ANALYSIS WITH MULTI-SAMPLE ChIP-Seq DATA

Qian Wu*, BioStat Solution

Kyoung-Jae Won, University of Pennsylvania

Hongzhe Li, University of Pennsylvania

Chromatin immunoprecipitation sequencing (ChIP-seq) technology is a powerful tool for analyzing protein interactions with DNA. The genes with differential binding region under two or multiple conditions are important and could be used to predict the gene expression changes. In the previous study, we proposed a kernel based nonparametric method ChIPtest to solve this problem under two conditions. In this article, we develop a new nonparametric testing method NonpChIP without considering any smoothing method, where the test statistics will not depend on the choice of bandwidth as ChIPtest. In addition, the methods are limited for detecting genes with differential binding region among more than two conditions, such as for multiple time-course ChIP-Seq data. We further investigate the time-course changes of the genes between four time points by defining multivariate test statistics as the mean (TSmean) or maximum (TSmax) of three adjacent pair-wise ChIPtest statistics. This new method provided the variance estimation under the assumption of equal or unequal error variance. We compared the performance of ChIPtest and NonpChIP via ROC curves and True Positive Rate (TPR) curves in both two conditions and multiple conditions. Both real data and simulation results shows the TSmax with kernel smoothing dominate the other methods. All these results indicate that the identified differential binding regions are indeed biologically meaningful. 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 in gene promoter regions between proliferating preadipocytes and mature adipocytes in murine 3T3-L1 cells. 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**

6p. ANALYSIS OF MASS SPECTROMETRY DATA AND PREPROCCESING METHODS FOR METABOLOMICS

Leslie Myint*, Johns Hopkins University

Kasper Hansen, Johns Hopkins University

The genetic basis of disease has been a popular source of scientific inquiry in recent years due to the rapid increase in efficiency of sequencing technologies. However, disruptions in biochemical pathways and metabolite concentrations are also key factors in disease etiology and progression, and growing numbers of researchers have started performing metabolic profiling to shed light on their biological condition of interest. Metabolomics analysis typically involves separating the compounds present in biological samples through some form of chromatography and fragmenting the compounds into distinctive patterns using a mass spectrometer. This process creates noisy and complex data, which must be processed in several ways before subsequent analysis. This processing should ideally produce a list of metabolites and their abundances for all samples, which will ultimately be used to identify the metabolites that are differentially abundant between groups. However, this preprocessing stage is a difficult task that has not been extensively explored. We identified interesting features of mass spectrometry data and investigated their relationship with and impact on the results of the widely used preprocessing software package, XCMS. We evaluate the impact of the change in results on differential abundance analysis.

email: **leslie.myint@gmail.com**

6q. ACCOUNTING FOR MEASUREMENT ERROR IN GENOMIC DATA AND MISCLASSIFICATION OF SUBTYPES IN THE ANALYSIS OF HETEROGENEOUS TUMOR DATA

Daniel Nevo, Hebrew University, Jerusalem, Israel

David Zucker*, Hebrew University, Jerusalem, Israel

Molin Wang, Harvard School of Public Health

Donna Spiegelman, Harvard School of Public Health

A common paradigm in dealing with heterogeneity across tumors in cancer analysis is to cluster the tumors according to subtypes using gene expression data on the tumor and then to analyze each of the clusters separately. A more specific target is to investigate the association between risk factors and specific subtype and to utilize the results for personalized treatment. This task is usually carried out by two steps – clustering and risk factor assessment. However, two sources of measurement error arise in these problems. The first is the error in the gene expression measurements. The second is the misclassification error when inaccurately assigning observations to clusters. We consider the case with a specified set of relevant genes and propose unified single- likelihood approach for normally distributed gene expressions. As an alternative, we consider a two-step procedure with the tumor type misclassification error taken into account in the second-step risk factor analysis. We describe our method for multinomial data and also for survival analysis data using a modified version of the Cox model. The results of a simulation study indicate that our methods significantly lower the bias with a small price being paid in terms of variance. We also analyze breast cancer data from the Nurses' Health Study to demonstrate the utility of our method.

email: **david.zucker@mail.huji.ac.il**

7. POSTERS: Methodology and Applications in Epidemiology, Environment, and Ecology

7a. CARPE DIEM! BIOSTATISTICIANS IMPACTING THE CONDUCTING AND REPORTING OF CLINICAL STUDIES

Sally Morton*, University of Pittsburgh

Standards for conducting and guidelines for reporting clinical studies have evolved and proliferated. Inconsistency, and in some cases disagreement, may be due to differences in whether or not the standard or guideline is meant as a best practice recommendation or a mandatory requirement. In addition, the targeted study design may play a role (for example, randomized trials versus observational studies). In this poster, we compare and contrast standards and guidelines using, as examples, the Patient-Centered Outcomes Research Institute (PCORI) standards for conducting patient-centered comparative effectiveness research, as well as standards and guidelines for systematic

reviews. We will consider the potential impact on innovation and propose that standards and guidelines provide potential tools of influence and education for the discipline. Biostatisticians can, and should, play an important role in the process of constructing, validating, and disseminating standards and guidelines.

email: **scmorton@pitt.edu**

7b. ON STRATIFIED BIVARIATE RANKED SET SAMPLING WITH OPTIMAL ALLOCATION FOR NAIVE AND RATIO ESTIMATORS

Lili Yu, Georgia Southern University

Hani Samawi, Georgia Southern University

Daniel Linder, Georgia Southern University

Arpita Chatterjee, Georgia Southern University

Yisong Huang*, Georgia Southern University

Robert Vogel, Georgia Southern University

The purpose of the current work is to introduce stratified bivariate ranked set sampling (SBVRSS) and investigate its performance for estimating the population means using naive and ratio methods. The properties of the proposed estimator are derived along with the optimal allocation with respect to stratification. We conduct a simulation study to demonstrate the relative efficiency of SBVRSS as compared to stratified bivariate simple random sampling (SBVSRS) for ratio estimation. Data that consist of weights and bilirubin levels in the blood of 120 babies used to illustrate the procedure on a real data set, with our results indicating that SBVRSS for ratio estimation is more efficient than using SBVSRS in all cases presented in the simulations.

email: **yh00049@georgiasouthern.edu**

7c. COMPARISONS OF THE CANCER RISK ESTIMATES BETWEEN EXCESS RELATIVE RISK AND RELATIVE RISK MODELS: A CASE STUDY

Shu-Yi Lin*, Taipei City Hospital, Taiwan

Relative risk models (RR Model) are commonly used in cancer risk assessment in public health. However, in radiation research, linear relative risk model is often used to estimate excess relative risk (ERR Model), such as the studies of Japanese atomic bomb survivors. The purpose of this study is to compare the estimates of the cancer risks between the ERR models and the RR models using the Taiwan radiation-contaminated buildings cohort follow-up data from 1983 to 2005. The analyses were based on 6,242 subjects who had ever lived in radio-contaminated buildings, and 117 cancer cases were identified. The study compares and assesses the estimates of cancer risks by Cox Proportional Hazard Model, Poisson Log-linear Model and ERR model. The study verifies that the excess relative risk estimated by ERR model are equivalent to the relative risk minus one estimated by Poisson models. Our analysis shows that the results by Cox model (hazard ratios) are more conservative than those by ERR model and Poisson model (rate ratios). The relative risks estimated by the models using attained age or the ones using time from exposure to event as the time scales were similar. Adjusting for different covariates does not change the estimates substantially.

email: **A3810@tpech.gov.tw**

7d. A REGRESSION BASED SPATIAL CAPTURE-RECAPTURE MODEL FOR ESTIMATING SPECIES DENSITY

Purna S. Gamage*, Texas Tech University

Souparno Ghosh, Texas Tech University

Philip S. Gipson, Texas Tech University

Gregory Pavur, Texas Tech University

Data obtained from capture-recapture studies are essentially spatial in nature. The spatial proximity of the activity centre, of an animal, and the trap location determines how likely the concerned individual will be captured. In order to incorporate the spatial information in the inference about the relative abundance of a species in the study region, Borchers and Efford (2008) proposed the spatially explicit capture-recapture (SECR) model. In its original form, SECR allowed the state-space of the activity centers of the individuals to arise from a non-homogeneous Poisson process (NHPP). However, in practice, complete spatial randomness (CSR) is generally assumed for the distribution of the activity centers. However, in many situations, covariates, such as vegetation characteristics, heavily influence the location of these activity centers. To accommodate such information, we envision an NHPP, with covariate dependent

intensity function, driving the location of the activity centers in the study region. We perform simulation studies to compare the robustness of CSR and NHPP specifications of the state-space, particularly under model-misspecification. We then illustrate our methodology on the abundance data obtained during a survey of the mesocarnivores in north-west Texas.

email: **purna.s.gamage@ttu.edu**

7e. APPLICATION OF THE USE OF PERCENTAGE DIFFERENCE FROM MEDIAN BMI TO OVERCOME CEILING EFFECTS IN ADIPOSITY CHANGE IN CHILDREN

Christa Lilly*, West Virginia University

Lesley Cottrell, West Virginia University

Karen Northrup, Wood County School System

Richard Wittberg, Wood County School System

Researchers are alert to the needs of morbidly obese children given the increasing incidence of obesity. However, concern arose over the sensitivity and power of select adiposity measures and analytic approaches for examining change in the top percentiles of BMI. In the present study, standard statistical techniques found no differences when assessing 4,058 students using both weight and BMI percentile outcomes in morbidly obese children with Acanthosis Nigricans (AN). The children's sex-age-specific median BMI was used to calculate the percentage difference from BMI (BMI%). Kruskal-Wallis one-way ANOVA by ranks test determined differences of change scores among 4 groups of students (those who gained, lost, maintained or never had AN). Significant effects were found with BMI% but not in the weight or BMI percentile outcomes. Findings provide evidence supporting the use of BMI% rather than BMI and weight percentiles when examining adiposity change among children with morbid obesity.

email: **cice@hsc.wvu.edu**

7f. A MULTI-PATHOGEN HIERARCHICAL BAYESIAN MODEL FOR SPATIO-TEMPORAL TRANSMISSION OF HAND, FOOT AND MOUTH DISEASE

Xueying Tang*, University of Florida

Nikolay Bliznyuk, University of Florida

Yang Yang, University of Florida

Ira Longini, University of Florida

Mathematical modeling of infectious diseases plays an important role in the development and evaluation of intervention plans. These plans, such as the development of vaccines, are usually pathogen-specific, but laboratory confirmation of all pathogen-specific infections is rarely available. If an epidemic is a consequence of co-circulation of several pathogens, it is desirable to jointly model these pathogens in order to study the transmissibility of the disease. Our work is motivated by the hand, foot and mouth disease (HFMD) surveillance data in China from 2008 to 2009. The data set consists of counts of reported cases in 334 prefectures and 53 consecutive weeks and the laboratory test data for a small subset of the reported cases. We build a hierarchical Bayesian multi-pathogen model by using a latent process to link the disease counts and the lab test data. Our model explicitly accounts for spatio-temporal disease patterns. The inference and prediction are carried out by a computationally tractable MCMC algorithm. We study operating characteristics of the algorithm on simulated data and apply it to the HFMD in China data set.

email: **xytang@ufl.edu**

7g. EVALUATING RISK-PREDICTION MODELS USING DATA FROM ELECTRONIC HEALTH RECORDS

Le Wang*, University of Pennsylvania

Pamela A. Shaw, University of Pennsylvania

Hansie Mathelier, University of Pennsylvania

Stephen E. Kimmel, University of Pennsylvania

Benjamin French, University of Pennsylvania

Currently, there is particular clinical and economic interest in developing and evaluating models that predict adverse events (e.g., short-term hospital readmission) among patients with chronic diseases (e.g., heart failure). Accurate risk-prediction models can be used to inform personalized treatment strategies for individual patients. As interest in individualized prediction has grown, so too has the availability of large-scale clinical information systems. The increasing availability of data from electronic health records facilitates the development

of prediction models, but estimation of prediction accuracy could be limited by outcome misclassification, which can arise if events are not captured by the electronic system. In simulation studies, we evaluate the performance of receiver operating characteristic curves and risk-reclassification methods in the presence of outcome misclassification. We consider situations in which events are not included in the electronic health record, with and without dependence on covariate values. We illustrate the impact of outcome misclassification on estimation of prediction accuracy using data from the University of Pennsylvania Health System electronic health record to evaluate alternative prognostic models for 30-day readmission among patients with a diagnosis of heart failure.

email: **lwang0217@gmail.com**

7h. A BAYESIAN MODEL FOR IDENTIFYING AND PREDICTING THE SPATIO-TEMPORAL DYNAMICS OF RE-EMERGING URBAN INSECT INFESTATIONS

Erica Billig*, University of Pennsylvania

Michael Levy, University of Pennsylvania

Michelle Ross, University of Pennsylvania

Jason Roy, University of Pennsylvania

Analyses of epidemics are complicated by several factors, including the fact that the true dispersal mechanism of disease agents and the precise infection times of patients are often unknown. Instead, we often observe the infection state of each unit at discrete time intervals. For example, consider a recent study of the Chagas disease vector Triatoma infestans in Arequipa, Peru. The fact that the epidemic is slow-moving and the counts of infested houses are small leads to analytic challenges. The data are limited to observed vector presence at each household at three time points over several years. In addition, streets are major barriers to T. infestans movement, resulting in a complex, spatial structure of the epidemic. To address these challenges, we propose a susceptible-infected-observed-removed model that uses informative priors and a novel spatial function that incorporates the complex dispersal dynamics observed in Arequipa. The fully Bayesian method is used to augment the data, estimate the dispersal parameters, and determine posterior infestation risk probabilities of households for future treatment. We investigate the properties of the model with simulation studies. Finally, the proposed methods are illustrated with an analysis of the Chagas disease vector data.

email: **ebillig@mail.med.upenn.edu**

7i. SEMI-MARKOV MODELS FOR INTERVAL CENSORED TRANSIENT COGNITIVE STATES WITH BACK TRANSITIONS AND A COMPETING RISK

Shaoceng Wei*, University of Kentucky

Richard Kryscio, University of Kentucky

Continuous-time multi-state stochastic processes are useful for modeling the flow of subjects from intact cognition to dementia with mild cognitive impairment and global impairment as intervening transient, cognitive states and death as a competing risk (Figure 1). Each subject's cognition is assessed periodically resulting in interval censoring for the cognitive states while death without dementia is not interval censored. We apply a Semi-Markov process in which we assume that the waiting times are Weibull distributed except for transitions from the baseline state, which are exponentially distributed and in which we assume no additional changes in cognition occur between two assessments. We apply our model to the Nun Study.

email: **swe225@uky.edu**

7j. GROWTH CURVES FOR CYSTIC FIBROSIS INFANTS VARY IN THE ABILITY TO PREDICT LUNG FUNCTION

Yumei Cao*, Medical College of Wisconsin

Raymond G. Hoffmann, Medical College of Wisconsin

Evans M. Machogu, Indiana University School of Medicine

Praveen S. Goday, Medical College of Wisconsin

Pippa M. Simpson, Medical College of Wisconsin

Introduction. Monitoring children routinely is especially important in the early years of life and for patients with chronic illness. The Centers for Disease and Control (CDC) Growth charts have been typically used for this purpose but now the World Health Organization (WHO) charts are recommended for infants 0-24 months and the CDC charts after that

age. However, the charts do not match at 24 months. Studies have been conducted setting goals using the CDC charts for all ages. Our aim was to show how (1) the charts differed, (2) they might be reconciled to track cystic fibrosis patients and (3) previous CDC growth goals for CF patients needed modification to account for the differences from WHO parameters. Methods. CF registry data for patients born 2001-2004 were used. Bland Altman plots were used to compare CDC and WHO growth parameters. In addition, the ability to predict good lung function at 6 years based on the different growth measures at 2 years of age were compared using generalized linear models. Results. There is a considerable difference among the different measures for the CF patients. However, the ability to predict is adequate for all measures.

email: **yucao@mcw.edu**

7k. AN EXAMINATION OF THE CONCEPT OF FRAILTY IN THE ELDERLY

Felicia R. Griffin*, Florida State University

Daniel L. McGee, Florida State University

Elizabeth H. Slate, Florida State University

Frailty has been defined as a state of increased vulnerability to adverse outcomes. The concept of frailty has been centered around counting the number of deficits in health, which can be diseases, disabilities, or symptoms. However, there is no consensus on how it should be quantified. Frailty has been considered synonymous with functional status and comorbidity, but these may be distinct concepts requiring different management. We compared two methods of defining a frailty phenotype, a count of deficits and a weighted score of health deficits incorporating the strength of association between each deficit and mortality. The strength of association was estimated using proportional hazards coefficients. The study uses data from the NHANES III. We compared the two methodologies: frailty was associated with age, gender, ethnicity, and having comorbid chronic diseases. This study introduces a weighted score for defining a frailty phenotype that is more strongly predictive of mortality, and has potential to improve targeting and care of today's elderly.

email: **fgriffin@stat.fsu.edu**

7l. EFFICIENCIES FROM USING ENTIRE UNITED STATES RESPONSES IN PREDICTING COUNTY LEVEL SMOKING RATES FOR WEST VIRGINIA USING PUBLICLY AVAILABLE DATA

Dustin M. Long*, West Virginia University

Emily A. Sasala, West Virginia University

Smoking rates, as well as other risk factors, tend to vary geographically, specifically by county within each state. The Behavioral Risk Factor Surveillance System (BRFSS) collects data from across the United States on different risk factors, including smoking. However, many counties are not represented in BRFSS reports due to small populations or low number of responses. Thus, missing counties smoking rates must be predicted using some modeling scheme. West Virginia's between county smoking rates have high variability and contains a high percentage (20%) of missing counties. Using other publicly available county level variables from 2010 as covariates, two modeling frameworks, a generalized linear model and a Bayesian model, were constructed to predict smoking rates for counties without 2010 BRFSS estimates in West Virginia. These models used only West Virginia data to predict the missing values in addition to using the entire US data. We found that the data using the entire US was more efficient, i.e., stronger prediction in both types of model and better overall convergence in the Bayesian model.

email: **dmlong@hsc.wvu.edu**

7m. OPTIMALLY COMBINED ESTIMATION FOR TAIL QUANTILE REGRESSION

Kehui Wang*, North Carolina State University

Huixia Judy Wang, The George Washington University

Quantile regression offers a convenient tool to access the relationship between a response and covariates in a comprehensive way and it is appealing especially in applications where interests are on the tails of the response distribution. However, due to data sparsity, the finite sample estimation at tail quantiles often suffers from high variability. To improve the tail estimation efficiency, we consider modeling multiple quantiles jointly for

cases where the quantile slope coefficients tend to be constant at tails. We propose two estimators, the weighted composite estimator that minimizes the weighted combined quantile objective function across quantiles, and the weighted quantile average estimator that is the weighted average of quantile-specific slope estimators. By using extreme value theory, we establish the asymptotic distributions of the two estimators at tails, and propose a procedure for estimating the optimal weights. We show that the optimally weighted estimators improve the efficiency over equally weighted estimators, and the efficiency gain depends on the heaviness of the tail distribution. The performance of the proposed estimators is assessed through a simulation study and the analysis of a precipitation downscaling data.

email: **kwang6@ncsu.edu**

8. POSTERS: Variable Selection and Methods for High Dimensional Data

8a. BAYES FACTOR CONSISTENCY UNDER G-PRIOR LINEAR MODEL WITH GROWING MODEL SIZE

Ruoxuan Xiang*, University of Florida

Malay Ghosh, University of Florida

Kshitij Khare, University of Florida

In this paper, we examine Bayes factor consistency in the context of Bayesian variable selection for normal linear regression models. We take a hierarchical approach using a hyper g prior (Liang et al.(2008) J. Amer. Statist. Assoc.). There are two regimes for computing Bayes factors, which differ in the choice of the base model. We study conditions under which Bayes factors are consistent for both regimes when the number of all potential regressors grows with sample size. This situation is not fully understood in the current literature, but gains increasing importance recently. In the present case, Bayes factors are not analytically tractable and are calculated via Laplace approximation. Results for other priors on g (e.g., the Zellner-Siow prior) can be obtained in a similar manner.

email: **rxxiang@ufl.edu**

8b. VARIABLE SELECTION FOR COX PROPORTIONAL HAZARD FRAILTY MODEL

Ioanna Pelagia*, The University of Manchester, United Kingdom

Jianxin Pan, The University of Manchester, United Kingdom

Extending the Cox Proportional Hazard (PH) model to Cox PH frailty model may increase the dimension of variable components and become a very challenging task in terms of the significance and estimation of the parameter coefficients. On the other hand, variable selection has always been one of the fundamental problems when it comes to statistical modelling with high dimension variables and has attracted a remarkable attention. Various techniques of variable selection have been proposed such as the best subset variable selection and stepwise elimination, but suffer from several drawbacks in contrast with penalty functions. However, the method proposed here, is to overcome the problem of high dimension under the Cox PH frailty model by considering a simultaneous variable selection of both fixed effects and frailty components through penalty functions such as LASSO and SCAD. Simulation studies show that the proposed procedure works well in selecting and estimating significant fixed and frailty terms. The proposed method is also applied to real data analysis for Diabetes of Type 2.

email: **ioanna.pelagia@manchester.ac.**

8c. FUSED LASSO APPROACH TO ASSESSING DATA COMPARABILITY WITH APPLICATIONS IN MISSING DATA IMPUTATION

Lu Tang*, University of Michigan

Peter X. K. Song, University of Michigan

Missing data imputation is a highly sought-after approach for missing data problems under big-data settings due to the curse of computing burden. Popular imputation methods include single imputation such as mean imputation, regression imputation, stochastic imputation and hot-deck imputation, as well as multiple imputation which takes the average of the outcomes from multiple imputed data sets. It is known that most of the model based imputation methods are sensitive to the assumed distributions to generate plausible values for the replacement of missing data. In effect, when the assumed model is misspecified, the imputation method may yield

"consistently biased" imputation values, which will cause misleading statistical estimation and inference. In this paper, we propose a method to evaluate the discrepancy between the distribution of original data set and that of the imputed data set, which will provide guidance on the selection of appropriate imputation techniques, for example, making a choice between model based imputation and nearest neighbor imputation. Our evaluation is built upon the combination of fused lasso and bootstrap resampling techniques, for both of which statistical software is readily available. We use extensive simulation studies to demonstrate the performance of the new method under different situations. A real data application example is also provided.

email: **lutang@umich.edu**

8d. MULTIPLE IMPUTATION USING SPARSE PCA FOR HIGH-DIMENSIONAL DATA

Domonique Watson Hodge*, Emory University

Qi Long, Emory University

Missing data presents challenges in the statistical analysis phase of research. Common naive analyses such as complete-case and available-case analysis may introduce bias, loss of efficiency, and produce unreliable results. Multiple imputation is one of the most widely used method for handling missing data which can be attributed to its ease of use. However, more research needs to be conducted to determine the best strategy to conduct multiple imputation (MI) in the presence of high-dimensional data. To address this concern, we evaluate several approaches for MI based on sparse principal component analysis (SPCA). The performance of these methods is assessed through numerical studies.

email: **domonique.watson@emory.edu**

8e. TOPIC MODELING FOR SIGNAL DETECTION OF SAFETY DATA FROM ADVERSE EVENT REPORTING SYSTEM DATABASE

Weizhong Zhao*, U.S. Food and Drug Administration

Wen Zou, U.S. Food and Drug Administration

James J. Chen, U.S. Food and Drug Administration

The FDA centers receive reports from consumers, health care professionals, manufacturers, and others regarding the safety of various regulated products, such as drugs, vaccines, artificial hearts, surgical lasers, and nutritional supplements. It is a challenge to extract the information in these reports for better assessment of product safety and rapid detection of adverse event signals. In this study, we applied topic modeling approach, which is a hierarchical Bayesian model, to analyzing FDA adverse event databases. Topic model can reveal "hidden" patterns between adverse events and products, such as drugs, vaccines, or consumption products, to detect potential safety signals. Drug groups, for example, inducing similar adverse events can be identified, and adverse events reported simultaneously with similar drugs are clustered into groups as well. The proposed method was evaluated by an analysis of a FDA drug adverse event dataset, which included 193 cardiovascular drugs with 8453 adverse events. The results identified some new signals as well as those clarified by other commonly used approaches.

email: **weizhong.zhao@fda.hhs.gov**

8f. BUILDING RISK MODELS WITH CALIBRATED MARGINS

Paige Maas*, National Cancer Institute, National Institutes of Health

Yi-Hau Chen, Academia Sinica

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 based on data from large representative cohorts and thoroughly validated in independent studies. As new risk factors are identified, there is a need to update existing risk models to include up-to-date information in predicting disease risk. It is often not reasonable to conduct an entirely 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 a new model. First, we explore the use of a regression calibration approach, utilizing a method from sample-survey literature which is traditionally used for increasing the efficiency of parameter estimation from a survey by leveraging

information from external data sources. Second, we investigate a constrained maximum likelihood approach, leveraging a key constraint identified in our work with the regression calibration method. We present analytic and numerical results that reveal the performance of these approaches in various relevant scenarios.

email: **pmaas@jhsph.edu**

8g. CATEGORICAL PREDICTORS AND PAIRWISE COMPARISONS IN LOGISTIC REGRESSION VIA PENALIZATION AND BREGMAN METHODS

Tian Chen*, North Carolina State University

Howard Bondell, North Carolina State University

Logistic regression is widely used to study the relationship between a binary response and a set of covariates. When the covariates in the logistic regression are categorical, two goals are determining the important factors, and detecting differences among the levels of these important categorical factors. In this paper, we propose a penalization based approach to conduct these pairwise comparisons among the levels. Within a single procedure, the irrelevant factors can be removed, while the levels within the important factors can be collapsed into groups. We propose an algorithm based on Split Bregman iterations, which transforms the constrained problem into a series of simple unconstrained problems. Because of the logistic structure, Iteratively Reweighted Least Squares (IRLS) is applied for optimization. It is shown that the method has the oracle property indicating that asymptotically it performs as well as if the true structure were known in advance. Simulation studies show the superiority of this procedure over the traditional post hoc multiple comparison hypothesis tests. The utility of the method itself, as well as the computational approach, are also examined via a real data analysis.

email: **tchen8@ncsu.edu**

8h. COMPARISON OF STEP-WISE VARIABLE SELECTION, BlmmLasso, AND GMMBoost FOR IDENTIFICATION OF PREDICTOR INTERACTIONS ASSOCIATED WITH DISEASE OUTCOME

Yunyun Jiang*, Medical University of South Carolina

Bethany Wolf, Medical University of South Carolina

Predicting patients' disease risk, severity, or response to treatment often necessitates modeling complex interactions among genetic and environmental variables measured over time. Generalized linear mixed models (GLMM) can model interactions in data with repeated measures, however without an a priori hypothesis, identification of higher-order interactions can be cumbersome. Predictors can be selected using a step-wise variable selection technique comparing models using statistics such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC). However, such variable selection techniques are known to produce unstable estimates. Lasso regression is an alternative predictor selection algorithm that yields sparse estimators by including a "shrinkage" penalty parameter. Lasso regression has been adapted for GLMMs, a method referred to as glmmLasso. GMMBoost, a boosted generalized linear mixed modeling algorithm, is another variable selection technique that yields a sparse solution through reweighting of model residuals. Both glmmLasso and GMMBoost effectively handle large numbers of predictors and achieve sparse prediction models. We conducted a simulation study comparing the ability of step-wise selection in GLMM, glmmLasso, and GMMBoost to correctly identify variables and interactions associated with a disease outcome. We apply these techniques to identify variables and variable interactions associated with treatment response in patients with lupus nephritis.

email: **jiany@musc.edu**

8i. SHRINKAGE PRIORS FOR BAYESIAN LEARNING FROM HIGH DIMESIONAL GENETICS DATA

Anjishnu Banerjee*, Medical College of Wisconsin

Shrinkage priors are widely used in high dimensional settings for variable selection, prediction and learning. There are currently two generic flavors for shrinkage priors - the first being having a global shrinkage parameter and the second, having individual shrinkage for each of the variables in question. There has been a lot of interest of late, notably with the

horse-shoe prior of Scott and Polson, 2010, which belongs to the second category with them being shown to outperform the global shrinkage parameter paradigm in experiments and theoretical settings. We argue in this article that neither approach is optimal - both from theoretical settings and computational perspectives. We propose a new variant of shrinkage priors - which is a "middle-path" between the global and local approaches and show superior empirical performance and significant gains in computational efficiency. We apply our proposed algorithm to the setting of high dimensional genetic data and compare it against competing approaches.

email: **anjishnu@gmail.com**

8j. FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS TO FIFTY-EIGHT MOST TRADED CURRENCIES BASED ON EURO

Jong-Min Kim, University of Minnesota, Morris

Ali H. AL-Marshadi, King Abdulaziz University

Junho Lim*, University of Minnesota, Morris

This research is for investigating the recent trend of fifty eight most traded currencies based on Euro by using functional principal component analysis since January, 2013. We also performed functional linear regression of Brent crude oil given on the currencies which were selected by Bayesian variable selection.

email: **limxx338@morris.umn.edu**

9. POSTERS: Bayesian Methods and Computational Algorithms

9a. NONPARAMETRIC BAYES MODELS FOR MODELING LONGITUDINAL CHANGE IN ASSOCIATION AMONG CATEGORICAL VARIABLES

Tsuyoshi Kunihama, Duke University

Amy Herring*, University of North Carolina, Chapel Hill

David Dunson, Duke University

Carolyn Halpern, University of North Carolina, Chapel Hill

Modeling and computation for multivariate longitudinal data has proven challenging, particularly when data are not all continuous but contain discrete measurements. Approaches based on generalized linear mixed modeling, and related exponential family hierarchical models, have been criticized due to a lack of robustness. In particular, problems arise due to the dual role of the random effects structure in controlling the dependence and shape of the marginal distributions. Motivated by an interesting application to sexual preference data, we propose a novel approach based on a Dirichlet process mixture of Gaussian latent factor models. The proposed model uses a rounded kernel method to allow data to be mixed scale, with a longitudinal factor structure incorporating dependence within-subjects in their repeated measurements. Survey weights are incorporated into the model to facilitate generalizability. Parameter interpretation is considered, and an efficient Markov chain Monte Carlo algorithm is proposed. The methods are assessed through simulation studies and applied to the National Longitudinal Study of Adolescent Health.

email: **amy_herring@unc.edu**

9b. REGRESSION MODEL ESTIMATION AND PREDICTION INCORPORATING COEFFICIENTS INFORMATION

Wenting Cheng*, University of Michigan

Jeremy M. G. Taylor, University of Michigan

Bhramar Mukherjee, University of Michigan

We consider a situation where there is a rich amount of historical data available for the coefficients and their standard errors in a regression model of E(Y|X) from large studies, and we would like to utilize this summary information for improving inference in an expanded model of interest, say, E(Y|X, B). The additional variables B could be thought of as a set of new biomarkers, measured on a modest number of subjects in a new dataset. We formulate the problem in an inferential framework where the historical information is translated in terms of non-linear constraints on the parameter space. We propose several frequentist and Bayes solutions to this problem. In particular, we show that the transformation approach proposed in Gunn and Dunson, 2005 is a simple and effective computational method to conduct Bayesian inference in this constrained

parameter situation. Our simulation results comparing the methods indicate that historical information on E(Y|X) can indeed boost the efficiency of estimation and enhance predictive power in the regression model of interest E(Y|X, B).

email: **chengwt@umich.edu**

9c. CROSS-CORRELATION OF CHANGE POINT PROBLEM

Congjian Liu*, Georgia Southern University

In general, a location or time, the observations or data follow two different models before and after it, which is change point. Change point problems are problems with chronologically ordered data collected over a period of time during which there is known (or suspected) to have been a change in the underlying data generation process. Interest then lies in, retrospectively, making inferences about the time or position in the sequence that the change occurred. (Everitt, Brian, 2010). See Fig1, many change points are shown in the plot of data. In Change point problems, we have series of observations or samples. In most cases, these observations appear in chronological order of their time. In the other hand, sample with a spatial distribution is also possible, which the change point is in space and the position of the interface. In the one-dimensional space, this is time variables, same with above.

email: **cL03124@georgiasouthern.edu**

9d. BAYESIAN NETWORK MODELS FOR SUBJECT-LEVEL INFERENCE

Sayantan Banerjee*, University of Texas MD Anderson Cancer Center

Han Liang, University of Texas MD Anderson Cancer Center

Veerabhadran Baladandayuthapani, University of Texas MD Anderson Cancer Center

We develop Bayesian models to analyze proteomic networks in different cancer types. Our primary aim is to predict patient-specific network structure leveraging multi-domain genomic data using Directed Acyclic Graphical (DAG) models. We infer the prior DAG network based on the training data and obtain the corresponding posterior network based on sparse Bayesian regression methods on each of the nodes, incorporating gene-level information (mRNA, miRNA and methylation) for each of the proteins. Bayesian model averaging is used to predict the responses for patients in the test data, along with obtaining the predictive density for each of the proteins corresponding to each test patient. A network-score is proposed for each patient as a measure of activation of the patient-specific network, based on probabilities of protein-activation for each of the proteins. The network scores are used to fit a survival model for the patients. The methods are motivated by and applied to Reverse Phase Protein Array (RPPA) data for two different cancer types, namely Kidney Renal Clear Cell Carcinoma (KIRC) and Lung Squamous Cell Carcinoma (LUSC), for prediction of the proteins in the PI3K/AKT pathway.

email: **SBanerjee@mdanderson.org**

9e. ALGORITHMS FOR CONSTRAINED GENERALIZED EIGENVALUE PROBLEM

Eun Jeong Min*, North Carolina State University

Hua Zhou, North Carolina State University

The generalized Rayleigh quotient R(x)= (x^t A x) / (x^t B x) for symmetric and positive semi-definite matrices A and B appears as the objective function in many multivariate statistics problems such as principal component analysis, canonical correlation analysis, and partial least squares. Maximizing or minimizing the Rayleigh quotient yields the generalized eigenvector corresponding to the maximal or minimal generalized eigenvalue respectively. In many applications, parameter constraints such as non-negativity and sparsity are necessary for better interpretability and conditioning. We investigate three classes of algorithms for the constrained generalized eigenvector problem: gradient based methods, coordinate descent, and alternating direction method of multipliers. Their numerical efficiency and convergence properties are evaluated by simulation studies and a real data arising from imaging genetics.

email: **emin2@ncsu.edu**

9f. CycloPs: A CYCLOSTATIONARY ALGORITHM FOR AUTOMATIC WALKING RECOGNITION

Jacek K. Urbanek*, Johns Hopkins Bloomberg School of Public Health

Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health

Tamara B. Harris, National Institute on Aging, National Institutes of Health

Nancy W. Glynn, University of Pittsburgh

Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Jaroslaw Harezlak, Indiana University School of Medicine

We develop an algorithm (CycloPs) for automatic recognition of walking periods based on modeling of local cyclostationarity in high-frequency time series obtained from wearable accelerometers. The algorithm uses advanced spectral analysis to recognize walking and to describe its properties at a sub-second level such as walking instantaneous energy (WalE) expressed in earths' gravity units and instantaneous walking frequency (IWF) expressed in steps per second. CycloPs is robust against within- and between-subject variability and it automatically adapts to the length of recording, type of device, and configuration set-up and can be applied to data collected by wrist-, hip- and ankle-worn accelerometers. CycloPs uses a one-pass exhaustive search algorithm that can process a week of data (~150M measurements) in an hour, allowing for efficient processing of very large datasets. We apply our algorithm to free-living data obtained from the Developmental Cohort Study (DECOS) where 50 elderly subjects were monitored for one week (~ 300 GB of data). Our results show that both WalE and IWF are strongly associated with subjects' gender and age.

email: **jkurbane@iupui.edu**

9g. SIMULATION-BASED ESTIMATION OF MEAN AND VARIANCE FOR META-ANALYSIS VIA APPROXIMATE BAYESIAN COMPUTATION (ABC)

Deukwoo Kwon*, University of Miami

Isildinha M. Reis, University of Miami

In meta-analysis crucial inputs are mean effect size and its corresponding variance from the studies in order to obtain pooled estimate. Hozo et al. (2005) proposed the sample standard deviation formulas using median, low and high end of the range, and the sample size. Wan et al. (2014) proposed a new estimation method to estimate standard deviation using same descriptive statistics in Hozo et al. along with inter-quartile range (IQR). These summary statistics are commonly reported in most studies. However, some literature provided different descriptive statistics (and/or summary statistics) other than median, range, and IQR such as 95% confidence interval or just mean and p-value. In longitudinal meta-analysis, we are often given mean and standard deviation at baseline and mean differences and corresponding standard deviations for specific time points relative to baseline. In this study we propose a simulation-based estimation approach using Approximate Bayesian Computation (ABC) technique for estimating mean and variance based on any types of summary statistics found in the published studies. We conduct simulation study to compare the existing methods with the proposed method. We also include an illustrative example of longitudinal meta-analysis of quality-of-life (QoL) data in prostate cancer patients.

email: **DKwon@med.miami.edu**

9h. THE EFFECTS OF SPARSITY CONSTRAINTS ON INFERENCE OF BIOLOGICAL PROCESSES IN STOCHASTIC NON-NEGATIVE MATRIX FACTORIZATION OF EXPRESSION DATA

Wai S. Lee*, Johns Hopkins University

Alexander V. Favorov, Johns Hopkins University

Elana J. Fertig, Johns Hopkins University

Michael F. Ochs, The College of New Jersey

Non-negative matrix factorization (NMF) and related methods, such as PCA, ICA, and Factor Analysis, model gene expression data as a mixture of underlying expression patterns. It has been established that sparsity is a powerful constraint for recovering biological information from these analyses. For example, the CoGAPS matrix factorization algorithm uses a Markov chain Monte Carlo approach that incorporates a prior distribution, which enforces both non-negativity and sparsity constraints. We present results using our new CoGAPS R/C++ Bioconductor package to explore the effects of different levels of sparsity on the recovery of biological information

from a well-studied cancer data set. As expected, we observe increased \chi-squared values as sparsity is increased along with reduced structure in the estimated matrix decomposition. However, in terms of recovery of biological processes previously validated, we find that there is an optimal range of sparsity that provides more reliable estimation of biological process activity.

email: **wlee70@jhmi.edu**

9i. BAYESIAN SAMPLE SIZE DETERMINATION FOR HURDLE MODELS

Joyce Cheng*, Baylor University

David Kahle, Baylor University

John W. Seaman, Baylor University

In many areas of research, count data containing a large number of zero outcomes is common. Hurdle models are often presented as an alternative to zero-inflated models for such data. 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. Sample size determination is an important aspect of experimental design for clinical trials. This is not a new problem in the realm of zero-inflated count data and has been addressed in the literature in a frequentist context. We consider a Bayesian approach to sample size determination for hurdle models and show its application to a hypothetical sleep disorder study.

email: **joyce_cheng@baylor.edu**

9j. FAST COVARIANCE ESTIMATION FOR SPARSE FUNCTIONAL/LONGITUDINAL DATA

Luo Xiao*, Johns Hopkins University

David Ruppert, Cornell University

Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health

Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Covariance function estimation is essential in functional/longitudinal data analysis. While covariance estimation is a bivariate smoothing problem, no bivariate smoother has been tailored to it. In this work, we propose a fast bivariate penalized spline smoother for estimating covariance functions from sparsely observed data. We select the smoothing parameter through leave-one-subject-out cross validation and derive a fast algorithm to overcome computational difficulties. Simulation results show that the proposed method works well. We illustrate the method with an application to a children growth data.

email: **lxiao@jhsph.edu**

9k. PRIOR ELICITATION FOR LOGISTIC REGRESSION WITH DATA EXHIBITING MARKOV DEPENDENCY

Michelle S. Marcovitz*, Baylor University

John Seaman Jr., Baylor University

We model data from a questionnaire with three binary questions, some of which are sensitive. The three questions exhibit first order Markov dependency so that the answer to the second question depends on the answer to the first and the answer to the third question depends on the answer to the second. For example, in a population of female sex workers admitted for treatment of STDs, the questions might be (1) "Do you engage in unprotected sex"?, (2) "Have you been arrested for prostitution?", and (3) "Do you have dependents living with you?" Participants are randomized to different versions of the questionnaire to protect privacy. We offer a Bayesian logistic regression model for analyzing such data where the parameters of interest are marginal and conditional probabilities of answering "yes". We construct power priors and conditional means priors. We consider the issue of induced priors for the marginal and conditional probabilities of "yes" answers when priors are elicited on regression parameters. Finally, we implement a Bayesian sample size determination method based on the two-priors approach for the logistic regression model.

email: **michelle_marcovitz@baylor.edu**

10. Advances in Patient-Centered Outcomes (PCOR) Methodology

PCORI FUNDING OPPORTUNITIES FOR BIOSTATISTICIANS

Jason Gerson*, Patient-Centered Outcomes Research Institute (PCORI)

This talk will provide an overview of the PCORI funding opportunities for biostatisticians and the methodology projects currently funded by PCORI.

email: **jgerson@pcori.org**

CAUSAL INFERENCE FOR EFFECTIVENESS RESEARCH IN USING SECONDARY DATA

Sebastian Schneeweiss*, Harvard University

The routine operation of the US healthcare system produces an abundance of electronically stored data that capture the care of patients as it is provided in settings outside of controlled research environments. The potential for utilizing these data to inform future treatment choices and improve patient care and outcomes of all patients in the very system that generates the data is widely acknowledged. Particularly for elderly multi-morbid patients and most other vulnerable patient groups who are often excluded from randomized trials, these data, properly analyzed, are key to improving care. Further, such secondary data reflect the health outcomes as they occur in routine care, a main goal of effectiveness research. Given these key properties of secondary data and the abundance of electronic healthcare databases covering millions of patients, it is critical to strengthen the rigor of analyses of such data. Highly innovative analytic approaches have recently been developed that (1) are solidly grounded in the principles of science and (2) are made to best fit any electronic healthcare data source. With the involvement of top researchers, patients, doctors, and other decision makers, we plan to evaluate how much better these new methods perform. To prove this, we use several large databases of electronic medical records and health insurance records. We will test the relationship between two newer and frequently used cardiovascular therapies. We will also use computer-generated artificial data in which we can impose a known association. In such simulation studies, we can further understand and improve the performance of these new analytic methods.

email: **schneeweiss@post.harvard.edu**

OPTIMAL, TWO STAGE, ADAPTIVE ENRICHMENT DESIGNS FOR RANDOMIZED TRIALS, USING SPARSE LINEAR PROGRAMMING

Michael Rosenblum*, Johns Hopkins Bloomberg School of Public Health

Xingyuan Fang, Princeton University

Han Liu, Princeton University

Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on accruing data in a randomized trial. These designs can be useful when it is suspected that treatment effects may differ in certain subpopulations, such as those defined by a biomarker or risk factor at baseline. Two critical components of adaptive enrichment designs are the decision rule for modifying enrollment, and the multiple testing procedure. We provide a general method for simultaneously optimizing both of these components for two stage, adaptive enrichment designs. The optimality criteria are defined in terms of expected sample size and power, under the constraint that the familywise Type I error rate is strongly controlled. It is infeasible to directly solve this optimization problem since it is not convex. The key to our approach is a novel representation of a discretized version of this optimization problem as a sparse linear program. We apply advanced optimization tools to solve this problem to high accuracy, revealing new, optimal designs.

email: **mrosenbl@jhsph.edu**

TREATMENT EFFECT INFERENCES USING OBSERVATIONAL DATA WHEN TREATMENTS EFFECTS ARE HETEROGENEOUS ACROSS OUTCOMES: SIMULATION EVIDENCE

John M. Brooks*, University of South Carolina

Cole G. Chapman, University of South Carolina

Helping patients make patient-centered treatment decisions requires treatment effect evidence aligned to the circumstances of individual patients. If treatment effects are heterogeneous

across patients, randomized control trials are impractical for this purpose and many have recognized the necessity of using observational data to generate evidence more closely aligned to individual patients. However, the treatments observed in observational databases are real world treatment decisions that often involve complex assessments of treatment effects across multiple outcomes valued by patients. Risk adjustment (RA) estimators and instrumental variable (IV) estimators are available to estimate treatment effectiveness using observational data. When treatment effects are heterogeneous across patients, though, it is critical to understand that these estimators yield parameter estimates applicable to distinct patient subsets and improper interpretation could lead to dramatic policy mistakes. Here we further conjecture that these interpretive distinctions are less clear when real world treatment decisions are complex and affect more than one outcome. Additional methodological research is needed to understand the proper interpretations of RA and IV estimates in complex treatment scenarios to avoid treatment and policy mistakes. In this study we use simulation modeling to assess the properties of the parameters produced by RA and IV estimators under various relationships between treatment benefits and risks across outcomes.

email: **john-brooks@sc.edu**

11. Looking Under the Hood: Assumptions, Methods and Applications of Microsimulation Models to Inform Health Policy

INTRODUCTION TO THE CISNET PROGRAM AND POPULATION COMPARATIVE MODELING

Eric J. Feuer*, National Cancer Institute, National Institutes of Health

CISNET is a consortium of NCI-sponsored investigators that use simulation modeling to understand past trends in cancer incidence and mortality, and to guide public health research and priorities. In this talk we describe the role of simulation modeling in understanding the population impact of interventions (i.e. screening, treatment and prevention). We review some of the unique aspects of modeling as carried out by CISNET, i.e. the development of flexible broad-based disease models, the ability to model multiple birth cohorts, comparative modeling, transparency in model structure and assumptions, and outreach to partners to make the modeling relevant. Finally, we summarize some of the major accomplishments of CISNET.

email: **rf41u@nih.gov**

MICROSIMULATION MODELING TO INFORM HEALTH POLICY DECISIONS ON AGE TO BEGIN, AGE TO END, AND INTERVALS OF COLORECTAL CANCER SCREENING

Ann G. Zauber*, Memorial Sloan Kettering Cancer Center

Microsimulation modeling is increasingly being used to inform health policy decisions but there is a lack of understanding in the public health and statistical community regarding when these models are needed, what kind of questions they can uniquely address, and what are their strengths and weaknesses. We provide an example of microsimulation modeling to inform a health policy decision when randomized controlled trials could not be conducted for the number of options under consideration. The colorectal cancer models from the Cancer Intervention and Surveillance Modeling Network (CISNET) were used to assess the age to begin screening (ages 40, 50, or 60), age to end screening (75 or 85) and intervals of repeat screening (5, 10, or 20 years for endoscopic tests and 1, 2 or 3 years for fecal occult blood tests). We used a natural history model of the adenoma carcinoma sequence for colorectal cancer and overlaid screening interventions on a large simulated population. The recommended screening strategy was to begin at age 50 and stop at age 75 provided the patient had been consistently screened with negative findings. This was the best strategy to balance life years gained with the resources required and complications associated with screening.

email: **zaubera@mskcc.org**

ROLE OF CALIBRATION AND VALIDATION IN DEVELOPING MICROSIMULATION MODELS

Carolyn M. Rutter*, RAND Corporation

Microsimulation models are an important tool for informing health policy. Models provide a structure for combining a wide range of evidence that represents the current understanding of both disease and interventions to prevent or treat disease. This structure includes a description of heath states that describe key events in a disease processes and rules describing transitions between states. Parameter associated with transitions rules are selected to achieve good fit to observed statistics through a process of model calibration. Once calibrated, models are used to predict population-level outcomes under different policy scenarios. Model validation, evaluation of model predictions for data not used for calibration, is critical for developing confidence in model predictions. This presentation focuses on issues related to microsimulation model validation, using three models for colorectal cancer screening as an example. We evaluated the accuracy of model predictions across a range of natural history landmarks to gain insight into the accuracy of model assumptions. Models generally provided good predictions of observed data, and between-model comparisons supported longer preclinical duration assumptions. Validation is important, but complicated, especially when evaluating fit to multiple targets, when using observed data that may be prone to bias, and when translating models to different target populations.

email: **crutter@rand.org**

USING MICROSIMULATION TO ASSESS THE RELATIVE CONTRIBUTIONS OF SCREENING AND TREATMENT IN OBSERVED REDUCTIONS IN BREAST CANCER MORTALITY IN THE UNITED STATES

Donald A. Berry*, University of Texas MD Anderson Cancer Center

More randomized trials have addressed improvements in treating and screening for breast cancer over the past 30 years than in any other cancer. Over time since the 1980s these interventions have been incorporated into clinical practice in the US; breast cancer mortality has since dropped by 30%, with comparable decreases in many European countries. Are the decreases due to treatment or screening or both? The 7 (now 6) Breast CISNET models specifically address this question and attribute relative benefits to the two types of interventions. Having 7 modeling teams addressing the same question using common data sources is unique. It allows for addressing the variability of conclusions across modeling approaches. In the present case it enables robust conclusions regarding a fundamental scientific and medical question. A New York Times editorial put it thusly: “What seems most important is that each team found at least some benefit from mammograms. The likelihood that they are beneficial seems a lot more solid today than it did four years ago, although the size of the benefit remains in dispute.” I will describe various improvements in Breast CISNET models, especially how they address heterogeneity of the molecular characteristics of breast cancer.

email: **dberry@mdanderson.org**

SYNTHESIS OF RANDOMIZED CONTROLLED TRIALS OF PROSTATE CANCER SCREENING TO ASSESS IMPACT OF PSA TESTING USING MICROSIMULATIONS

Ruth Etzioni*, Fred Hutchinson Cancer Research Center

Roman Gulati, Fred Hutchinson Cancer Research Center

Alex Tsodikov, University of Michigan

Eveline Heijnsdijk, Erasmus University

Harry de Koning, Erasmus University

Randomized trials are the gold standard for evidence regarding the efficacy of cancer screening tests. In the case of PSA screening for prostate cancer, two trials conducted in the US and Europe produced apparently conflicting results, with the European (ERSPC) trial indicating a significant benefit and the US (PLCO) trial showing no benefit. Recognizing that the two trials had different populations, protocols and compliance rates we used simulation modeling to replicate the trials as conducted to determine whether a common screening efficacy could be identified. Three different models of disease natural history, screening and mortality were used. The models showed that under efficacy similar to that in the ERSPC trial, the null result produced by US trial result would not be unexpected (15-28% probability across the models) given the extreme contamination observed on the control arm of the US trial. Further, by modeling differences between the trials one at a

time we were able to identify the main factors that explain the different results. We conclude that differences in implementation explain much if not most of the reported differences in screening efficacy across the trials, but note that our results are subject to a large degree of uncertainty.

email: **retzioni@fhcrc.org**

12. Optimal Inference for High Dimensional Problems

A NON-PARAMETRIC NATURAL IMAGE FOR DECODING VISUAL STIMULI FROM THE BRAIN

Yuval Benjamini*, Stanford University

Bin Yu, University of California, Berkeley

Brain decoding refers to extracting the experimental stimulus - in our case a natural image (photo) or video - from brain activity. For a subject that watches images or video while being scanned in an function MRI, the goal is to reconstruct what they saw from their brain scans. In other words, can we display what they were seeing? For this inverse problem, we consider a so-called Bayesian decoder combining three sources of information: (a) a forward model, using training data, relating the image or video to the evoked brain activity, (b) the estimated multivariate distribution of the prediction errors derived from the regressions, and (c) a prior for the natural stimuli to constrain the inverse operation. In the talk, we will focus on the problem of determining a non-parametric multivariate prior for natural stimuli that will be best suited for reconstruction. We use regularity in the images and the abundance of image databases to estimate a patch-wise density prior. As we show, explicitly accounting for the modeling error in the prior improves the achieved reconstructions. This work is a collaboration with the Gallant lab in UC Berkeley.

email: **yuvalben@stanford.edu**

DOES ℓ_q MINIMIZATION OUTPERFORM ℓ_1 MINIMIZATION

Arian Maleki*, Columbia University

In many application areas ranging from bioinformatics to imaging we are faced with the following question: Can we recover a sparse vector $\beta_o \ in \mathbb{R} ^ p$ from its undersampled set of noisy observations $y\in\ mathbb{R} ^ n$, $y=X\beta+\epsilon$? The last decade has witnessed a surge of algorithms and theories to address this question. One of the most popular algorithms is the ℓ_q-penalized least squares given by the following formulation: $\hat \beta(\lambda,q)=\arg\min_\ beta \| y - X\beta\|_2 ^ 2+\lambda \| \ beta \|_q ^ q.$ Despite the non-convexity of these optimization problems for $0 \leq q<1$, they are still appealing because of the following folklores in the high-dimensional statistics: (i) $\hat \beta(\ lambda,q)$ is closer to β_o than $\ hat{\beta}(\lambda,1)$. (ii) If we employ iterative methods that converge to a local minima of $ \| y - X\beta\|_2 ^ 2 + \lambda \| \beta \|_q ^ q$, then under good initialization these algorithms converge to a solution that is still closer to $\ beta_o$ than $\hat{\beta}(\lambda,1)$. We establish the scope of validity of these folklore theorems. Starting with message passing algorithm as a heuristic method for solving ℓ_q penalized least squares, we study the following questions in the asymptotic settings: (i) What is the impact of initialization on the performance of the algorithm? (ii) When does the algorithm converge to the sparsest solution regardless of the initialization? Studying these questions leads us to the answer of the first folklore theorem, i.e., the performance of the global minimizer $\hat \beta(\lambda,q)$.

email: **arian.maleki@gmail.com**

INFERENCE IN HIGH-DIMENSIONAL VARYING COEFFICIENT MODELS

Mladen Kolar*, University of Chicago

Damian Kozbur, ETH, Zurich

Varying coefficient models have been successfully applied in a number of scientific areas ranging from economics and finance to biological and medical science. Varying coefficient models allow for flexible, yet interpretable, modeling when traditional parametric models are too rigid to explain heterogeneity of sub-populations collected. Currently, as a result of technological advances, scientists are collecting large amounts of high-dimensional data from complex systems which require new analysis techniques. We focus on the high-dimensional linear varying-coefficient model and develop a novel procedure for estimating the coefficient functions in the model based on penalized local linear smoothing. Our procedure works for regimes which allow the number of explanatory variables to be much larger than the sample size,

under arbitrary heteroscedasticity in residuals, and is robust to model misspecification as long as the model can be approximated by a sparse model. We further derive an asymptotic distribution for the normalized maximum deviation of the estimated coefficient function from the true coefficient function. This result can be used to test hypotheses about a particular coefficient function of interest, for example, whether the coefficient function is constant, as well as construct confidence bands for covering the true coefficient function. Construction of the uniform confidence bands relies on a double selection technique that guards against omitted variable bias arising from potential model selection mistakes. We demonstrate how these results can be used to make inference in high-dimensional dynamic graphical models.

e-mail: **mkolar@chicagobooth.edu**

FEATURE AUGMENTATION VIA NONPARAMETRICS AND SELECTION (FANS) IN HIGH DIMENSIONAL CLASSIFICATION

Jianqing Fan, Princeton University

Yang Feng, Columbia University

Jiancheng Jiang, University of North Carolina, Charlotte

Xin Tong*, University of Southern California

We propose a high dimensional classification method that involves nonparametric feature augmentation. Knowing that marginal density ratios are most powerful univariate classifiers, we use the ratio estimates to transform the original feature measurements. Subsequently, penalized logistic regression is invoked, taking as input the newly transformed or augmented features. This procedure trains models equipped with local complexity and global simplicity, thereby avoiding the curse of dimensionality while creating a flexible nonlinear decision boundary. The resulting method is called Feature Augmentation via Nonparametrics and Selection (FANS). We motivate FANS by generalizing the Naive Bayes model, writing the log ratio of joint densities as a linear combination of those of marginal densities. It is related to generalized additive models, but has better interpretability and computability. Risk bounds are developed for FANS. In numerical analysis, FANS is compared with competing methods, so as to provide a guideline on its best application domain. Real data analysis demonstrates that FANS performs very competitively on benchmark email spam and gene expression data sets. Moreover, FANS is implemented by an extremely fast algorithm through parallel computing.

email: **xint@marshall.usc.edu**

13. Lifetime Data Analysis Highlights

MODELING THE "WIN RATIO" IN CLINICAL TRIALS WITH MULTIPLE OUTCOMES

David Oakes*, University of Rochester

Recently the "win ratio" has been popularized as a simple method of statistical analysis for controlled clinical trials with multiple endpoints (see for example Finkelstein and Schoenfeld, 1999 and Pocock et al., 2012). This approach is based on pairwise comparisons between patients in the treatment and control groups using a primary outcome (say, for example, mortality) with ties broken using a secondary outcome (say, occurrence of a cardiac event) when a ranking based on the primary outcome cannot be determined. In interpreting such analyses for studies involving prolonged follow-up it is important to recognize that the observed pairwise preferences and the weight they attach to the component rankings will change over time. We study some properties of this procedure under various models for the treatment effect on each outcome and the dependence between them.

email: **oakes@bst.rochester.edu**

A MODEL FOR TIME TO FRACTURE WITH A SHOCK STREAM SUPERIMPOSED ON PROGRESSIVE DEGRADATION: THE STUDY OF OSTEOPOROTIC FRACTURES

Xin He*, University of Maryland, College Park

G. A. Whitmore, McGill University

Geok Yan Loo, University of Maryland, College Park

Marc C. Hochberg, University of Maryland, Baltimore

Mei-Ling Ting Lee, University of Maryland, College Park

Osteoporotic hip fractures in the elderly are associated with a high mortality in the first year following fracture and a high incidence of disability among survivors. We study first and second fractures of

elderly women using data from the Study of Osteoporotic Fractures (SOF). We present a new conceptual framework, stochastic model and statistical methodology for time to fracture. Our approach gives additional insights into the patterns for first and second fractures and the concomitant risk factors. Our modeling perspective involves a novel time-to-event methodology called threshold regression which is based on the plausible idea that many events occur when an underlying process describing the health or condition of a person or system encounters a critical boundary or threshold for the first time. In the parlance of stochastic processes, this time to event is a first hitting time of the threshold. The underlying process in our model is a composite of a chronic degradation process for skeletal health combined with a random stream of shocks from external traumas, which taken together trigger fracture events.

email: **xinhe@umd.edu**

JOINT RATE MODELS FOR BIVARIATE RECURRENT EVENTS WITH FRAILTY PROCESSES

Mei-Cheng Wang*, Johns Hopkins University

Bivariate or multivariate recurrent event data are often collected in longitudinal studies as the primary outcome measurements for research. We consider statistical modeling for bivariate recurrent events, where the association between two types of recurrent events is characterised by frailty processes and hence allows for time-dependent association. This forms a contrast with those conventional models for bivariate recurrent events where the association is characterised solely by baseline frailty variables. A composite likelihood approach is developed to estimate parameters in the joint rate models in semiparametric setting. The proposed model and method can be used to identify biomarkers or risk factors for recurrent events that could be used to tailor preventive strategies and treatment plans. To illustrate the applicability of the methods, the proposed approaches are applied to data arising from a youth violence study.

email: **mcwang@jhsph.edu**

EFFICIENT ESTIMATION OF THE COX MODEL WITH AUXILIARY LANDMARK TIME SURVIVAL INFORMATION

Chiung-Yu Huang*, Johns Hopkins University

Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Huei-Ting Tsai, Georgetown University

Assessing heterogeneity of treatment effects is of great importance in patient-centered outcomes research, as it is critical to identify subgroups of patients who are likely to benefit from the treatment. However, clinical trials are usually not powered to detect the interaction between treatment and patient characteristics. In this research, we propose a novel approach to improve efficiency in estimating the survival time distribution by synthesizing information from the individual-level data in clinical studies with that from the aggregate survival information at a landmark time. We propose a double-empirical likelihood method to combine published landmark survival information obtained from different sources such as disease registers. We also propose an empirical likelihood ratio test to examine whether the aggregate information is consistent with the individual-level data. Simulation studies show that the proposed estimator yields a substantial gain in efficiency over the conventional partial likelihood approach. A data analysis illustrates the methods and theory. (Joint work with Jing Qin and Huei-Ting Tsai).

email: **cyhuang@jhmi.edu**

14. Recent Advances and Challenges in the Design of Early Stage Cancer Trials

MOTIVATING SAMPLE SIZES IN ONE- AND TWO-AGENT PHASE I DESIGNS VIA BAYESIAN POSTERIOR CREDIBLE INTERVALS

T**homas M. Braun***, University of Michigan

Simulation remains the primary method for which sample sizes are derived for early-phase Bayesian adaptive clinical trials, which is unappealing both due the time needed to program the simulations, as well as the subjective means by which the final sample size is determined. We apply the idea of Bayesian posterior credible intervals as a way to quickly generate a sample size for both one- and two-agent trials that is determined through an objective decision rule. Our methods are also useful for examining the sensitiv-

ity of any design to the prior distribution selected for the model parameter(s) and the operational values assigned to doses, i.e the "skeleton." We compare our approach to that proposed by Cheng for the CRM, and we also use our methods to compare the sample sizes necessary for several models that have been proposed for two-agent designs.

email: **tombraun@umich.edu**

BEYOND THE MTD: PERSONALIZED MEDICINE AND CLINICAL TRIAL DESIGN

Daniel Normolle*, University of Pittsburgh

Brenda Diergaarde, University of Pittsburgh

Julie Bauman, University of Pittsburgh

Cancer therapy is arriving at a crossroads where the half-century paradigm of cytotoxic therapy development will become irrelevant. Recent discoveries based on high-throughput sequencing indicate that the genomics of metastatic disease are an order of magnitude more complex than that of primary disease, implying that treatments for metastatic disease will require combinations of targeted therapies that will be unique to each patient. Validation and optimization of therapy strategies will be, accordingly, much more complex than the design and assessment of monotherapies. I will discuss recent advances in cancer genomics that affect the design of personalized therapies, and speculate on trial designs and endpoints that will be required to move beyond n=1 analyses.

email: **dpn7@pitt.edu**

UNDERSTANDING THE TOXICITY PROFILE OF NOVEL ANTICANCER THERAPIES

Shing M. Lee*, Columbia University

The methods developed for estimating the maximum tolerated dose for chemotherapeutic agents may not be appropriate for novel targeted therapies and immunotherapies. While toxicities from chemotherapy generally arises soon after treatment, there is increasing literature to suggest that this may not be true for novel anticancer therapies. Moreover, toxicities may also be cumulative, with patients experiencing mild toxicities in earlier cycles and progressing into more severe ones in later cycles. Before, we can suggest good designs for these therapies it is necessary to better understand the toxicity profile of these newer therapies. We analyzed the data from several phase I trials on targeted therapies to illustrate the toxicity profile and propose methods to address the complexities of these data. The methods are compared to standard approaches to illustrate the deficiencies of conventional methods and the need for better designs for novel targeted therapies and immunotherapies.

email: **sml2114@columbia.edu**

SIMPLE BENCHMARK FOR PLANNING AND EVALUATING COMPLEX DOSE FINDING DESIGNS

Ken Cheung*, Columbia University

While a general goal of early phase clinical studies is to identify an acceptable dose for further investigation, modern dose finding studies and designs are highly specific to individual clinical settings. In addition, as outcome-adaptive methods often involve complex algorithm, it is crucial to have diagnostic tools at the planning stage to evaluate the plausibility of a method's simulated performance and the adequacy of the algorithm. In this talk, I will introduce a simple technique that provides an upper limit, or a benchmark, of accuracy for dose finding methods for a given design objective. The proposed benchmark is nonparametric optimal, and is demonstrated by examples to be a practical accuracy upper bound for model-based dose finding methods. We illustrate the implementation of the technique in the context of phase I trials that consider multiple toxicities and phase I/II trials where dosing decisions are based on both toxicity and efficacy, and apply the benchmark to several clinical examples considered in the literature. By comparing the operating characteristics of a dose finding method to that of the benchmark, we can form quick initial assessments of whether the method is adequately calibrated and evaluate its sensitivity to the dose-outcome relationships.

email: **yc632@columbia.ed**

15. Large Scale Data Science for Observational Healthcare Studies

BEYOND CRUDE COHORT DESIGNS: PHARMACOEPIDEMIOLOGY AT SCALE

Marc A. Suchard*, University of California, Los Angeles

Massive longitudinal healthcare databases enable development of surveillance solutions to identify and evaluate drug risk at unprecedented scale. Recent comparative drug safety analyses using administrative claims data continue to rely on unadjusted incidence rate ratios. We develop a large-scale regularized regression framework to control for drug exposure-assignment and estimate adjusted incidence rate ratios at scale. Our framework uses advancing computing technology for Big Data to fit statistical models involving 1,000,000s of patients and enables automatic adjustment via stratification, propensity score matching and doubly-robust estimators. These models involve conditioned likelihoods that were previously computationally impractical in observational healthcare. In our framework, we include all clinical information available about patients up to their time of indication diagnosis and treatment exposure, such as all possible drug prescriptions, medical conditions, procedures and other demographics. The number of covariates stands in the 10,000s, regularization helps us avoid overfitting and algorithmic optimization provides estimates in real-time. We apply our method to examine incidence rates of in-patient gastrointestinal bleeding among atrial fibrillation patients taking dabigatran or warfarin in a database that covers over 227M patient-years.[joint work with the Observational Healthcare Data Sciences and Informatics program]

email: **msuchard@ucla.edu**

HONEST INFERENCE FROM OBSERVATIONAL DATABASE 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. We have 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. I describe this work, explore opportunities to expand the use of observational data to further our understanding of medical products, and highlight areas for future research and development.

email: **david.madigan@columbia.edu**

SAFETY ANALYSIS STRATEGIES FOR COMPARING TWO COHORTS SELECTED FROM HEALTHCARE DATA USING PROPENSITY SCORES

William DuMouchel*, Oracle Health Sciences

Rave Harpaz, Oracle Health Sciences

Propensity scores provide a way to select two cohorts from a longitudinal healthcare database that are matched by their estimated probability of exposure to two therapies. This is designed to minimize potential biases caused by the non-randomized treatment assignment. This balance theoretically protects against bias when comparing all outcomes observed after treatment assignment, so that, for example, the two cohorts can be compared across a wide variety of safety risks. We focus on the use of the high dimensional propensity score method, and discuss and illustrate how longitudinal data from the two cohorts can be imported into a general purpose tool for comparisons across many safety outcomes. Fecundity is defined as the biologic potential of men and women for reproduction, and is often measured by estimating the probability of pregnancy in each menstrual cycle among couples having regular unprotected intercourse. Estimating fecundity is challenging, in part, given the effect that varying patterns of sexual intercourse may have on the length of pregnancy attempt. Clinical guidance is sometimes sought to aid couples in timing intercourse acts around ovulation to minimize the time needed to achieve pregnancy. Empirical evidence delineating the timing of intercourse

relative to ovulation are few, resulting in a generalized clinical recommendation to have intercourse every other day (Practice Committee of the American Society for Reproductive Medicine, 2013). Understanding the relation between fecundity, intercourse behavior and other relevant covariates is increasingly relevant given population level changes in the sociodemographic characteristics of reproductive aged couples such as an increase in age at first pregnancy. This may be associated with reduced intercourse activity, longer time-to-pregnancy, an increased prevalence of infertility or a combination of all these factors. Our main objective is to jointly model intercourse behavior, a binary longitudinal process (measured on day level), menstrual cycle characteristic (measured on monthly level and TTP, a survival outcome (on monthly timescale), with a view towards prediction of both longitudinal processes on differing timescales and time to pregnancy. This is achieved using an empirical bayes approach of joint modeling of multivariate longitunidal processes and time to event.

email: **bill.dumouchel@oracle.com**

INTERPRETABLE FEATURE CREATION AND MODEL UNCERTAINTY IN OBSERVATIONAL MEDICAL DATA

Tyler McCormick*, University of Washington

Rebecca Ferrell, University of Washington

Large-scale observational health databases (such as electronic medical records or administrative claims data) capture continuous-time, unsolicited recordings of patient experiences. As with many emerging data sources without a formal sampling design, these data require substantial pre-processing before using standard statistical tools. For observational health databases, pre-processing often involves coding for characteristics present at a designated baseline period through discretization of the temporal element of the records, e.g. coarsening the health event timelines over a specified "lookback period" into a binary or count feature to capture prior disease history. Though there is rich literature examining model selection, very little work examines these pre-processing, "feature creation" choices. We propose a model uncertainty framework to address this problem in the context of medical event prediction. Through simulations and an application to health claims data, we demonstrate the effect of decisions to encode time-varying information as static baseline covariates on predictive performance and discuss approaches to account for uncertainty in defining lookback periods.

email: **tylermc@uw.edu**

16. CONTRIBUTED PAPERS: Competing Risks

EXTENDING FINE AND GRAY'S MODEL: GENERAL APPROACH FOR COMPETING RISKS ANALYSIS

Anna Bellach*, University of Copenhagen and University of North Carolina, Chapel Hill

Jason Peter Fine, University of North Carolina, Chapel Hill

Ludger Rüschendorf, Albert Ludwigs University of Freiburg im Breisgau

Michael R. Kosorok, University of North Carolina, Chapel Hill

We introduce a pseudo likelihood function that can be used to derive estimators for the hazard rate of the subdistribution in competing risks settings for a broad class of semiparametric regression models. Two important special cases of our approach are the Fine and Gray model and the proportional odds model for the hazard rate of the subdistribution. For a general class of semiparametric transformation models we prove the consistency and asymptotic normality of the estimators. Our estimates are directly interpretable as we target on the hazard rate of the subdistribution. Our model is efficient for administrative censored data. In simulation studies we show that also for right censored data our model performs well with respect to the variances even for very small sample sizes. We apply the method to a bone marrow

transplant data set to demonstrate its practical utility. We illustrate how our proposed method improves the precision of prediction for the individual event types if the appropriate link function is selected by the Akaike information criterion.

email: **annabella@sund.ku.dk**

NON-PARAMETRIC CUMULATIVE INCIDENCE ESTIMATION UNDER MISCLASSIFICATION IN THE CAUSE OF FAILURE

Giorgos Bakoyannis*, Indiana University

Menggang Yu, University of Wisconsin

Constantin T. Yiannoutsos, Indiana University

Constantine Frangakis, Johns Hopkins University

The fundamental identifiable quantities in cohort studies and clinical trials with competing risks are the cause-specific hazard and the cumulative incidence function. However, in many clinical settings, the cause of failure is diagnosed with error. This type of misclassification is expected to lead to biased cumulative incidence estimates. In this work we evaluate the effect of cause of failure misclassification in cumulative incidence estimates and propose a weighted version of the Aalen-Johansen non-parametric estimator to adjust for such a misclassification. The weights are functions of the misclassification probabilities, which can be estimated through double-sampling techniques of a random sample of subjects whose true cause of failure is unequivocally ascertained through possibly more expensive diagnostic methods. Consistency and asymptotic normality of the estimator is established and finite-sample properties are studied through simulation experiments. The method is illustrated using data from HIV-1 seropositive individuals in sub-Saharan Africa, where serious death under-reporting (with deceased patients being misclassified as dropouts) affects the estimates of the cumulative incidence of mortality and of non-retention in HIV care.

email: **gbakogia@iu.edu**

EFFICIENT ESTIMATION OF SEMI-PARAMETRIC TRANSFORMATION MODELS FOR THE CUMULATIVE INCIDENCE OF COMPETING RISKS

Lu Mao*, University of North Carolina, Chapel Hill

Danyu Lin, University of North Carolina, Chapel Hill

For analysis of competing risks data, interest has centered on the cumulative incidence because of its practical relevance and direct interpretation.
A semiparametric regression model proposed by Fine and Gray (1999) has become the method of choice for formulating the effects of covariates on the cumulative incidence. Its estimation, however, requires modeling of the censoring distribution and is not statistically efficient. In this article, we present a broad class of semiparametric transformation models which extends the Fine and Gray model, and we derive the nonparametric maximum likelihood estimators (NPMLEs). We develop a simple and fast algorithm for computing the NPMLEs through the profile likelihood. We establish the consistency, asymptotic normality, and semiparametric efficiency of the NPMLEs. In addition, we construct graphical and numerical procedures to evaluate and select models. Finally, we demonstrate the advantages of the proposed methods over the existing ones through extensive simulation studies and an application to a major study on bone marrow transplantation.

email: **lmao@unc.edu**

JOINT DYNAMIC MODELING OF RECURRENT COMPETING RISKS AND A TERMINAL EVENT

Piaomu Liu*, University of South Carolina, Columbia

Edsel Peña, University of South Carolina, Columbia

Recurrent events and terminal events occur in many areas in the biomedical and public health settings. In this talk a joint model for recurrent competing risks and a terminal event will be described. Associations among the recurrent competing risks are induced by both a frailty variable and the impact of previous recurrent event occurrences. The recurrent competing risks also impact the occurrence of the terminal event. In addition, further association between the terminal event and the recurrent competing risks is induced by a frailty variable. To dynamically model the impact of interventions after each event occurrence on the recurrent competing risks, an effective age process is introduced. The impact of the increasing number of

recurrent event occurrences and covariate processes are also incorporated into the semiparametric model. Estimators of the parameters of the proposed model will be described. Some finite-sample and large-sample properties of estimators will be presented.

email: **piaomuliu@gmail.com**

DYNAMIC PREDICTION OF SUBDISTRIBUTION FUNCTIONS FOR DATA WITH COMPETING RISKS

Qing Liu*, University of Pittsburgh

Chung-Chou H. Chang, University of Pittsburgh

To be able to dynamically predict a patient's prognosis based on the disease progression is very helpful to the physician for medical decision making. Landmark Cox models have great potential for serving the purpose of dynamic prediction but the use of such models becomes much more challenging when competing risks are present. Several studies have extended the landmark method to competing risks regression models, however, the resulting models are either sensitive to the proportional subdistribution hazards (PSH) assumption, or somewhat difficult to handle time-dependent covariate values and time-varying covariate effects simultaneously. In this study, we developed a landmark PSH model and a more comprehensive landmark PSH supermodel. Our proposed models have four advantages over other dynamic predictive models in addressing competing risks. First, they are robust against violations of the PSH assumption. Second, the landmark PSH supermodel enables users to make predictions with a set of landmark points in one step. Third, the proposed models can incorporate various types of time-varying information. Finally, our models are not computationally intensive and can be easily implemented with existing statistical software. We assessed the performance of our models via simulations and applied the proposed models to a data set from a multicenter clinical trial for breast cancer patients.

email: **qil18@pitt.edu**

COMPETING RISKS REGRESSION USING PSEUDO-VALUES UNDER RANDOM SIGNS CENSORING

Tianxiu Wang*, University of Pittsburgh

Chung-Chou H. Chang, University of Pittsburgh

In medical studies, investigators are often interested in estimating marginal survival distributions of latent failure times when competing risks exist. Without further assumption, marginal survival functions are not identifiable (Tsiatis, 1975). In this study, we incorporate the random signs censoring principle (Cooke, 1993; Yabes, 2012) in estimating marginal survival functions. The random signs censoring (RSC) principle is verifiable from the observed data. We propose an estimator of the effect of a covariate on marginal survival functions. The proposed estimator is based on pseudo-values for inverse-probability-censoring-weighted (IPCW) Kaplan-Meier functions and the corresponding marginal survival function can be written in the form of a standard generalized linear regression model. The proposed estimator is easy to implement and it also has desirable asymptotic properties. We evaluated the finite-sample performance of the estimator via simulation studies. In the application, we applied the proposed method to identify potential risk factors for 90-day mortality without transplantation for pediatric patients with end-stage liver diseases.

email: **tiw17@pitt.edu**

KERNEL SCORE TEST FOR PROGRESSION FREE SURVIVAL

Matey Neykov*, Harvard University

Tianxi Cai, Harvard University

Recently papers have emerged, for testing whether certain genetic information has effect on disease progression. The kernel methods, that these papers use are highly flexible, allowing for the genetic information to have nonlinear and non-additive effects on the disease progression. In this paper we are interested in utilizing these methods to obtain a test in a semi-competing risk situation. The main advantage of using such methods, is that we can capture of the informative censoring which might be lost by simply using a progression free survival (PFS) model, which is typically used as a gold-standard in the literature. However, our simulations demonstrate that the PFS can have a really poor performance in some situations.

email: **mneykov@g.harvard.edu**

17. CONTRIBUTED PAPERS: Applications and Methods in Environmental Health

METHODOLOGY FOR QUANTIFYING THE CHANGE IN MORTALITY ASSOCIATED WITH FUTURE OZONE EXPOSURES UNDER CLIMATE CHANGE

Stacey E. Alexeeff*, National Center for Atmospheric Research

Gabriele G. Pfister, National Center for Atmospheric Research

Doug Nychka, National Center for Atmospheric Research

Climate change is expected to have many impacts on the environment, including changes in ozone concentrations at the surface level. A key public health concern is the potential increase in ozone-related summertime mortality if surface ozone concentrations rise in response to climate change. Previous health impact studies have not incorporated the variability of ozone into their prediction models. We propose a Bayesian posterior analysis and Monte Carlo estimation method for quantifying health effects of future ozone. The key features of our methodology are (i) the propagation of uncertainty in both the health effect and the ozone projections and (ii) use of the empirical distribution of the daily ozone projections to account for their variation. The use of interpolation to improve the accuracy of averaging over irregular shaped regions helps to derive average exposure for the regions where mortality and demographic information is reported. We also derive an analytic expression for the integral with respect to the mortality parameter, which is useful to reduce the Monte Carlo computational burden associated with this parameter. Using our proposed approach, we quantify the expected change in ozone-related summertime mortality in the contiguous United States between 2000 and 2050 under a changing climate. We also illustrate the results when using a common technique in previous work that averages ozone to reduce the size of the data, and contrast these findings with our own.

email: **salexeef@ucar.edu**

ESTIMATION OF ENVIRONMENTAL EXPOSURE DISTRIBUTION ADJUSTING FOR DEPENDENCE BETWEEN EXPOSURE LEVEL AND DETECTION LIMIT

Yuchen Yang*, University of Kentucky

Brent Shelton, University of Kentucky

Tom Tucker, University of Kentucky

Li Li, Case Western Reserve University

Richard Kryscio, University of Kentucky

Li Chen, University of Kentucky

In environmental exposure studies, it is common to observe a portion of exposure measurements to fall below experimentally determined detection limits (DLs). The reverse Kaplan-Meier (RKM) estimator, which mimics the well-known Kaplan-Meier estimator for right-censored survival data with the scale reversed, has been recommended for estimating the exposure distribution for the data subject to DLs. However, the RKM estimator requires the independence assumption between the exposure level and DL and can lead to biased results when this assumption is violated. We propose a kernel-based nonparametric estimator for the exposure distribution without imposing any independence assumption between the exposure level and DL. We show the proposed estimator is consistent and asymptotically normal. Simulation studies demonstrate that the proposed estimator performs well in practical situations. A colon cancer study is provided for illustration.

email: **yuchen.y@uky.edu**

SPATIAL CONFOUNDING, SPATIAL SCALE AND THE CHRONIC HEALTH EFFECTS OF COARSE THORACIC PARTICULATE MATTER

Helen Powell*, Johns Hopkins Bloomberg School of Public Health

Roger D. Peng, Johns Hopkins Bloomberg School of Public Health

Spatial confounding occurs when unmeasured spatially varying confounders make it difficult to distinguish the effect of the exposure from residual spatial variation in the outcome. Studies which aim to investigate the long-term health effects of air pollution typically include a spatial term in the regression model to account for the inherent bias associated with spatial confounding. However, this bias can only be reduced if the spatial scale of the exposure is smaller than that of

the unmeasured confounder, a concept which has thus far not been considered by air pollution studies. We aim to investigate the long-term health effects of coarse thoracic particulate matter (PM), a metric of PM which is currently unregulated by the US Environmental Protection Agency, taking into account the spatial scale of the exposure. By allowing the spatial scale of the pollutant to inform spatial terms in the regression we aim to reduce bias in the estimated health effects resulting from spatial confounding.

email: **hpowell6@jhu.edu**

ESTIMATING THE CAUSAL EFFECT OF COAL BURNING POWER PLANTS ON CO2 EMISSIONS

Georgia Papadogeorgou*, Harvard School of Public Health

Corwin Zigler, Harvard School of Public Health

Francesca Dominici, Harvard School of Public Health

Prior literature on evaluating the effect of policy regulations on power plants through the Acid Rain Control Program (ARCP) is scarce and has been restricted to simple methodological and statistical approaches. Potential flaws in these approaches could be a failure to control for confounders of the relationship between regulations and emissions or health outcomes, unobserved confounding, misinterpretation of estimates, and comparison of implausible interventions. One compliance strategy following these regulations is the switch of power plant units primary fuel from coal to other fuels (primarily natural gas). Using the monthly aggregated ARCP data, we identify important confounders and estimate the monthly causal effect of coal on CO2 emissions in 2012 using three statistical procedures: Coarsened Exact matching, Nearest Neighbor Propensity Score matching, and Nearest Neighbor Distance Adjusted Propensity Score (DAPS) matching. The first two methods are well-established methods of estimating the causal effect of an intervention using observational data, and we propose DAPS as a way to further adjust for confounding by incorporating the distance between power plant locations to control for spatially varying, unobserved confounding. For each method we fit 3 different models to estimate the effect. All models give similar estimates of the causal effect of coal (~31,500-40,000 tons of CO2 emissions per month of 2012) and revealed similar patterns of seasonality.

email: **gpapadogeorgou@fas.harvard.ed**

TEMPORAL ASPECTS OF AIR POLLUTANT MEASURES IN EPIDEMIOLOGIC ANALYSIS: A SIMULATION STUDY

Laura F. White*, Boston University

Jeffrey Yu, Boston University

Bernardo Beckerman, University of California, Berkeley

Michael Jerrett, University of California, Berkeley

Patricia Coogan, Boston University

Background. Numerous observational studies have assessed the association between ambient air pollution and chronic disease incidence. There is however no uniform approach to create an exposure metric that captures the variability in air pollution through time and determines the most relevant exposure window for determining risk of chronic illness. Methods. In this study we use simulation to assess nine exposure metrics that incorporate the time trends in ambient air pollution and make use of different exposure windows. We simulate observational data based on the characteristics of the Black Women's Health Study and use observed values for particulate matter < 2.5 microns (PM2.5) for this cohort to create the nine exposure metrics. Results. When we fit Cox proportional hazards models using the nine different metrics, we observe that time-invariant metrics perform poorly and tend to underestimate the true hazard ratio. Time varying metrics that average previous values tend to perform well. Conclusions. Our simulation study indicates that the use of averaged time-varying exposure metrics provide the least biased results.

e-mail: **lfwhite@bu.edu**

BAYESIAN MODELS FOR MULTIPLE OUTCOMES IN DOMAINS WITH APPLICATION TO THE SEYCHELLES CHILD DEVELOPMENT STUDY

Luo Xiao, Johns Hopkins Bloomberg School of Public Health

Sally W. Thurston*, University of Rochester

David Ruppert, Cornell University

Tanzy M.T. Love, University of Rochester

Philip W. Davidson, University of Rochester

The Seychelles Child Development Study examines the effects of prenatal methylmercury exposure on central nervous system functioning. The data include 20 outcomes measured on 9-year old children that can be classified into four "domains": cognition, memory, motor, and social behavior. Previous analyses and scientific theory suggest that some outcomes may belong to more than one domain. We develop a framework in which each domain is defined by a sentinel outcome preassigned to that domain only, while all other outcomes may belong to multiple domains and are not preassigned. Our model allows us to learn about assignment of outcomes to domains, while allowing exposure and covariate effects to differ across domains and across outcomes within domains. We take a Bayesian MCMC approach. Results from the Seychelles study and from extensive simulations show that our model can effectively determine sparse domain assignment, and give increased power to detect overall, domain-specific and outcome-specific exposure and covariate effects than separate models. When fit to the Seychelles data, several outcomes were classified as partly belonging to several domains. Checks of model misspecification were improved relative to a model that assumes each outcome is in a single domain.

e-mail: **thurston@bst.rochester.edu**

ANALYSIS OF 26 MILLION AREA VOC OBSERVATIONS FOR THE PREDICTION OF PERSONAL THC EXPOSURE USING BAYESIAN MODELING

Caroline P. Groth*, University of Minnesota

Sudipto Banerjee, University of California, Los Angeles

Gurumurthy Ramachandran, University of Minnesota

Ian Reagen, University of Minnesota

Richard Kwok, National Institute of Environmental Health Sciences, National Institutes of Health

Aaron Blair, National Cancer Institute, National Institutes of Health

Dale Sandler, National Institute of Environmental Health Sciences, National Institutes of Health

Lawrence Engel, National Institute of Environmental Health Sciences, National Institutes of Health

Mark Stenzel, Stewart Exposure Assessments, LLC

Patricia Stewart, Stewart Exposure Assessments, LLC

Environmental health researchers often develop estimates of personal exposure to chemicals where exposure data are limited. In the NIEHS Gulf STUDY, an epidemiologic study of the health of workers who participated in the 2010 Deepwater Horizon oil spill clean-up, researchers are using monitoring data from the time of the spill to develop task, time, and location-specific exposure estimates for several oil-related chemicals. One data set contains 4300 full work shift measurements of personal total hydrocarbon (THC) measurements and another has over 26,000,000 short-term area measurements of volatile organic compounds (VOC) collected on 38 vessels assisting in the clean-up. We present a Bayesian model and framework for estimating personal THC airborne exposures from area VOC data when personal THC data are missing. We first summarize the 26,000,000 area VOC observations in hourly averages by vessel. Then, we correlate the VOC hourly averages that overlap with the time of each THC sample. From the relationship between VOC and THC, we develop a model for predicting personal THC exposure. Throughout this analysis, we employ methods to account for values below the limit of detection in both VOC and THC measurements. Using this framework we present preliminary findings on a subset of this analysis.

e-mail: **groth203@umn.edu**

18. CONTRIBUTED PAPERS: Statistical Methods for Genomics

IDENTIFICATION OF CONSISTENT FUNCTIONAL MODULES

Xiwei Chen*, State University of New York at Buffalo

David L. Tritchler, State University of New York at Buffalo

Jeffrey C. Miecznikowski, State University of New York at Buffalo

Daniel P. Gaile, State University of New York at Buffalo

It is often of scientific interest to find a set of genes that may represent an independent functional module or network, such as a functional gene expression module causing a biological response, a transcription regulatory network, or a constellation of mutations jointly causing a disease. In this paper we are specifically interested in identifying modules that control a particular outcome variable such as a disease biomarker. We discuss the statistical properties that functional networks should possess and introduce the concept of network consistency which should be satisfied by real functional networks of cooperating genes, and directly use the concept in the pathway discovery method we present. Our method gives superior performance for all but the simplest functional networks.

e-mail: **xiweiche@buffalo.edu**

A MEDIATION-BASED INTEGRATIVE GENOMIC ANALYSIS OF LUNG CANCER

Sheila Gaynor*, Harvard University

Xihong Lin, Harvard University

Genetic association methods have traditionally been used to analyze the relationship between SNP or sequencing data and disease outcomes. This standard approach often fails to explain a significant proportion of disease and elucidate the complete relationship between SNPs and complex diseases. It has thus been suggested that studies may be improved by jointly analyzing SNP and gene expression data to analyze phenotypes of complex diseases. Huang, VanderWeele and Lin (2014) proposed that this can be approached using a mediation model, where the association between SNP sets or whole genome sequences and a disease outcome is mediated by gene expression or epigenetic data. In our analysis we explore this framework, leveraging the lung adenocarcinoma and squamous cell data sets (n=1025) from The Cancer Genome Atlas. We utilize the whole genome SNP data, gene expression and methylation data, and clinical data on disease phenotypes such as survival. We show the application of such a mediation framework can quantitatively and qualitatively supplement a traditional genetic association study by providing explanation of the mechanisms leading to disease phenotypes. Further, we provide empirical evidence suggesting for which genomic settings this framework is useful.

e-mail: **sgaynor@fas.harvard.edu**

NONPARAMETRIC FAILURE TIME ANALYSIS WITH GENOMIC APPLICATIONS

Cheng Cheng*, St. Jude Children's Research Hospital

Genome-wide Association Study (GWAS) has become routine in cancer genomic translational research, which often requires genome-wide screening to identify ordinal genomic features that are associated with treatment outcome, for example, single nucleotide polymorphisms associated with time to relapse. The estimated coefficient of a hazard rate regression model (HRRM) is often used as the association test statistic. It will be demonstrated in this talk that in certain cases the HRRM approach is problematic. A robust, completely nonparametric alternative using rank correlation is then proposed. This method, called correlation profile test (CPT), consists of the correlation profile statistic and a very efficient hybrid permutation test combining permutation and asymptotic theory. Statistical performances are compared with several established methods, by a simulation study and analysis of real genomics data. It is shown that CPT performs much better than the HRRM approach in terms of maintaining the power and nominal significance level, especially in cases where the proportional hazard model does not hold.

e-mail: **cheng.cheng@stjude.org**

AN OMNIBUS TEST FOR DIFFERENTIAL ABUNDANCE ANALYSIS OF MICROBIOME DATA

Jun Chen*, Mayo Clinic, Rochester

Emily King, Iowa State University

Diane Grill, Mayo Clinic, Rochester

Karla Ballman, Mayo Clinic, Rochester

One central goal of microbiome studies is to identify taxa that show differentiation between sample groups. The identified taxa can provide insights into disease etiology as well as be used as biomarkers for disease diagnosis and prevention. Many methods have been developed to address this statistical problem ranging from simple adaptation of the t test (Metastat) to more sophisticated statistical test based on zero-inflated Gaussian model (metagenomeSeq) and count based methods (DESeq2). However, none of the statistical methods have taken into account all the features of the taxa data, which are zero-inflated overdispersed count data. Moreover, most of the methods focus on detecting the change of the mean of the taxa abundance. In real situation, disease could affect not only the abundance mean but also the prevalence and the variance. Both dysbiosis and disease heterogeneity can lead to differential variance. We therefore develop an omnibus test based on a zero-inflated count model that jointly tests the equality of mean, zero probability and variance between two sample groups. Both simulations and real data applications demonstrated the increased power of the omnibus test as well better control of the type I error than existing methods.

e-mail: **jchen1981@gmail.com**

SPARSE ANALYSIS FOR HIGH DIMENSIONAL DATA WITH APPLICATION TO DATA INTEGRATION

Sandra Addo Safo*, Emory University

Jeongyoun Ahn, University of Georgia

A core idea of most multivariate data analysis methods is to project higher dimensional data vectors on to a lower dimensional subspace spanned by a few meaningful directions. Many multivariate methods, such as canonical correlation analysis (CCA), multivariate analysis of variance (MANOVA), and linear discriminant analysis (LDA), solve a generalized eigenvalue problem. We propose a general framework, called substitution method, with which one can easily obtain a sparse estimate for a solution vector of a generalized eigenvalue problem. We employ the idea of direct estimation in high dimensional data analysis and suggests a flexible framework for sparse estimation in all statistical methods that use generalized eigenvectors to find interesting low-dimensional projections in high dimensional space. We illustrate the framework with sparse CCA for joint analysis of two high dimensional datasets- gene expression measurements and copy number variations, to study the idea that changes in expression profiles may be associated with copy number variations and that copy number variations may be related to gene expression measurements.

e-mail: **seaddosafo@gmail.com**

ROBUST INFERENCE OF CHROMOSOME 3D STRUCTURE USING HI-C CHROMATIN INTERACTION DATA

Kai Wang*, University of Iowa

Kai Tan, University of Iowa

DNA-DNA spacial contact counts revealed by chromosome conformation capture (3C) techniques contain valuable information regarding chromosome 3D structure. Popular consensus approaches for inferring chromosome 3D structure include multidimensional scaling (MDS) and likelihood-based modeling (LM). MDS method employs a pre-determined contact-to-distance transfer function. However, there are mounting evidences against the existence of a universal transfer function. Although LM does not require specification of a transfer function, it needs a distribution for contact counts which is typically assumed to be Poisson or negative binomial neither of which is empirically justified. Most importantly, spatial coordinates in these methods do not seem to be uniquely identifiable as they are not invariant to rotation or shifting. Hence the effort to search for the global optimal solution is severely compromised. We propose a novel variation of the MDS method by focusing on the topological similarity between the inferred spatial 3D structure and the structure of contact counts. Unlike a recent MDS-based variant, our method allows for a more general transfer function. In addition, among the 3n spatial coordinates for the n loci, 3n-7 of them can be uniquely identified after fixing the other 7 coordinates. The usefulness of the proposed method is demonstrated by simulation studies and an empirical study.

e-mail: **kai-wang@uiowa.edu**

19. CONTRIBUTED PAPERS: Spatial and Spatio-Temporal Methods and Applications

A SEMIPARAMETRIC APPROACH FOR SPATIAL POINT PROCESS WITH GEOCODING ERROR IN CASE-CONTROL STUDIES

Kun Xu*, University of Miami

Yongtao Guan, University of Miami

When conducting risk estimation in epidemiological studies, residence of subjects is commonly geocoded in geographic information systems software by converting residential addresses to geographic coordinates. The ignorance of geocoding error in spatial analysis usually results in biased parameter estimates, inflated standard errors and reduced statistical power to detect spatial cluster and trends. In this article, we propose a novel bias-correction method for such data, where only a small portion of true case and control locations are observed. We construct score vector at each location without any distribution assumptions on error distribution. We study spatial correlation of those score vectors and establish our estimating equation. We show consistency and asymptotical normality of our estimator. We illustrate our method through simulation and Iowa Carroll County childhood asthma data.

email: **kunxu0609@gmail.com**

SEMIPARAMETRIC NONSEPARABLE SPATIAL-TEMPORAL SINGLE INDEX MODEL

Hamdy Fayez Farahat Mahmoud*, Virginia Tech

Inyoung Kim, Virginia Tech

In this paper, we propose two semiparametric single index models for spatially-temporally correlated data. One model has the nonparametric function separable from spatially correlated random effects and time effects. We call thus model semiparametric spatio-temporal separable single index model (SSTS-SIM), while the other does not separate the nonparametric function and spatially correlated random effects but separates the time effects, we call it semiparametric spatio-temporal nonseparable single index model (SSTN-SIM). Two algorithms based on Markov Chain Expectation Maximization algorithm are introduced to estimate the models parameters, spatial effects and times effects. The proposed models are applied to the mortality data set of six major cities in South Korea. The data covers the period from January, 2000 to December, 2007. It is found that Busan city has the highest mortality and Seoul and Daejeon have the lowest mortality. SSTS-SIM enforces the unknown mortality functions of all cities to have the same shape but SSTN-SIM is more flexible. In terms of estimation, SSTN-SIM is better than SSTSSIM. In terms of prediction, in case we have enough data, SSTN-SIM is better.

email: **ehamdy@vt.edu**

STATISTICAL ANALYSIS OF FEED-FORWARD LOOPS ARISING FROM AGING PHYSIOLOGICAL SYSTEMS

Jonathan (JJ) H. Diah*, Columbia University

Feiran Zhong, Columbia University

Arindam RoyChoudhury, Columbia University

We define a feed-forward loop as a multivariate continuous-valued discrete-time stochastic process where one variable influences another variable, and in turn the former variable is influenced by the latter variable at a later time. Specifically, a bivariate feed-forward loop is defined as a stochastic process {Xt, Yt}, where Xt is associated with Y(t+1), after taking into account the effects Yt, and Yt is associated with X(t+1), after taking into account the effects Xt. A trivariate feed-forward loop is similarly defined for three variables. One way of performing inference on feed-forward loops is structural equation modeling. Application of a feed-forward loop process can come from any branch of science with multiple interacting systems; the aging physiological system is one such example. We have modeled aging physiological measures data from two major aging cohorts and concluded that there are significant evidences of feed-forward loop relationships between certain variables. In particular, we found evidence that physical functionality, lean muscle mass, and physical performance measures interact in a feed-forward loop. Such results are important for understanding how physiological systems interrelate with each other and lead to aging. Thus, our novel modeling of the feed-forward loop has far reaching applications in biostatistics of geriatrics.

email: **jhdiah@gmail.com**

BAYESIAN COMPUTATION FOR LOG-GAUSSIAN COX PROCESSES: A COMPARATIVE ANALYSIS OF METHODS

Ming Teng*, University of Michigan

Farouk S. Nathoo, University of Victoria

Timothy D. Johnson, University of Michigan

The Log-Gaussian Cox Process (LGCP) is a commonly used model for the analysis of spatial point pattern data. Different methods have been proposed for inference including traditional likelihood-based approaches as well as methods based on the Bayesian framework. The computation of such inference is intensive due to the doubly stochastic property, i.e., the model is a hierarchical combination of a Poisson process and a Gaussian Process (GP), which leads to an intractable integral over an infinite-dimensional random function. As a result of these challenges a number of computational techniques have been proposed for Bayesian inference. These include Hamiltonian Monte Carlo (HMC), Integrated Nested Laplace Approximation (INLA) and Variational Bayes (VB). Taylor and Diggle (2012) compared MCMC based on the Metropolis Adjusted Langevin algorithm (MALA) and INLA for this model; however, comparisons between HMC, INLA, and VB have not been considered previously. In this talk we describe these comparisons in terms of accuracy and computational efficiency using simulation studies as well as through applications to ecology and brain imaging.

email: **tengming@umich.edu**

THE JOINT ASYMPTOTICS FOR ESTIMATING THE SMOOTHNESS PARAMETERS OF BIVARIATE GAUSSIAN RANDOM PROCESS

Yuzhen Zhou*, Michigan State University

Yimin Xiao, Michigan State University

Characterizing the dependence structure of the multivariate random field plays a key role in multivariate spatial model setting. Usually, the covariance structure for each component of the multivariate process is highly related to the smoothness of the surface. The estimation of smoothness parameters in univariate model has been studied extensively. Yet, there is few work in the multivariate case. In this paper, we first propose an estimation procedure for the smoothness parameters of bivariate Gaussian process. Then we investigate the joint asymptotics of the estimators and study how the cross dependence structure would affect the performance of the estimators.

email: **zhouyuzh@msu.edu**

COVARIANCE TAPERING FOR ANISOTROPIC NONSTATIONARY GAUSSIAN RANDOM FIELDS WITH APPLICATION TO LARGE SCALE SPATIAL DATA SETS

Abolfazl Safikhani*, Michigan State University

Yimin Xiao, Michigan State University

Estimating the covariance structure of spatial random processes is an important step in spatial data analysis. Maximum likelihood estimation is a popular method in spatial models based on Gaussian random fields. But calculating the likelihood in large scale data sets is computationally infeasible due to the heavy computation of the precision matrix. One way to mitigate this issue, which is due to Furrer et al. (2006), is to "taper" the covariance matrix. While most of the results in the current literature focus on isotropic tapering for stationary Gaussian processes, there are many cases in application that require modeling of anisotropy and/or nonstationarity. In this article, we propose a nonstationary parametric model, in which the underlying Gaussian random field may have different regularities in different directions, thus can be applied to model anisotropy. Using the theory of equivalence of Gaussian measures under nonstationary assumption, strong consistency of the tapered likelihood based estimation of the variance component under fixed domain asymptotics are derived by putting mild conditions on the spectral behavior of the tapering covariance function. The procedure is illustrated with numerical simulation.

email: **safikhan@stt.msu.edu**

DYNAMIC NEAREST NEIGHBOR GAUSSIAN PROCESS MODELS FOR LARGE SPATIO-TEMPORAL DATASETS

Abhirup Datta*, University of Minnesota

Sudipto Banerjee, University of California, Los Angeles

Andrew O. Finley, Michigan State University

Gaussian process models for analyzing large spatial or spatio-temporal datasets involve large dense matrix computations rendering them infeasible. Nearest Neighbor Gaussian Process (NNGP) models based on local neighborhoods provide a scalable alternative for large spatial datasets. We extend this idea to construct dynamic local neighborhoods in a continuous spatio-temporal domain using strength of a correlation function as a proxy for distance. We develop a dynamic Nearest Neighbor Gaussian Process which yields finite dimensional Gaussian densities with sparse precision matrices. We use the dynamic NNGP as a sparsity inducing prior in a hierarchical spatio-temporal setup. We provide an algorithm for fast updates of the dynamic neighborhoods and show that the total storage and computation costs of a Markov Chain Monte Carlo (MCMC) iteration for this model are proportional to the size of the dataset thereby ensuring massive scalability. We demonstrate the computational and inferential benefits of the dynamic NNGP over other competing methods using real and synthetic datasets.

email: **datta013@umn.edu**

20. CONTRIBUTED PAPERS: Case Studies in Longitudinal Data Analysis

USING THE SIGMOID MIXED MODELS FOR LONGITUDINAL COGNITIVE DECLINE

Ana W. Capuano*, Rush University Medical Center

Robert S. Wilson, Rush University Medical Center

Sue E. Leurgans, Rush University Medical Center

Jeffrey D. Dawson, University of Iowa

Donald Hedeker, University of Chicago

Random-effects linear mixed models are widely used to analyze longitudinal cognitive decline. Often, however, trajectories are non-linear. For example, terminal cognitive decline is characterized by a decline that is faster proximate to death. Adding a quadratic term for time, or considering two linear slopes (e.g. random change point model) may not properly characterize the trajectories. We describe a random-effects non-linear mixed model with covariates for such longitudinal data, based on sigmoidal logistic curves. The most general of the models include five parameters, representing: level at death, level before decline, rate of decline, decline midpoint, and asymmetry. To illustrate the applicability of the approach, we fit a random-effects sigmoid model to cognitive data from deceased participants in two longitudinal studies: Rush Religious Order Study, and Rush Memory and Aging Project. We present simulation results that illustrate the model empirical properties.

email: **ana_capuano@rush.edu**

SHORT-TERM BLOOD PRESSURE VARIABILITY OVER 24 HOURS USING MIXED-EFFECTS MODELS

Jamie M. Madden*, University College Cork, Ireland

Xia Lee, University College Cork, Ireland

Patricia M. Kearney, University College Cork, Ireland

Anthony P. Fitzgerald, University College Cork, Ireland

The benefits of using ambulatory blood pressure measurements (ABPM) in addition to clinic measurements in the management of hypertension are well established. As well as mean day, night and dip values, measures of short-term blood pressure variability (BPV) can also be obtained from ABPM. Long term BPV has been associated with cardiovascular events but the prognostic significance of short-term BPV remains uncertain. The majority of studies have focused on summary measures of BPV such as standard deviation but there is uncertainty in how accurately these indexes capture the true variability. We obtained data from the Mitchelstown Study, a cross-sectional study of Irish adults aged 47-73 years (n=2,047). A subsample (1,207) underwent 24-h ABPM. In addition to using traditional measures of variability

such as standard deviation this analysis makes full use of the longitudinal and circadian nature of ABPM data by applying mixed-effects models to determine subject-specific trajectories over time. The variation about subject-specific trajectories was taken as a measure of an individual's BPV. Additionally, the association between this measure of variability and subclinical target organ damage (documented by microalbuminuria and ECG left ventricular hypertrophy) was then examined using logistic regression. Results will be presented and the findings will be discussed.

email: **jamiem1234@gmail.com**

A LONGITUDINAL MODELLING CASE STUDY IN RENAL MEDICINE AND AN ASSOCIATED R PACKAGE

Ozgur Asar*, Lancaster University

Peter J. Diggle, Lancaster University and University of Liverpool

James Ritchie, University of Manchester

Philip A. Kalra, University of Manchester

Kidney health is monitored by blood biomarkers, principally serum creatinine level. An increase in creatinine level is indicative of worsening kidney function. Acute kidney injury (AKI) is defined as a sudden fall in kidney function. For instance, stage 1 AKI is defined as a 1.5-fold increase in creatinine level within 48 hours. The influence of AKI occurrence on subsequent kidney health is still an open research area. In this study, our main aim is to compare the level and slope of kidney function regarding pre and post an AKI event. For this purpose, we developed a continuous-time linear mixed model with three random components, namely random intercept, a non-stationary Gaussian process and measurement error. We also provide an R package, lmenssp, to fit this class of models. The case-study uses data from the Chronic Renal Insufficiency Standards Implementation Study, an ongoing cohort study based at Salford Royal Hospital, Greater Manchester.

email: **o.asar@lancaster.ac.uk**

A LIKELIHOOD RATIO TEST FOR NESTED PROPORTIONS

Yi-Fan Chen*, University of Illinois, Chicago

Jonathan Yabes, University of Pittsburgh

Maria Brooks, University of Pittsburgh

Sonia Singh, Royal Children's Hospital

Lisa Weissfeld, Statistics Collaborative Inc.

For policy and medical issues, it is important to know if the proportion of an event changes after an intervention. When the later proportion can only be calculated in a portion of the sample used to compute the previous proportion, the two proportions are nested. The motivating example is to test whether admission rates in emergency departments are different between the first and a return visit. Here, relatively small subjects who contribute to the admission rate at the return visit must be included in the first rate and also return, but not vice versa. This conditionality makes existing methods, such as longitudinal data analysis, not directly applicable, and researchers can only explore this question by using descriptive statistics. We propose a likelihood ratio test to compare two nested proportions by using the product of conditional probabilities. This test accommodates the conditionality, subject dependencies and cluster effects and can be implemented in SAS PROC NLMIXED easily. We evaluate the properties of our approach and compare it with the two-sample proportion z-test and the Cochran–Mantel–Haenszel test via simulations. An example based on readmission rates through an emergency department is used to illustrate the proposed method.

email: **yfchen2@uic.edu**

BAYESIAN NONPARAMETRIC QUANTILE REGRESSION MODELS: AN APPLICATION TO A FETAL GROWTH STUDY WITH 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

The appropriate interpretation of monitored fetal growth throughout pregnancy in individual fetus and population is dependent on the availability of adequate standards. The focus of this paper is on developing Bayesian nonparametric quantile regression models to develop contemporary U.S. fetal growth standards for racial/ethnic groups of pregnant women. The proposed method relies on assuming the asymmetric Laplace distribution as auxiliary error distribution. We also consider the covariates-dependent

random partition models that the probability of any particular partition is allowed to depend on covariates. This leads to random clustering models indexed by covariates, i.e., quantile regression models with the outcome being a partition of the experimental units. 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.

email: **kims2@mail.nih.gov**

MODELING REPEATED LABOR CURVES IN CONSECUTIVE PREGNANCIES: INDIVIDUALIZED PREDICTION OF LABOR PROGRESSION FROM PREVIOUS PREGNANCY DATA

Olive D. Buhule*, 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

Alexander C. McLain, University of South Carolina

Katherine Grantz, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Measuring cervical dilation in the late stage of pregnancy is a commonly used technique for monitoring the progression of labor. Recent statistical methodology has been developed to address the analytical challenges for such data when only a single labor curve is observed on each woman (McLain and Albert, 2014, Biometrics). These challenges include conducting valid inference and prediction when there is not a time zero (i.e., when women enter the hospital at different stages of their labor). Motivated by the NICHD Consecutive Pregnancy Study (CPS), a unique cohort study that collected repeat labor data on over 50,000 women, we propose new methodology for analyzing labor curves across multiple pregnancies. Our focus is on using the cervical dilation data from prior pregnancies to predict subsequent labor curves. We propose a hierarchical random effects model with random change points that characterizes repeated labor curves within and between women. We employ Bayesian methodology (MCMC) for parameter estimation and prediction. The methodology was used in analyzing the CPS data, and in developing a predictor for labor progression that can be used in clinical practice.

email: **odb3@pitt.edu**

AN EXAMPLE OF UNCONSTRAINED MODEL FOR COVARIANCE STRUCTURE FOR MULTIVARIATE LONGITUDINAL DATA: MAJOR LEAGUE BASEBALL BATTER'S SALARY WITH THE WEIGHTED OFFENSIVE AVERAGE

Chulmin Kim*, University of West Georgia

The positive-definiteness requirement for the covariance matrix may impose complicated nonlinear constraints on the parameters. Kim (2012) proposed an unconstrained parameterization for the covariance structure for the multivariate longitudinal data, and then to model its parameters parsimoniously. Kim (2013) also introduced the weighted offensive average (WOA) as a variation of on base plus slugging (OPS) which explains not only a batter's hitting performance but also his non-hitting performance to generate runs for his team such as stolen bases, walks, and etc. We adopt Kim's unconstrained model for the covariance structure for Major League Baseball batter's salary with the Weighted Offensive Average.

email: **ckim@westga.edu**

21. CONTRIBUTED PAPERS: Meta Analysis

META-ANALYSIS SPARSE K-MEANS FRAMEWORK FOR DISEASE SUBTYPE DISCOVERY WHEN COMBINING MULTIPLE TRANSCRIPTOMIC STUDIES

Zhiguang Huo*, University of Pittsburgh

George 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 the first step towards this goal. In this paper, we extend a sparse K-means method towards a meta-analytic 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 signatures across studies. The method was evaluated by leukemia and breast cancer data sets. The identified disease subtypes from meta-analysis were characterized with improved accuracy and stability compared to single study analysis. The breast cancer model was applied to an independent METABRIC dataset and generated improved survival difference between subtypes. These results provide a basis for diagnosis and development of targeted treatments for disease subgroups.

email: **zhh18@pitt.edu**

META ANALYSIS: A CAUSAL FRAMEWORK, WITH APPLICATION TO RANDOMIZED STUDIES OF VIOXX

Michael E. Sobel*, Columbia University

David Madigan, Columbia University

Wei Wang, Columbia University

We construct a framework for meta-analysis that helps to clarify and empirically examine the sources of between study heterogeneity in treatment effects. The key idea is to consider, for each of the treatments under investigation, the subject's potential outcome in each study were he to receive that treatment. We consider four sources of heterogeneity: 1) response inconsistency, whereby a subject's response to a given treatment varies across different studies, 2) the grouping of non-equivalent treatments, where two or more treatments are grouped and treated as a single treatment under the incorrect assumption that a subject's responses to the different treatments would be identical, 3) non-ignorable treatment assignment, and 4) response related variability in the composition of subjects in different studies. We then examine the implications of these assumptions for heterogeneity/homogeneity of conditional and unconditional treatment effects. To illustrate the utility of our approach, we re-analyze individual patient data from 29 randomized placebo controlled studies of Vioxx on the cardio-vascular risk of Vioxx, a Cox-2 selective non- steroidal anti-inflammatory drug approved by the FDA in 1999 for the management of pain and withdrawn from the market in 2004.

email: **mes105@columbia.edu**

A BAYESIAN HIERARCHICAL MODEL FOR NETWORK META-ANALYSIS OF DIAGNOSTIC TESTS

Xiaoye Ma*, University of Minnesota

Haitao Chu, University of Minnesota

Yong Chen, University of Texas Health Science Center, Houston

Joseph Ibrahim, University of North Carolina, Chapel Hill

To compare the accuracy of multiple tests in a single study, three designs are commonly used: 1) the multiple test comparison design; 2) the randomized design and 3) the non-comparative design. Existing meta-analysis methods of diagnostic tests (MA-DT) have been focused on evaluating the performance of a single test by comparing it with a reference test. The increasing number of available diagnostic instruments for a disease condition and the different study designs being used have generated the need to develop efficient and flexible meta-analysis framework to combine all designs for simultaneous inference. In this paper, we develop a missing data framework and a Bayesian hierarchical model for network meta-analysis of diagnostic tests (NMA-DT) and offer important promises over the traditional MA-DT: 1) it combines studies using all three designs; 2) it pools both studies with or without a gold standard; 3) it combines studies with different sets of candidate tests; and 4) it accounts for heterogeneity across studies and complex correlation structure among multiple tests. We illustrate our method through a NMA of deep vein thrombosis tests. Finally, we evaluate the performance of the proposed method through simulation studies.

email: **maxxx372@umn.edu**

INFERENCE FOR CORRELATED EFFECT SIZES USING MULTIPLE UNIVARIATE META-ANALYSES

Yong Chen, University of Texas Health Science Center, Houston

Yi Cai*, University of Texas Health Science Center, Houston

Chuan Hong, University of Texas Health Science Center, Houston

Dan Jackson, Cambridge Institute of Public Health

Multivariate meta-analysis, which involves jointly analyzing multiple and correlated outcomes from separate studies, has received a great deal of attention. One reason to prefer the multivariate approach is because of its ability to account for the

dependence between multiple estimates from the same study. However, nearly all the existing methods for analyzing multivariate meta-analytic data require the knowledge of the within-study correlations, which are usually unavailable in practice. We propose a simple non-iterative method that can be used for the analysis of multivariate meta- analysis datasets that has no convergence problems and does not require the use of within-study correlations. Our approach uses standard univariate methods for the marginal effects but also provides valid joint inference for multiple parameters. The proposed method can directly handle missing outcomes under missing completely at random assumption. Simulation studies show that the proposed method provides unbiased estimates, well-estimated standard errors and confidence intervals with good coverage probability. Furthermore, the proposed method is found to maintain high relative efficiency compared to conventional multivariate meta-analyses where the within-study correlations are known. We illustrate the proposed method through two real meta-analyses where functions of the estimated effects are of interest.

email: **yi.cai@uth.tmc.edu**

DETECTING OUTLYING STUDIES IN META-REGRESSION MODELS USING A FORWARD SEARCH ALGORITHM

Dimitris Mavridis, University of Ioannina

Irini Moustaki*, London School of Economics

Melanie Wall, Columbia University

Georgia Salanti, University of Ioannina

Meta-analysis summarizes evidence from many studies addressing the same research hypothesis and is one of the most influential and powerful techniques underpinning evidence-based practice. When considering data from many trials, it is likely that some of them present a markedly different intervention effect or exert an undue influence on the summary results and, subsequently, on policy decision-making. The Cochrane Collaboration recommends the application of a random-effects meta-analysis both with and without outlying studies. Here, we develop a forward search algorithm for identifying outlying studies in meta-analysis models. The forward search algorithm starts by fitting the hypothesized model to a small subset of studies and proceeds by adding studies that are determined to be close to the fitted model. We monitor estimated parameters, measures of fit and Cook's distances and identify outliers by sharp changes in their forward plots. The suggested methodology allows us to test if a change in a statistic being monitored is caused by the study entering the search or can be attributed to random variation. We apply the method to a meta-analysis that examines the effect of writing-to-learn interventions on academic achievement adjusting for three possible effect modifiers and compare results to other outlier detection strategies and to data from medical research.

email: **i.moustaki@lse.ac.uk**

COMPARING MULTIPLE IMPUTATION METHODS FOR SYSTEMATICALLY MISSING SUBJECT-LEVEL DATA

David M. Kline*, The Ohio State University

Eloise E. Kaizar, The Ohio State University

Rebecca R. Andridge, The Ohio State University

When conducting research synthesis, the collection of studies that will be combined often do not measure the same set of variables, which creates missing data. Traditionally, the focus of missing data methods for longitudinal data has been on missing observation-level (time-varying) variables. In this paper, we focus on missing subject-level (non-time-varying) variables and compare two multiple imputation approaches, a joint modeling approach and a sequential conditional modeling approach, for modeling missing data of this type. We find the joint modeling approach to be preferable to the sequential conditional approach except when the covariance structure of the repeated outcome for each individual has homogenous variance and exchangeable correlation. Specifically, the regression coefficient estimates from an analysis incorporating imputed values based on the sequential conditional method are attenuated and less efficient than those from the joint method. Remarkably, the estimates from the sequential conditional method are often less efficient than a complete case analysis, which, in the context of research synthesis, implies that we lose efficiency by combining studies.

email: **kline.273@osu.edu**

22. CONTRIBUTED PAPERS: Semi-Parametric Methods

UNDERSTANDING GAUSSIAN PROCESS FITS USING AN APPROXIMATE FORM OF THE RESTRICTED LIKELIHOOD

Maitreyee Bose*, University of Minnesota

James S. Hodges, University of Minnesota

Gaussian processes (GPs) are widely used in statistical modeling. A GP is often used as the random effect in a linear mixed model, with its unknowns estimated by maximizing the log restricted likelihood or using a Bayesian analysis, which are closely related. However, it is unclear how the process variance, range, and error variance are fit to features in the data. In this paper, we aim to gain a better understanding of how GP parameters are fit to data. To do so, we need a simple, interpretable, and fast-computing form of the restricted likelihood. This is achieved by applying the spectral approximation to the GP and representing it as a linear mixed model. The log restricted likelihood from this approximate model has a scalarized form and is identical to the log likelihood arising from a gamma-errors generalized linear model (GLM) with the identity link. We use this GLM representation to make conjectures about how GP parameters are fit to data, and investigate our conjectures by introducing features in simulated data, like outliers and mean-shifts, and observing how introduction of these features affects the GP parameter estimates.

email: **bosex020@umn.edu**

MITIGATING BIAS IN GENERALIZED LINEAR MIXED MODELS: THE CASE FOR BAYESIAN NONPARAMETRICS

Joseph L. Antonelli*, Harvard School of Public Health

Sebastien Haneuse, Harvard School of Public Health

Lorenzo Trippa, Harvard School of Public Health

Generalized linear mixed models (GLMMs) use random effects to account for correlation in clustered or longitudinal data. The random effects follow some unknown distribution, G, and in practice this is taken to be a Normal distribution. If this assumption does not hold, however, the model is misspecified and estimation/inference may be invalid. An alternative is to adopt a Dirichlet process (DP) prior for G in a Bayesian analysis. Conventional wisdom suggests that the increased flexibility reduces bias, although this has not been thoroughly examined. Furthermore, the extent to which the increased flexibility confers a bias-variance trade-off has not been examined. Under a range of `true' random effects distributions, we examine operating characteristics for estimation of fixed and random effects in a GLMM using a DP prior for G. Strategies for the specification of the precision parameter in the DP prior are also investigated. We conclude that while no single model is likely to work well in all settings, the use of the DP prior in a GLMM mitigates much if not all of the bias that arises when one incorrectly assumes a Normal distribution, with little-to-no penalty paid in terms of efficiency.

email: **jantonelli@fas.harvard.edu**

AN ESTIMATED LIKELIHOOD ESTIMATOR BY EXTRACTING AUXILIARY INFORMATION UNDER OUTCOME DEPENDENT SAMPLE DESIGN

Wansuk Choi*, University of North Carolina, Chapel Hill

Haibo Zhou, University of North Carolina, Chapel Hill

Outcome dependent sampling(ODS) has been studied by many researchers because it is a cost effective design. In case of easily obtainable outcome, researchers can have responses of every member in a study-population. However, it can be difficult to have all covariate of interest information from a study-population. In this situation, even though missing data in covariates problem exists, researchers can obtain auxiliary covariates information from all members in population. Weaver and Zhou(2005) showed that, rather than simple random sampling, ODS design could improve of estimator's property in terms of efficiency. And ODS design provided unbiaseness and consistency to estimators. In this article, we propose a method under the situation that we can only have ODS samples but not a whole population. We assume that SRS part of ODS has missing covariate and supplemental sample part of ODS has missing covariate. In addition, every member of ODS sample has a binary auxiliary variable, which is related to covariates of interest. The proposed method use auxiliary information to estimate nonparametric parts in the likelihood to derive an estimated likelihood. The finite sample performance of the proposed method is studied, compared to other existing methods.

email: **choiwan@live.unc.edu**

ESTIMATION, IID REPRESENTATION AND INFERENCE FOR THE AVERAGE OUTCOME UNDER STOCHASTIC INTERVENTION ON DEPENDENT DATA

Oleg Sofrygin*, University of California, Berkeley

Mark J. van der Laan, University of California, Berkeley

We describe targeted minimum loss-based estimation (TMLE) for the mean outcome under joint stochastic intervention in dependent data. Suppose data on N units is observed, each unit i is O_i=(F_i,W_i,A_i,Y_i), where F_i – set of units making up i's "network", W_i - baseline covariates, A_i - binary exposure, Y_i - binary outcome. We assume A_i is a function of W_i; Y_i is a function of (A_i,W_i). Dependence between units is added by assuming A_i depends on covariates of units in F_i; Y_i depends on covariates and exposures of units in F_i. We propose a semi-parametric model for the observed data and focus on estimating the expected value of the average outcome \bar{Y}, where \bar{Y} is the average of Y_i over N units and expectation is taken wrt some known joint stochastic intervention on exposures. We describe TMLE for above quantity and demonstrate it is a doubly robust, asymptotically efficient and asymptotically linear estimator. We also demonstrate how our statistical quantity of interest can be represented as a mapping from certain IID data distribution, which happens to be a function of the true distribution of O and stochastic intervention. This insight leads to a simplified estimator of the asymptotic variance for above TMLE, performance of which is then evaluated in a simulation study.

email: **sofrygin@berkeley.edu**

EMPIRICAL LIKELIHOOD-BASED INFERENCE FOR PARTIALLY LINEAR MODELS

Haiyan Su*, Montclair State University

We propose an empirical likelihood (EL)-based inference for the linear component coefficient in partially linear models and partially linear mixed-effect models. The proposed method combines the projection method with the EL method. The project method is used to remove the nuisance parameter in the model and then EL method is used to construct confidence intervals for the linear component. Bartlett correction method is used to correct the EL-based confidence intervals. The test statistic is shown to follow regular chi-square distribution asymptotically. The numerical performance of the method under normal and non-normal error terms is evaluated through simulation studies.

email: **suh@mail.montclair.edu**

BAYESIAN NONPARAMETRIC METHODS FOR TESTING SHAPE CONSTRAINT FOR LONGITUDINAL DATA

Yifang Li*, North Carolina State University

Sujit Ghosh, North Carolina State University & Statistical and Applied Mathematical Sciences Institute

In various applications of longitudinal data analysis, we often have subject matter knowledge about the population that may suggest a specific shape of the unknown mean curve over a given time period of interest. However, due to the variability across subjects or sites or lack of experimental scientific evidence, it may not be obvious to detect a specific shape of the population level trend based on sparsely observed data. For example, it is widely believed and debated that global temperature might be on rise over the last century based on observations taken at various locations around the globe, but a definitive answer is still lacking. Mixed-effect model is a commonly used tool to account for variations across different subjects or sites in longitudinal analysis. This paper develops a nonparametric Bayesian method to test various shape constraint of the population level mean trend based on approximating a Gaussian process using a sequence of penalized splines whose coefficients are allowed with vary with subjects or sites. Posterior consistency of the test procedure is established under a set of regularity conditions and numerical illustrations are presented based on simulated and real data sets.

email: **yli40@ncsu.edu**

HYPOTHESIS TESTING IN SEMI-PARAMETRIC DISCRETE CHOICE MODEL

Yifan Yang*, University of Kentucky

Mai Zhou, University of Kentucky

Discrete choice model is widely used in social sciences and economics. (Wang and Zhou, 1995) proposed a least square type of estimation, which is difficult to derive the confidence interval or perform a hypothesis test on the regression coefficients. In this paper, a semi-parametric approach was described to solve these

problems. The proposed method was based on empirical likelihood method (Own, 2001) and used the interpretation of the Expectation and Maximization (EM) principle (Zhou, 2002). Simulation work shows that the log likelihood ratio is standard chi-square distributed hence can be used to construct confidence interval for regression coefficients.

email: **yifan.yang@uky.edu**

23. Trends and Innovations in Clinical Trial Statistics: "The Future ain't what it Used to be"

"THE FUTURE AIN'T WHAT IT USED TO BE" (YOGI BERRA). HAVE STATISTICIANS RECEIVED THE MEMO?

Nevine Zariffa*, AstraZeneca Pharmaceuticals

We will review key contributions from statisticians in healthcare advancement: past, present and future. A brief tour of history and an analysis of the current situation will allow us to consider future directions and challenges. We will explore opportunities to fundamentally transform the healthcare system, and how our discipline best evolves. Key topics will include: the patient perspective, quality in all its incarnations, the evolving data eviron-ment, and effective communication.

email: **nevine.zariffa@astrazeneca.com**

PANELISTS:

Sara Hughes, GlaxoSmithKline

Dominic Labriola, Bristol-Myers Squibb

Lisa LaVange, U.S. Food and Drug Administration

Shiferaw Mariam, Janssen R&D

Jerry Schindler, Merck

Venkat Sethuraman, Bristol-Myers Squibb

Frank Shen, AbbVie

Anastasios (Butch) Tsiatis, North Carolina State University

24. Causal Inference in HIV/AIDS Research

REPRESENTING UNMEASURED CONFOUNDING IN CAUSAL MODELS FOR OBSERVATIONAL DATA

Joseph W. Hogan*, Brown University

Dylan Small, University of Pennsylvania

Arguably the most important assumption needed for drawing causal inference from observational data is "ignorable treatment assignment" or "no unmeasured confounding." In potential outcomes formulations of causal effect, "no unmeasured confounders" states that, conditionally on a specific set of measured covariates (confounders), the potential outcomes are independent of exposure. When the assumption holds, valid inferences about causal effect can be obtained using methods such as inverse probability weighting, propensity score methods, and g-estimation. However, because "no unmeasured confounding" is untestable, researchers in statistics, epidemiology, and econometrics have developed a variety of methods for representing unmeasured confounding and quantifying potential biases. We review, compare, and draw connections between two distinct approaches. The first represents unmeasured confounding as a latent random variable that has a specific dependence structure with the outcome and exposure. The second treats the unobserved potential outcome as the sole unmeasured confounder. Each approach gives rise to specific methods for sensitivity analysis and bias quantification; these will be illustrated and compared using both simulation and data analysis. We also examine whether and how estimated sensitivity to bias from unmeasured confounding depends on the set of measured confounders being used for adjustment.

email: **jhogan@stat.brown.edu**

INVERSE PROBABILITY OF CENSORING WEIGHTS UNDER MISSING NOT AT RANDOM WITH APPLICATION TO CD4 OUTCOMES IN HIV-POSITIVE PATIENTS IN KENYA

Judith J. Lok*, Harvard School of Public Health

Constantin T. Yiannoutsos, Indiana University Fairbanks School of Public Health

Agnes Kiragga, Infectious Diseases Institute, Kampala, Uganda

Ronald J. Bosch, Harvard School of Public Health

Right-censoring is Missing Not At Random (MNAR) when the prognosis of patients after censoring is different from the prognosis of patients in follow-up, even given observed characteristics just prior to the dropout time. Analyzing MNAR data is complicated. Often,

bounds and sensitivity analyses are the only option. We propose a method to obtain point estimates for the trajectory of a mean/median over time when dropout is MNAR, if so-called outreach data are available: additional data on a subsample of patients lost-to-follow-up, successfully located afterwards. We propose an extension of Inverse Probability of Censoring Weighting to this setting. We illustrate our method by estimating the response to antiretroviral therapy (ART) among HIV-positive patients in Kenya. The available data are MNAR: more patients in the outreach sample died shortly after dropout than expected on the basis of their initially observed covariates, and more patients in the outreach sample were off treatment, in part because of limited access to ART outside the program evaluated. Taking MNAR into account leads to a substantial downward adjustment of the response to ART.

email: **jlok@hsph.harvard.edu**

DOUBLY ROBUST INSTRUMENTAL VARIABLE ESTIMATION FOR OUTCOME MISSING NOT AT RANDOM

BaoLuo Sun*, Harvard School of Public Health

Lan Liu, Harvard School of Public Health

James Robins, Harvard School of Public Health

Eric Tchetgen Tchetgen, Harvard School of Public Health

Missing data occurs frequently in epidemiologic practice, compromising our ability to make accurate inferences. In particular, the outcome of interest may not be observed for a subset of the sample. The outcome is said to be missing not at random (MNAR) if, conditional on the observed variables, it is still dependent on the unobserved outcomes. Under such settings, identification is generally not possible without imposing additional assumptions. Identification is sometimes possible, however, if an exogeneous instrumental variable (IV) is observed for all subjects such that it satisfies exclusion restriction, and that the IV affects the missing data process without directly influencing the outcome. In this presentation, we propose a nonparametric method for identification, followed by inverse probability weighted (IPW) and regression estimators, which give consistent estimates for the average response if either the propensity score or the outcome regression model is correct, respectively. Lastly, we propose a doubly robust estimator that remains consistent if either of the above models is true.

email: **bluosun@gmail.com**

ESTIMATING PREVENTION EFFICACY AMONG COMPLIERS IN HIV PRE-EXPOSURE PROPHYLAXIS (PrEP) TRIALS

James Dai*, Fred Hutchinson Cancer Research Center and University of Washington

Elizabeth Brown, Fred Hutchinson Cancer Research Center and University of Washington

Adherence is vital to success of PrEP trials. Pharmacological measures of adherence have been widely used, typically in case-control or case-cohort samples in the active product arms. Using the MTN-003 trial as an example, we investigate several analytical strategies to infer prevention efficacy among a group of women who used study product based on the plasma TFV detection, including causal potential outcomes methods and confounding adjustment. Merits and limitations of these approaches will be discussed. In particular, we show that when adherence is moderate or low, the power of the potential outcomes approach can be low. Using the exclusion restriction assumption, we show that the conventional confounding-adjustment approach can be useful in assessing whether selection bias has been adequately removed.

email: **jdai@fhcrc.org**

25. Open Problems and New Directions in Neuroimaging Research

OPEN PROBLEMS IN STRUCTURAL BRAIN IMAGING: WAVELETS AND REGRESSIONS ON NON-EUCLIDEAN MANIFOLDS

Moo K. Chung*, University of Wisconsin, Madison

Structural brain images such as MRI and DTI are inherently geometric in nature. However, many processing and analysis techniques used in the field are Euclidean and do not explicitly incorporate the non-Euclidean nature of data and space. We present a unified differential geometric framework for performing kernel smoothing, regressions and wavelet transforms on non-Euclidean space such

as the space of positive definite symmetric matrices, curved 3D curved surfaces and 4D hypersphere. The framework is applied in modeling brain networks and subcortical brain surfaces.

email: **mkchung@wisc.edu**

OPEN PROBLEMS AND NEW DIRECTIONS IN MODELING ELECTROENCEPHALOGRAMS

Hernando Ombao*, University of California, Irvine

Quantitative neuroscience is a flourishing discipline where statistics plays a critical role. With studies collecting repeated measurements on thousands of subjects over multiple years, the size of data sets are becoming larger and more complex. Given the complexity and size of neuroimaging data, simple reproducible data-analytic methods, data exploration and careful design of experiments will become increasingly important and will require the expertise of statisticians. This is creating significant new demand and unmatched opportunity for statisticians working in the field. This talk will focus on open problems for analyzing electroencephalograms (EEG) which are non-invasive measurements of brain electrical activity. These are projections of the unobserved neuronal electrical activity on the scalp. With excellent temporal resolution (approximately 1000 samples per second), they capture both oscillatory activity at the cortex and connectivity between the cortical regions. Under designed experiments, there is a need to develop flexible models that correctly capture variation in brain responses across subjects and differences across conditions and patient groups. Moreover, it is important to model and estimate the latent neuronal sources by combining information across several trials for each subject. Finally, since EEG data is massive, it is important to develop computationally efficient algorithms for estimation and inference.

email: **hernando.ombao@gmail.com**

OPEN PROBLEMS AND NEW DIRECTIONS IN FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI)

Martin A. Lindquist*, Johns Hopkins University

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive technique for studying brain activity. During the past two decades fMRI has provided researchers with an unprecedented access to the inner workings of the brain, leading to countless new insights into how the brain processes information. The field that has grown around the acquisition and analysis of fMRI data has experienced a rapid growth in the past several years and found applications in a wide variety of fields. In this talk we will discuss several new directions in the analysis of fMRI data. These include open problems in brain connectivity, brain decoding and multi-modal data analysis.

email: **mlindqui@jhsph.edu**

EMPIRICAL BAYES METHODS LEVERAGING HERITABILITY FOR IMAGING GENETICS

Wesley Kurt Thompson*, University of California, San Diego

Brain imaging and genetics both produce very high dimensional data. Associating brain phenotypes with genetics data thus leads to the curse of dimensionality squared. We tackle this problem from two directions. First, using data from twin imaging studies, we reduce the dimensionality of the brain phenotype by clustering voxels that share genetic influences. Second, for associating genetic variation to these new phenotypes we propose a novel resampling-based methodology that obtains non-parametric estimates of replication effect sizes from Genome-Wide Association (GWA) data. From replication effect sizes we can compute a number of parameters of interest, including the local false discovery rate, the tagged heritability, and polygenic estimates of predicted phenotypic values in de novo subjects. Low locfdr indicates high probability of being non-null. We also develop an extension of this methodology, termed “covariate modulated local false discovery rate” (cmlocfdr),that incorporates functional annotations and pleiotropic relationships of one phenotype with another to leverage information from multiple GWAS. We demonstrate these new methodologies on a GWAS of several thousand subject using brain morphology phenotypes (cortical thickness, surface area, and subcortical volumes).

email: **wktwktwkt@gmail.com**

26. Statistical Methods for Understanding Whole Genome Sequencing

GROUP ASSOCIATION TEST USING A HIDDEN MARKOV MODEL FOR SEQUENCING DATA

Charles Kooperberg*, Fred Hutchinson Cancer Research Center

Yichen Cheng, Fred Hutchinson Cancer Research Center

James Y. Dai, Fred Hutchinson Cancer Research Center

With next generation sequencing data, group association tests are of great interest, because the power of testing for association of a single genomic feature at a time is often very small, as well as the small effect sizes, due to the overwhelming number of individual genomic features. Many methods have been proposed to test association of a trait with a group of features, e.g. all variants in a gene, yet few of these methods account for the fact that a substantial proportion of the features are not associated with the trait. We propose to model the association for each feature in the group as a mixture of no association or a constant non-zero association to account for the fact that a fraction of features may not be associated with the trait even if other features in the group are. The observed individual associations are first estimated by generalized linear models; the sequence of these estimated associations are then modeled by a hidden Markov chain. To test for association, we use a modified likelihood ratio test based on an independence log-likelihood with additional penalty term, for which we derive the asymptotic distribution under the null hypothesis.

email: **clk@fhcrc.org**

VARIANT CALLING AND BATCH EFFECTS IN DEEP WHOLE-GENOME SEQUENCING DATA

Margaret A. Taub*, Johns Hopkins University

Suyash S. Shringarpure, Stanford University

Rasika A. Mathias, Johns Hopkins University

Ingo Ruczinski, Johns Hopkins University

Kathleen C. Barnes, Johns Hopkins University and The CAAPA Consortium

Whole genome sequencing studies with thousands of samples are currently underway. Statistical challenges in working with these massive data sets arise at all phases of data analysis, from initial measurement of data quality to development of appropriate methods for testing genetic hypotheses and interpreting observed patterns of genetic variation. Here, we present work using 642 samples from the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) who were whole-genome sequenced to an average depth of 30x. We first performed a comparison of different variant calling algorithms, focusing on characteristics of variants called by different subsets of callers. We then developed a quality-control classifier which uses genotyping array data, typically collected for all sequenced individuals, as a gold standard training set to improve calibration of variant calls. One finding of interest is that there is little difference in overall quality between single-sample and multi-sample calling methods at the depth of coverage in our data set. In addition, we examined our data set for evidence of batch effects to detect possible confounders in our downstream analysis.

email: **mtaub@jhsph.edu**

FLEXIBLE PROBABILISTIC MODELING OF GENETIC VARIATION IN GLOBAL HUMAN STUDIES

John Storey*, Princeton University

email: **jstorey@princeton.edu**

ALLELE SPECIFIC EXPRESSION TO IDENTIFY CAUSAL FUNCTIONAL QTLs

Barbara Englehardt*, Princeton University

email: **bee@princeton.edu**

27. Doing Data Science: Straight Talk from the Frontline

DOING DATA SCIENCE

Rachel Schutt*, Newscorp

In this talk, I will explore the question "what is data science?" Many statisticians have understandably asked, "isn't statistics the science of data?" which suggests that data science is just a rebranding of the discipline of statistics. Yet data science is clearly emerging in job titles

and academic programs, and doesn't seem to be going away any time soon. We'll discuss possible definitions of data science, and some important concepts that suggest that data science is a new and distinct discipline in its own right. I'll describe more about my role at News Corp as the Chief Data Scientist and what that job involves, and the opportunities I think that exist for statisticians both from career and research problem perspectives.

email: **rschutt@newscorp.com**

28. IMS Medallion Lecture

UNCERTAINTY QUANTIFICATION IN COMPLEX SIMULATION MODELS USING ENSEMBLE COPULA COUPLING

Tilmann Gneiting*, Heidelberg Institute for Theoretical Studies (HITS) and Karlsruhe Institute of Technology (KIT)

Roman Schefzik, Heidelberg University

Thordis L. Thorarinsdottir, Norwegian Computing Center

Critical decisions frequently rely on high-dimensional output from complex computer simulation models that show intricate cross-variable, spatial and/or temporal dependence structures, with weather and climate predictions being key examples. There is a strongly increasing recognition of the need for uncertainty quantification in such settings, for which we propose and review a general multi-stage procedure called ensemble copula coupling (ECC), proceeding as follows. 1. Generate a raw ensemble, consisting of multiple runs of the computer model that differ in the inputs or model parameters in suitable ways. 2. Apply statistical postprocessing techniques, such as Bayesian model averaging or nonhomogeneous regression, to correct for systematic errors in the raw ensemble, to obtain calibrated and sharp predictive distributions for each univariate output variable individually. 3. Draw a sample from each postprocessed predictive distribution. 4. Rearrange the sampled values in the rank order structure of the raw ensemble, to obtain the ECC postprocessed ensemble. The use of ensembles and statistical postprocessing have become routine in weather forecasting over the past decade. We show that seemingly unrelated, recent advances can be interpreted, fused and consolidated within the framework of ECC, the common thread being the adoption of the empirical copula of the raw ensemble. In some settings, the adoption of the empirical copula of historical data offers an attractive alternative. In a case study, the ECC approach is applied to predictions of temperature, pressure, precipitation, and wind over Germany, based on the 50-member European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble. Joint work with Roman Schefzik and Thordis Thorarinsdottir.

29. Panel Discussion: In Memory of Marvin Zelen: Past, Present and Future of Clinical Trials and Cancer Research

PANELISTS:

Colin Begg, Memorial Sloan Kettering Cancer Center

Dave DeMets, University of Wisconsin, Madison

Ross Prentice, Fred Hutchison Cancer Center

Victor De Gruttola, Harvard School of Public Health

30. CONTRIBUTED PAPERS: Methods for Clustered Data and Applications

MULTIVARIATE MODALITY INFERENCE WITH APPLICATION ON FLOW CYTOMETRY

Yansong Cheng*, GlaxoSmithKline

Surajit Ray, University of Glasgow

The number of modes (also known as modality) of a kernel density estimator (KDE) draws lots of interests and is important in practice. In this presentation, we develop an inference framework on the modality of a KDE under multivariate setting using Gaussian kernel. We applied the modal clustering method proposed by Li et.al. (2007) for mode hunting. A test statistic and its asymptotic distribution are derived to assess the significance of each mode. The inference procedure is applied on real flow cytometry data.

email: **chengyansong85@gmail.com**

ESTIMATION OF THE PREVALENCE OF DISEASE AMONG CLUSTERS USING RANDOM PARTIAL-CLUSTER SAMPLING

Sarah J. Marks*, University of North Carolina, Chapel Hill

John S. Preisser, University of North Carolina, Chapel Hill

Anne E. Sanders, University of North Carolina, Chapel Hill

James D. Beck, University of North Carolina, Chapel Hill

The gold-standard for estimating population prevalence of dental conditions, such as periodontitis, relies on a full mouth exam for each individual, requiring examination of up to 168 tooth sites per person. Due to time constraints, partial mouth exams that select only a subset of sites have been used in epidemiological studies; these exams, however, may underestimate the population prevalence when disease is defined entirely in terms of the selected sites. We propose a model-based approach for estimating prevalence based on the conditional linear family for correlated binary data. For random site selection, our simple estimator requires specification of only two parameters in a working model: the marginal mean assumed constant across sites and an exchangeable correlation for pairs of sites within clusters (mouths). Using oral examination data from 6,793 participants in the Arteriolosclerosis Risk in Communities Study, our proposed partial sampling method estimator produces estimates of periodontitis prevalence that are very similar to those from full mouth exams. These estimates give good precision/reproducibility for a range of cluster sizes. Our method is applicable to many areas of health research where the use of partial cluster sampling is resource-preserving , such as estimating community-level prevalence of an infectious disease.

email: **sjmarks@live.unc.edu**

TESTING HOMOGENEITY IN A CONTAMINATED NORMAL MODEL WITH CORRELATED DATA

Meng Qi*, University of Kentucky

Richard Charnigo, University of Kentucky

In this talk, we consider the problem of testing homogeneity in a contaminated normal model, when the data is correlated under some known covariance structure. To address this problem, we developed a moment based homogeneity test, assuming the data has a known compound symmetric covariance structure, and designed the weight for test statistics to increase power. We did simulations to assess size and power of the test and established asymptotic properties. In a case study, we applied our test to microarray about Down's syndrome caused by an extra copy of chromosome 21. By assuming different covariance parameters, we got a contour plot of p-values from our test, showing that failing to take into account correlation may massively understate the p-value.

email: **meng.qi@uky.edu**

ON THE USE OF BETWEEN-WITHIN MODELS TO ADJUST FOR CONFOUNDING DUE TO UNMEASURED CLUSTER-LEVEL COVARIATES

Babette A. Brumback*, University of Florida

Zhuangyu Cai, University of Florida

Between-within models are generalized linear mixed effects models for clustered data that incorporate a random intercept as well as fixed effects for both a within-cluster covariate and a between-cluster covariate, wherein the between-cluster covariate represents the cluster means of the within-cluster covariate. One popular use of these models is to adjust for confounding of the effect of within-cluster covariates due to unmeasured between-cluster covariates. Previous research has shown via simulations that using this approach can yield an inconsistent estimator. We investigate this problem further.

email: **brumback@ufl.edu**

ESTIMATING THE EFFECTS OF CENTER CHARACTERISTICS ON CENTER OUTCOMES: A SYMBOLIC DATA APPROACH

Jennifer Le-Rademacher*, Medical College of Wisconsin

This paper introduces a symbolic data approach to evaluate the effects of center-level characteristics on center outcomes. The proposed method treats centers rather than patients as the units of observation when estimating the effects of center characteristics since centers are the entities of interest. To adjust for the differences in outcomes among

centers caused by varying patient load, the effects of patient-level characteristics are first modelled treating patients as the units of observation. The outcomes (adjusted for patient-level effects) of patients from the same center are then combined into a distribution of outcomes representing that center. The outcome distributions are symbolic-valued responses on which the effects of center-level characteristics are modelled. The proposed method provides an alternative framework to analyze clustered data. This method distinguishes the effects of center characteristics from the patient characteristics effects. It can be used to model the effects of center characteristics on the mean as well as the consistency of center outcome which classical methods such as the fixed-effect model and the random-effect model cannot. This method performs well even under scenarios where the data come from a fixed-effect model or a random-effect model. The proposed approach is illustrated using a bone marrow transplant example.

email: **jlerade@mcw.edu**

A ROBUST AND FLEXIBLE METHOD TO ESTIMATE ASSOCIATION FOR SPARSE CLUSTERED DATA

Lijia Wang*, Emory University

John J. Hanfelt, Emory University

It is challenging to conduct robust inference on sparse clustered data in heterogeneous populations. For example, in a study of drinking water, researchers wanted to know whether highly credible gastrointestinal illness (HCGI) episodes tended to aggregate within households, after adjustment for demographic variables and fine stratification by geographic area. Motivated by this study, we present a composite conditional likelihood approach that yields valid inference on the intracluster pairwise association along with the effects of covariates on the marginal responses. We use the general odds ratio function to measure the intracluster pairwise assocations, which accommodates responses of any type, is invariant under prospective or retrospective study design, and is unconstrained by the marginal univariate distributions of the responses. Theoretical and simulation results demonstrate the validity of our proposed method. We apply the method to investigate whether HCGI episodes tended to aggregate within households.

email: **lwang87@emory.edu**

31. CONTRIBUTED PAPERS: GWAS

GENE-DISEASE ASSOCIATIONS VIA SPARSE SIMULTANEOUS SIGNAL DETECTION

Sihai Dave Zhao*, University of Illinois at Urbana-Champaign

Tony Cai, University of Pennsylvania

Hongzhe Li, University of Pennsylvania

It is of great interest to identify genes whose expression levels are regulated by disease-associated variants, as these genes may be important in the functional mechanisms underlying the disease. One promising approach is to integrate genome-wide association and genetical genomics studies to test, for a given gene, whether there are SNPs that are simultaneously associated with its expression and with disease. In this paper a method is proposed to detect such simultaneous associations. The method allows the SNP-expression and SNP-disease associations to be calculated in independent datasets and is easy to implement and quick to compute. In addition, it is shown that the proposed method is asymptotically optimal under certain conditions, and a procedure for calculating p-values in finite samples is also provided. In simulations it is shown that the proposed procedure is more powerful than standard enrichment approaches, and in data analysis the procedure is used to identify genes whose regulation may play a role in Crohn's disease.

email: **sdzhao@illinois.edu**

STATISTICAL TESTS FOR THE DETECTION OF SHARED COMMON GENETIC VARIANTS BETWEEN HETEROGENEOUS DISEASES BASED ON GWAS

Julie Kobie*, University of Pennsylvania

Sihai Dave Zhao, University of Illinois at Urbana-Champaign

Yun R. Li, University of Pennsylvania

Hakon Hakonarson, University of Pennsylvania

Hongzhe Li, University of Pennsylvania

Studying complex diseases, such as autoimmune diseases or psychiatric disorders, can lead to the detection of pleiotropic loci with otherwise small effects. Through the detection of pleiotropic loci, the genetic architecture of these

related but clinically-distinct diseases can be better defined, allowing for subsequent improvements in their treatment and prevention efforts. We investigate the genetic relatedness of complex diseases through the detection of shared common genetic variants, utilizing data from readily available genome-wide association studies (GWAS). GWAS have the potential to identify additional SNPs associated with complex diseases with increased sample sizes, but standard meta-analysis approaches are not optimal for the study of these diseases. We present two tests for the detection of shared genetic variants between two diseases, including the global test proposed by Zhao et al. (2014), originally for the analysis of expression quantitative trait loci (eQTL), and a modified global test with an added level of dependency on the direction of the association signals. A procedure for obtaining an analytical p-value for the modified global test is proposed and validated using simulations. Both global tests identify pairs of related but clinically-distinct pediatric autoimmune diseases (pAIDs) that share at least one common genetic variant.

e-mail: **jkobie@mail.med.upenn.edu**

TESTING CLASS-LEVEL GENETIC ASSOCIATIONS USING SINGLE-ELEMENT SUMMARY STATISTICS

Jing Qian*, University of Massachusetts, Amherst

Eric Reed, University of Massachusetts, Amherst

Sara Nunez, University of Massachusetts, Amherst

Rachel Ballentyne, University of Pennsylvania

Liming Qu, University of Pennsylvania

Muredach P. Reilly, University of Pennsylvania

Andrea S. Foulkes, Mount Holyoke College

Characterization of the genetic determinants of complex diseases can be further augmented by incorporating knowledge of underlying structure or classifications of the genome, such as newly developed mappings of protein-coding genes, epigenomic marcs, enhancer elements and non-coding RNAs. In this manuscript, we derive two class-level test statistics and their theoretical distributions, and evaluate them in two publicly-available summary level datasets derived from genome-wide association study meta-analyses -- namely, the CARDIoGRAM and GIANT consortium meta-analysis data. The proposed Genetic Class Association Testing (genCAT) approach is intended to complement post-hoc characterization of class effects (e.g. genes) based on the minimum single element-level (e.g. single SNP level) p-value in the class. Additionally we address high degrees of redundancy in the genotype data. A simulation study is presented to characterize overall performance of this approach.

e-mail: **qian@schoolph.umass.edu**

SET-BASED TESTS FOR GENETIC ASSOCIATION IN LONGITUDINAL STUDIES

Zihuai He*, University of Michigan

Min Zhang, University of Michigan

Seunggeun Lee, University of Michigan

Jennifer A. Smith, University of Michigan

Xiuqing Guo, Harbor-UCLA Medical Center

Walter Palmas, Columbia University

Sharon L.R. Kardia, University of Michigan

Ana V. Diez Roux, University of Michigan

Bhramar Mukherjee, University of Michigan

Genetic association studies with longitudinal markers of chronic diseases provide a valuable opportunity to explore how genetic variants affect traits over time by utilizing the full trajectory of longitudinal outcomes. Since these traits are likely influenced by the joint effect of multiple variants in a gene, a joint analysis of these variants may help to explain additional phenotypic variation. We propose a longitudinal genetic random field model (LGRF), to test the association between a phenotype measured repeatedly during the course of an observational study and a set of genetic variants. Generalized score type tests are developed, which we show are robust to misspefication of within-subject correlation, a feature that is desirable for longitudinal analysis. A joint test incorporating gene-time interaction is further proposed. Computational advancement is made for scalable implementation of the proposed methods

in large-scale genome-wide association studies (GWAS). The methods are evaluated through extensive 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 with two commonly used existing alternatives: (i) single marker tests using longitudinal outcome and (ii) existing gene-based tests using the average value of repeated measurements as the outcome.

e-mail: **zihuai@umich.edu**

GPA: A STATISTICAL APPROACH TO PRIORITIZING GWAS RESULTS BY INTEGRATING PLEIOTROPY AND ANNOTATION

Dongjun Chung*, Medical University of South Carolina

Can Yang, Hong Kong Baptist University

Cong Li, Yale University

Joel Gelernter, Yale University

Hongyu Zhao, Yale University

Results from Genome-Wide Association Studies (GWAS) have shown that complex diseases are often affected by many genetic variants with small or moderate effects. Identifications of these risk variants remain a very challenging problem. Hence, there is a need to develop more powerful statistical methods to leverage available information to improve upon traditional approaches that focus on a single GWAS dataset without incorporating additional data. In this presentation, I will discuss our novel statistical approach, GPA (Genetic analysis incorporating Pleiotropy and Annotation), to increase statistical power to identify risk variants through joint analysis of multiple GWAS data sets and annotation information. Our approach is motivated by the observations that (1) accumulating evidence suggests that different complex diseases share common risk bases, i.e., pleiotropy; and (2) functionally annotated variants have been consistently demonstrated to be enriched among GWAS hits. GPA can integrate multiple GWAS datasets and functional annotations to identify association signals, and it can also perform hypothesis testing to test the presence of pleiotropy and enrichment of functional annotation. I will discuss the power of GPA with its application to real GWAS data with various functional annotations and the simulation studies.

email: **chungd@musc.edu**

OPTIMUM STUDY DESIGN FOR DETECTING IMPRINTING AND MATERNAL EFFECTS BASED ON PARTIAL LIKELIHOOD

Fangyuan Zhang*, The Ohio State University

Abbas Khalili, McGill University

Shili Lin, The Ohio State University

Despite spectacular advances in molecular genomic technologies in the past two decades, resources, especially those for family based genomic studies, are still finite. To maximally utilize limited resources to increase statistical power, an important study-design question is whether to genotype siblings of probands or to recruit more independent families. Numerous studies have attempted to address this issue for detecting imprinting and maternal effects. However, the question is far from settled, mainly due to the fact that results and recommendations in the literature are based on anecdotal evidence from simulation studies rather than based on a rigorous statistical analysis. In this paper, we propose a systematic approach to study various study designs for simultaneous detection of imprinting and maternal effects based on a partial likelihood formulation. We derive the asymptotic properties and obtain close-form formulas for computing the information contents of study designs. Our results show that, for a common disease, recruiting additional siblings is preferred; whereas if a disease is rare, additional families will be a better choice with a fixed amount of resources. Our work thus offers a practical strategy for investigators to select the optimum study design within a case-control family scheme before data collection.

email: **zhang.1243@osu.edu**

ANALYSIS OF GENOMIC DATA VIA LIKELIHOOD RATIO TEST IN COMPOSITE KERNEL MACHINE REGRESSION

Ni Zhao*, Fred Hutchinson Cancer Research Center

Michael C. Wu, Fred Hutchinson Cancer Research Center

Semiparametric kernel machine regression has emerged as a powerful and flexible tool in genomic studies in which genetic variants are grouped into biologically meaningful entities for association

testing. Recent advances have expanded the method to test for the effect of multiple groups of genomic features via a composite kernel that is constructed as a weighted average of multiple kernels. Variance component testing is used to evaluate the significance but requires fixing the weighting parameters or perturbation. In this paper, we focus on the (restricted) likelihood ratio test for kernel machine regression with composite kernels where instead of fixing the weighting parameter, we estimate the weighting parameter by maximizing the likelihood functions through the linear mixed model with multiple variance components. We derive the spectral representation of (R) LRT in linear mixed models with multiple variance components to obtain their finite sample distribution. We conduct extensive simulations to evaluate the power and type I error. Finally, we applied to proposed (R)LRT method to a real study to illustrate our methodology.

email: **nzhao@email.unc.edu**

32. CONTRIBUTED PAPERS: Applications, Simulations and Methods in Causal Inference

ESTIMATING THE FRACTION WHO BENEFIT FROM A TREATMENT, USING RANDOMIZED TRIAL DATA

Emily J. Huang*, Johns Hopkins University

Michael A. Rosenblum, Johns Hopkins University

Most analyses of randomized trials focus on the average treatment effect. The result is that, even when there is strong evidence of a positive average treatment effect for a population, analyses leave unanswered whether treatment benefits are widespread or limited to a select few. This problem affects many disease areas, since it stems from how randomized trials, often the gold standard for evaluating treatments, are designed and analyzed. The goal of this work is to estimate the fraction who benefit from a given experimental treatment, based on randomized trial data, when the primary outcome is continuous or ordinal. Because the fraction who benefit is non-identifiable, we develop a method to estimate sharp lower and upper bounds on it. A novel application of linear programming is used to compute the bounds, which allows fast, flexible, and easy implementation. The method allows incorporation of baseline data, and the user may impose restrictions based on subject matter knowledge. The method is applied to estimate lower and upper bounds on the fraction who benefit from a new surgical intervention for stroke, based on the MISTIE II randomized trial.

email: **emhuang@jhsph.edu**

SENSITIVITY ANALYSES IN THE PRESENCE OF EFFECT MODIFICATION IN OBSERVATIONAL STUDIES

Jesse Y. Hsu*, University of Pennsylvania

Dylan S. Small, University of Pennsylvania

Paul R. Rosenbaum, University of Pennsylvania

In observational studies of treatment effects, subjects are not randomly assigned to treatments, 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 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. In this talk, we will consider (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. A sensitivity analysis for a test of the global null hypothesis of no effect is converted to sensitivity analyses for subgroup analyses using closed testing.

email: **hsu9@mail.med.upenn.edu**

THE CAUSAL EFFECT OF GENE AND PERCENTAGE OF TRUNK FAT INTERACTION ON PHYSICAL ACTIVITY

Taraneh Abarin*, Memorial University

Literature has shown that a high level of physical activity reduces obesity-related traits, such as BMI and percentage of trunk fat. More recent research has also paid attention to the interaction between

physical activity and obesity-associated genes. However, exploring whether or not these associations, in part, reflect reverse causation remains a challenge. More importantly, whether or not carriers of the risk allele modify the causality, is yet to be discovered. Using data from the Complex Diseases in the Newfoundland Population: Environment and Genetics study, we aim to assess whether or not different levels of BMI and percentage of trunk fat in the population of adults in Newfoundland and Labrador, causally influence different levels of physical activity. Using Mandelian Randomization Analysis, we also aim to assess potential genetic effects on these causal relations.

email: **tabarin@mun.ca**

A SIMULATION STUDY OF A MULTIPLY-ROBUST APPROACH FOR CAUSAL INFERENCE WITH BINARY OR CONTINUOUS MISSING COVARIATES

Jia Zhan*, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health

Changyu Shen, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health

Xiaochun Li, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health

Lingling Li, Harvard Medical School and Harvard Pilgrim Health Care Institute

Confounding bias and missing data are two major barriers to valid comparative effectiveness studies using observational data. Each respective problem has been extensively studied. Lately a principled approach to causal inference on data with binary or continuous missing covariates is developed. It is a unified multiply-robust (MR) methodology which simultaneously handles both issues. The 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 with both binary and continuous missing covariates. We use the plug-in approach to deal with the continuous missing covariates, while with binary missing covariates we can explicitly get the expectation. 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@umail.iu.edu**

THE IMPACT OF UNMEASURED CONFOUNDING IN OBSERVATIONAL STUDIES

Zugui Zhang*, Christiana Care Health System

Paul Kolm, Christiana Care Health System

Unmeasured confounding has been a major source of bias in observational studies. In this study, we assessed the impact of unmeasured confounding factors on the 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 National Cardiovascular Data Registry in ASCERT from years 2004 to 2008. 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. Cost-effectiveness is expressed as the incremental cost effectiveness ratio (ICER), the difference in costs of the two groups divided by the difference in effectiveness (events prevented, life years gained or QALYs). The unmeasured confounder could have the cost difference, ranging from decrease 20% to increase 10%. If the prevalence due to the confounder in PCI was 50%, the ICER of CABG vs. PCI would change from $24,000, to $28,000, to $33,000, to $38,000 and $41,000 per QALY gained for the prevalence of the confounder in CABG with 5%, 10%, 20%, 30%, 40%, respectively.

email: **zzhang@christianacare.org**

FLEXIBLE MODELS FOR ESTIMATING OPTIMAL TREATMENT INITIATION TIME FOR SURVIVAL ENDPOINTS: APPLICATION TO TIMING OF cART INITIATION IN HIV/TB CO-INFECTION

Liangyuan Hu*, Brown University

Joseph W. Hogan, Brown University

Timing of combinational antiretroviral therapy (cART) initiation is important in HIV/TB co-infection. Early initiation during TB treatment increases drug toxicity, the risk of inflammatory immune reconstitution, and cost burden; late initiation increases risk for morbidity and mortality associated with HIV/AIDS. Evidence from recent RCTs generally supports early initiation. However, the RCT studies do not give specifics about optimal initiation time or precise recommendations for those with CD4>100. We use data from a large observational cohort to gain more detailed information about treatment effects in practical settings. We formulate a causal structural model that flexibly captures the joint effects of treatment initiation time and treatment duration using smoothing splines, and develop methods for fitting the model to observational data wherein both mortality and cART initiation times are subject to censoring. We fit the model to data from 4903 individuals in a large HIV treatment program in Kenya, and use it to estimate optimal initiation times by CD4 subgroups. Additionally, we derive rules that are consistent with RCTs but have higher resolution in the sense of generating CD4-specific rules that can be used to complement current treatment guidelines.

email: **liangyuan_hu@brown.edu**

DOUBLE ROBUST GOODNESS-OF-FIT TEST OF COARSE STRUCTURAL NESTED MEAN MODELS WITH APPLICATION TO INITIATING HAART IN HIV-POSITIVE PATIENTS

Shu Yang*, Harvard School of Public Health

Judith Lok, Harvard School of Public Health

Coarse Structural Nested Mean Models (SNMMs) provide useful tools to estimate treatment effects from longitudinal observational data in the presence of time-dependent confounders. Coarse SNMMs lead to a large class of estimators. An optimal estimator was derived within the class of coarse SNMMs under the conditions of well-specified models for the treatment effect, for treatment initiation, and for nuisance regression outcomes (Lok et al., 2014). The key assumption lies in a well-specified model for the treatment effect; however there is no existing guidance to specify the treatment effect model. Researchers often use simple models mainly because they do not want to estimate too many parameters. Misspecification of the treatment effect model leads to biased estimators, preventing valid inference. To test whether the treatment effect model matches the data well, we derive a goodness-of-fit (GOF) test procedure based on overidentification restrictions tests (Sargan, 1958; Hansen, 1982). Overidentification restrictions tests are widely used in the economic literature; they however seem to have been previously unnoticed in the statistics and biostatistics literature. We show that our GOF statistic is double-robust in the sense that with a correct treatment effect model, if either the treatment initiation model or the nuisance regression outcome model is correctly specified, the GOF statistic has a chi-squared limiting distribution with degrees of freedom equal to the number of overidentification restrictions. Moreover we show that the testing procedure is consistent. We demonstrate the empirical relevance of our methods using simulation designs based on an actual data set. Our simulation shows that the asymptotic distribution of the GOF statistic derived in this article provides an accurate approximation to the finite sample behavior of the GOF statistic. Our simulations show that the GOF statistic is extremely powerful in detecting misspecification of the treatment effect model. In addition, we apply the GOF test procedure in the study of the role of initiation timing of highly active antiretroviral treatment (HAART) after infection on one-year treatment effect in HIV-positive patients with acute and early infection.

email: **yangshuyounggirl@gmail.com**

33. CONTRIBUTED PAPERS: Adaptive Designs and Dynamic Treatment Regimes

A BAYESIAN OPTIMAL DESIGN IN TWO-ARM, RANDOMIZED PHASE II CLINICAL TRIALS WITH ENDPOINTS FROM EXPONENTIAL FAMILIES

Wei Jiang*, University of Kansas Medical Center

Jo A. Wick, University of Kansas Medical Center

Jianghua He, University of Kansas Medical Center

Jonathan D. Mahnken, University of Kansas Medical Center

Matthew S. Mayo, University of Kansas Medical Center

Frequentist optimal designs for two-arm randomized phase II clinical trials with outcomes from exponential dispersion families was proposed by Jiang et al. (2014), where the total sample size is minimized under multiple constraints on the standard error of the estimated group means. This design was generalized from approaches developed in Mayo et al. (2010) for dichotomous outcomes. Compared to frequentist methods, Bayesian approaches offer a flexible way to incorporate uncertainty in parameters of interest into considerations. In this paper, a Bayesian optimal deign for Phase II clinical trials with endpoints from the exponential families is developed from the two previous frequentist approaches. The proposed optimal design minimizes the total sample size under pre-specified constraints on the expected length of posterior credible intervals for both group means and their difference. Examples of method implementation are provided for different types of endpoints in the exponential families.

email: **wjiang@kumc.edu**

A NOVEL METHOD FOR ESTIMATING OPTIMAL TREE-BASED TREATMENT REGIMES IN RANDOMIZED CLINICAL TRIALS

Lisa L. Doove*, Katholieke Universiteit Leuven

Elise Dusseldorp, Leiden University

Katrijn Van Deun, Tilburg University

Iven Van Mechelen, Katholieke Universiteit Leuven

For many medical problems, multiple treatment alternatives are available. A major challenge in such cases pertains to identifying optimal treatment regimes that specify for each individual client the preferable treatment alternative, with the optimal regime being the one leading to the greatest expected potential outcome for the population under study. Estimating optimal regimes comes down to an unsupervised learning problem, with the goal being to find a set of unknown subgroups of patients each of which is associated with a preferable treatment alternative. Of particular interest for this problem are methods to construct tree-based treatment regimes, in which the subgroups that constitute the basis of the regimes are the leaves of a decision tree. However, the majority of methods for estimating tree-based treatment regimes either do not formally optimize an estimate of expected potential outcome, or use supervised learning techniques. In this paper we propose a novel unsupervised tree-based approach for estimating optimal treatment regimes in RCTs that directly maximizes an estimator of the overall expected outcome for the tree-based regimes under study. The performance of the proposed approach is assessed through simulation studies, and the approach is illustrated using data from an RCT on early-stage breast cancer.

email: **lisa.doove@ppw.kuleuven.be**

LONGITUDINAL BAYESIAN ADAPTIVE DESIGNS FOR THE PROMOTION OF MORE ETHICAL TWO ARMED RANDOMIZED CONTROLLED TRIALS: A NOVEL EVALUATION OF OPTIMALITY

Jo Wick*, University of Kansas Medical Center

Scott M. Berry, Berry Consultants

Byron J. Gajewski, University of Kansas Medical Center

Hung-Wen Yeh, University of Kansas Medical Center

Won Choi, University of Kansas Medical Center

Christina M. Pacheco, University of Kansas Medical Center

Christine Daley, University of Kansas Medical Center

Classical clinical trial designs focus on statistical power but pay little attention to optimizing other operating characteristics. This primary focus on statistical power results in suboptimal trial designs

that lead to a lower percentage of trial participants placed in the better intervention, have fewer trial responders, bigger sample size, and longer duration. For example, the balanced two-armed design has optimal power, but we show this design has suboptimal performance in other operating characteristics. Bayesian adaptive designs (BAD) are known for their flexibility in clinical trial design, allowing for modification to the design based on knowledge gained during the study. However, since BAD are often less powerful than traditional fixed designs, we consider a longitudinal variant, that uses interim results to adapt the randomization of subjects to treatment (BADL) to improve statistical power. A novel approach evaluates the designs based on both traditional operating characteristics and other subject-focused trial features, leaning us to an unbalanced two-armed design as the optimal design.

email: **jwick@kumc.edu**

IDENTIFYING A SET THAT CONTAINS THE BEST DYNAMIC TREATMENT REGIMES

Ashkan Ertefaie*, University of Pennsylvania

Tianshuang Wu, University of Michigan

Inbal Nahum-Shani, University of Michigan

Kevin Lynch, University of Pennsylvania

A dynamic treatment regime (DTR) is a treatment design that seeks to accommodate patient heterogeneity in response to treatment. DTRs can be operationalized by a sequence of decision rules that map patient information to treatment options at specific decision points. The Sequential Multiple Assignment Randomized Trial (SMART) is a trial design that was developed specifically for the purpose of obtaining data that informs the construction of good (i.e., efficacious) decision rules. One of the scientific questions motivating a SMART concerns the comparison of multiple DTRs that are embedded in the design. Typical approaches for identifying the best DTRs involve all possible comparisons between DTRs that are embedded in a SMART, at the cost of greatly reduced power to the extent that the number of embedded DTRs increase. Here, we propose a method that will enable investigators use SMART study data more efficiently to identify the set that contains the most efficacious embedded DTRs. Our method ensures that the true best embedded DTRs are included in this set with at least a given probability. Simulation results are presented to evaluate the proposed method and the Extending Treatment Effectiveness of Naltrexone SMART study data are analyzed to illustrate its application.

email: **ertefaie@wharton.upenn.edu**

OPTIMAL DYNAMIC TREATMENT REGIMES FOR TREATMENT INITIATION WITH CONTINUOUS RANDOM DECISION POINTS

Yebin Tao*, University of Michigan

Lu Wang, University of Michigan

Haoda Fu, Eli Lilly and Company

Identifying the optimal dynamic treatment regimes (DTRs) allows patients to receive the best treatment prescription given their own evolving disease status and medical history. We consider estimating the optimal DTRs for treatment initiation using observational data, where key biomarkers of disease severity are monitored continuously during follow-up and a decision of whether or not to initiate a specific treatment is made each time the biomarkers are measured. The goal is to find out the optimal DTR to initiate the treatment given the biomarker history. Instead of considering multiple fixed decision stages as in most DTR literature, our study undertakes the task of dealing with continuous random decision points for treatment modification given patients' up-to-date clinical records. Under each DTR, we employ a flexible survival model with splines for time-varying covariates to estimate the probability of adherence to that regime for all patients, given their own covariate history. We then use the estimated probability to construct inverse probability weighted estimators for the counterfactual mean utility, prespecified criteria for assessing each DTR. We conduct simulations to demonstrate the performance of our method and further illustrate the application procedure with the example of type 2 diabetes patients enrolled to initiate insulin therapy.

email: **yebintao@umich.edu**

STATISTICAL INFERENCE FOR THE MEAN OUTCOME UNDER A POSSIBLY NON-UNIQUE OPTIMAL TREATMENT STRATEGY

Alexander R. Luedtke*, University of California, Berkeley

Mark J. van der Laan, University of California, Berkeley

We consider challenges that arise in the estimation of the mean outcome under an optimal individualized treatment strategy defined as the treatment rule that maximizes the mean outcome, where the candidate treatment rules are restricted to depend on baseline covariates. Previous works have established regular root-n rate inference for this quantity in a large semiparametric model provided that the treatment has some effect, either beneficial or detrimental, in each strata of the measured covariates. Here we prove a necessary and sufficient condition for the pathwise differentiability of the mean outcome under an optimal rule, a key condition needed for such regular root-n rate inference, that is slightly more general than the previous condition implied in the literature. We then describe an approach to get confidence intervals for the mean outcome under the optimal individualized treatment strategy even when regular inference is not possible. This procedure requires that one be able to (at least asymptotically) bound the mean-squared error between an estimate of the strata-specific treatment effects and the true underlying strata-specific treatment effects. We show that bounding such a quantity is straightforward in a parametric model, and suggest an extension to the nonparametric case.

email: **aluedtke@berkeley.edu**

SEQUENTIAL ADVANTAGE SELECTION FOR OPTIMAL TREATMENT REGIME

Ailin Fan*, North Carolina State University

Wenbin Lu, North Carolina State University

Rui Song, North Carolina State University

Variable selection plays an important role in deriving practical and reliable optimal treatment regimes for personalized medicine, especially when there are a large number of predictors, and is getting more attention. Most existing variable selection techniques focus on selecting variables that are important for prediction, therefore some variables that are poor in prediction but are critical for decision-making may be ignored. A qualitative interaction of a variable with treatment arises when treatment effect changes direction as the value of this variable varies. Variables that have qualitative interactions with treatment are of clinical importance for decision-making. Gunter et al. proposed S-score to characterize the magnitude of qualitative interaction of individual variable with treatment. In this article, we developed a sequential advantage selection method based on modified S-score. Our method selects qualitatively interacted variables sequentially and allows multiple decision points. We also propose a BIC-type criterion based on sequential advantage to select the best candidate subset of variables for decision-making. The empirical performance of the proposed method is evaluated by simulation and an application to depression data from a clinical trial.

email: **afan@ncsu.edu**

34. CONTRIBUTED PAPERS: Survival Analysis and Cancer Applications

REGRESSION ANALYSIS OF INFORMATIVE CURRENT STATUS DATA UNDER CURE RATE MODEL

Yeqian Liu*, University of Missouri, Columbia

Tao Hu, Capital Normal University, China

Jianguo Sun, University of Missouri, Columbia

Current status data arise when a single follow-up inspection is made for each individual and the occurrence of events is only detected at inspection times. Motivated by medical studies in which patients could be cured and no longer suseptible to the disease. We consider the current status data under the cure rate model and assume a generalized linear model with log link function for the cure probability. The Cox proportional hazards models are used to model both the failure times and censoring times. To avoid the inference bias resulted from ignoring the informative censoring. We

propose a log-normal frailty to characterize the correlation between the censoring time and the failure time. An EM algorithm combining a sieve method by using Bernstein polynomials is developed for parameter estimation. Simulation studies are performed to evaluate the proposed estimates and suggest that the approach works well in finite sample situations. An illustrative example is provided.

email: **yldg5@mail.missouri.edu**

THE HISTORICAL COX MODEL

Jonathan E. Gellar*, Johns Hopkins Bloomberg School of Public Health

Fabian Scheipl, LMU Munich

Mei-Cheng Wang, 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

In this paper, we extend the Cox proportional hazards model to account for densely sampled time-varying covariates as historical functional terms. This approach allows the hazard function at any time t to depend not only on the current value of the time-varying covariate, but also on all previous values. The fundamental idea is to assume a bivariate coefficient function ?(s, t) that estimates a weight function that is applied to the full or partial covariate history up to t, and is allowed to change with t. Estimation is performed by maximizing the penalized partial likelihood, using a likelihood-based information criterion to optimize the smoothing parameter. Methods are applied to a study of in-hospital mortality among patients with acute respiratory distress syndrome in the intensive care unit.

email: **jgellar@jhsph.edu**

BAYESIAN ANALYSIS OF SURVIVAL DATA UNDER GENERALIZED EXTREME VALUE DISTRIBUTION WITH APPLICATION IN CURE RATE MODEL

Dooti Roy*, University of Connecticut

Vivekananda Roy, Iowa State University

Dipak Dey, University of Connecticut

This paper introduces both maxima and minima generalized extreme value (GEV) distribution to analyze right-censored survival data. We also use GEV distributions to construct flexible models for populations with a surviving fraction. Our proposed GEV model leads to extremely flexible hazard functions. We show that our Bayesian model has several nice properties. For example, we prove that even when improper priors are used, the resulting posterior distribution could still be proper under some weak conditions. We further provide theoretical and numerical results showing that our GEV models offer a richer class of models than the widely used Weibull models. Finally, a glioblastoma multiforme cancer data with a cure rate is analyzed to illustrate the proposed GEV model.

email: **dooti.roy@uconn.edu**

JOINT SEMIPARAMETRIC TIME-TO-EVENT MODELING OF CANCER ONSET AND DIAGNOSIS WHEN ONSET IS UNOBSERVED

John D. Rice*, University of Michigan

Alex Tsodikov, University of Michigan

In cancer research, interest frequently centers on factors influencing a latent event that must precede a terminal event. In practice it is often impossible to observe the latent event precisely, making inference about this process difficult. To address this problem, we propose a joint model for the unobserved time to the latent and terminal events, with the two events linked by the baseline hazard. Covariates enter the model parametrically as linear combinations that multiply, respectively, the hazard for the latent event and the hazard for the terminal event conditional on the latent one. The baseline hazard is estimated nonparametrically using the EM algorithm, which allows for closed-form Breslow-type estimators at each iteration, drastically reducing computational time compared with maximizing the marginal likelihood directly. The parametric part of the model is estimated by maximizing the profile likelihood. We derive asymptotic properties for the model, while simulation studies are presented to illustrate the finite-sample properties of the method. Its use in practice is demonstrated in the analysis of a prostate cancer data set.

email: **jdrice@umich.edu**

A MULTIPLE IMPUTATION APPROACH FOR SEMIPARAMETRIC CURE MODEL WITH INTERVAL CENSORED DATA

Jie Zhou*, University of South Carolina, Columbia

Jiajia Zhang, University of South Carolina, Columbia

Alexander C. McLain, University of South Carolina, Columbia

Bo Cai, University of South Carolina, Columbia

Interval censored data, where the exact event time is only known to lie in an observed time interval, are commonly encountered in practice. Such data analysis may be conducted under the setting where a fraction of patients can be considered as fully recovered and they will never experience the event of interest. We propose a semiparametric estimation method for the proportional hazards mixture cure model, which is easy to implement and is computationally efficient. A multiple imputation approach based on the asymptotic normal data augmentation is used to obtain parameter and variance estimates for both the cure probability and the survival probability of uncured patients. A simulation study is performed to evaluate the proposed method and the results are compared with a fully parametric approach. The proposed method is applied to 2000-2010 Greater Georgia breast cancer dataset from the Surveillance, Epidemiology, and End Results Program.

email: **zhou57@email.sc.edu**

A FLEXIBLE PARAMETRIC CURE RATE MODEL WITH KNOWN CURE TIME

Paul W. Bernhardt*, Villanova University

Models for survival data usually assume that all individuals will eventually experience the event of interest. However, in many applications, the event will never occur for a subset of individuals. Cure rate models have long been used for handling this type of data by modeling the probability of being "cured" as well as the probability of survival among those who are not cured. We propose an extension of standard parametric mixture cure rate models that allows for the incorporation of both a fixed, known cure time and different censoring distributions for the cured and uncured subgroups. We show through simulations that the proposed model performs well in a variety of scenarios.

email: **paul.bernhardt@villanova.edu**

CHANGE-POINT PROPORTIONAL HAZARDS MODEL FOR CLUSTERED EVENT DATA

Yu Deng*, University of North Carolina, Chapel Hill

Jianwen Cai, University of North Carolina, Chapel Hill

Donglin Zeng, University of North Carolina, Chapel Hill

Jinying Zhao, Tulane University

In the analysis of clustered time-to-event data, some continuous variable may possess a "change point", which violates the assumption of linear effects on the disease incidence in the standard Cox model. In this work, we propose a change-point proportional hazards model for clustered event data. The model incorporates the unknown threshold of the threshold variable as a change point in the regression. Pseudo-partial likelihood functions are maximized for estimating both regression coefficients and the change point in the model. Furthermore, we use the supremum test based on robust score statistics to check the existence of a change point. The m out of n bootstrap method is applied to make inference for the estimator of the change point, where m is determined by the extension of Bickel and Sakov (2008) method. We establish the consistency and asymptotic distributions of the proposed estimators. The small-sample performance of the proposed method is demonstrated via simulation studies. Finally, the Strong Heart Study dataset is analyzed to illustrate the method.

email: **yudeng@live.unc.edu**

35. INVITED AND CONTRIBUTED ORAL POSTERS: Methods and Applications in High Dimensional Data and Machine Learning

35a. INVITED POSTER: MACHINE LEARNING METHODS FOR CONSTRUCTING REAL-TIME TREATMENT POLICIES IN MOBILE HEALTH

Susan Murphy*, University of Michigan

Yanzhen Deng*, University of Michigan

Mobile devices are being increasingly used by health researchers to, in real-time, both collect symptoms and other information as well as provide interventions. These interventions are often provided via treatment policies. The policies specify how patient information should be used to determine when, where and which intervention to provide. Here we present generalizations of "Actor-Critic" learning methods from the field of Reinforcement Learning for use, with existing data sets, in constructing treatment policies.

email: **samurphy@umich.edu**

35b. INVITED POSTER: PREDICTING STROKES USING RELATIONAL RANDOM FORESTS

Zach Shahn, Columbia University

Patrick Ryan, Columbia University

David Madigan*, Columbia University

With increasingly widespread use of Electronic Health Records (EHRs), predicting health outcomes from high dimensional, longitudinal health histories is of central importance to healthcare. The medical literature has formalized such prediction problems in a few instances, and the resulting "risk calculators" attract widespread use. We have adapted a predictive modeling method originally developed in the context of speech recognition to the context of large-scale EHR data. "Relational Random Forests" (RRF) greedily construct informative labeled graphs representing temporal relations between multiple health events at the nodes of randomized decision trees. We have applied RRFs to the problem of predicting strokes in patients newly diagnosed with atrial fibrillation. Our approach compare favorably with the widely used "CHADS2" risk calculator.

email: **david.madigan@columbia.edu**

35c. NETWORK-CONSTRAINED GROUP LASSO FOR HIGH DIMENSIONAL MULTINOMIAL CLASSIFICATION WITH APPLICATION TO CANCER SUBTYPE PREDICTION

Xinyu Tian*, Stony Brook University

Jun Chen, Mayo Clinic

Xuefeng Wang, Stony Brook University

Classic multinomial logit model, commonly used in multiclass regression problem, is restricted to few predictors and no regard to the relationship among variables. Its usage is limited for genomic data, where the number of genomic features far exceeds the sample size. Also, genomic features such as gene expressions are usually related by an underlying biological network. Making use of the network information is crucial to improve classification performance as well as the biological interpretability. We proposed a penalized multinomial logit model that is capable to adjust for both the high-dimensionality of predictors and the underlying network information. In fact, group LASSO was involved to induce model sparsity and a network constraint was used to induce the smoothness of the coefficients subject to the underlying network structure. To deal with the non-convexity of the objective function in parameter estimation, we developed a proximal-gradient-based algorithm for efficient computation. The proposed models were compared to models with no prior structure information in both simulations and a problem of cancer subtype prediction with real data, and it outperformed the traditional ones in both cases.

email: **sarahtxy@gmail.com**

35d. TWO SAMPLE MEAN TEST IN HIGH DIMENSIONAL COMPOSITIONAL DATA

Yuanpei Cao*, University of Pennsylvania

Wei Lin, Peking University

Hongzhe Li, University of Pennsylvania

Compositional data arise naturally in many scientific applications; for example, in microbiome studies, only the composition of the bacterial taxa is observed. Recently studies also found that the differences in the composition of the microbiome are associated with disease

or treatment outcomes. Thus, detecting the differences of the composition is a potentially important issue in microbiome studies. However, the performance of the canonical generalized likelihood ratio test (Aitchison, 2003) on the additive log-ratio transformation (alr) of the composition is unsatisfactory under the high dimension setting. In this article, we introduce a global test based on the centered log-ratio transformation (clr) to detect the differences of the compositions. Under the assumption that the basis covariance matrix is sparse, we show that the limiting null distribution of the test statistic and the power of the test based on clr are the same as the test on the log transformation of the basis. Simulation studies demonstrate that such tests based on clr outperform some naive tests that ignore the unique features of the compositional data. We apply the proposed test to an analysis of microbiome data that compare normal and Crohn's disease gut microbiomes.

email: **yuanpeic@sas.upenn.edu**

35e. CLASSIFICATIONS BASED ON ACTIVE SET SELECTIONS

Wen Zhou*, Colorado State University

Stephen Vardeman, Iowa State University

Huaiqing Wu, Iowa State University

Max Morris, Iowa State University

Dataset shift is the phenomenon in predictive analytics where distributions of training and predicting (or test) data are different. This is encountered in developing classification methods. It is drawing growing attention as many practical applications of classification must cope with some degree of shift, and performances of theoretically well-behaved methods can suffer substantial degradation when it is present. We consider the covariate shift problem, a particular type of dataset shift where distributions of feature vectors are possibly different between training and predicting (test) sets. Inspired by kernel density estimations, we propose a classification method that involves the weighted bootstrap and ensemble learning. This procedure trains classifiers using subsets of the training data that are in some sense like the predicting (test) cases, thereby dealing with the covariate shift problems. The resulting method is called {\bf A}ctive {\bf S}et {\bf S} election {\bf C}lassification (ASSC). The basic procedure is flexible and can be used with existing methods of classification, such as support vector machines (SVM)s, linear discriminant analysis (LDA), and classification trees to improve their prediction accuracy. ASSC performs well on both simulated and real data sets. We preface application of ASSC with a preliminary screening step to deal with situations where the number of features is larger than the training set size.

email: **riczw@stat.colostate.edu**

35f. APPLICATION OF A GRAPH THEORY ALGORITHM IN SOFT CLUSTERING

Wenzhu Mowrey*, Albert Einstein College of Medicine

George C. Tseng, University of Pittsburgh

Lisa A. Weissfeld, Statistics Collaborative, Inc.

Clustering methods usually assign a hard cluster membership to indicate whether or not an observation belongs to a cluster. In situations where the underlying subgroups overlap with each other or there are outliers or noisy observations that may influence clustering results, soft clustering methods may be desirable since these methods allow for the assignment of a cluster membership probability to indicate the likelihood that an observation belongs to a cluster. These methods often involve resampling the dataset, where cluster memberships are summarized by a comembership matrix for each resampling run. The consensus matrix is then computed as the average of the comemberhsip matrices from all resampling runs. In this work we propose using the Bron-Kerbosch algorithm from graph theory to obtain clusters from the consensus matrix. This algorithm is ideal since obtaining clusters from the consensus matrix can be viewed as equivalent to the maximum clique problem in graph theory where the goal is to find the largest complete subgraph within a graph and by "complete" it means that any two nodes of the graph are connected.

email: **wenzhu.mowrey@einstein.yu.edu**

35g. TESTING FOR THE PRESENCE OF CLUSTERING

Erika S. Helgeson*, University of North Carolina, Chapel Hill

Eric Bair, University of North Carolina, Chapel Hill

Cluster analysis is an unsupervised learning strategy that can be employed to identify groups of observations in data sets of unknown structure. This strategy

is particularly useful for analyzing high-dimensional data such as microarray gene expression data. Many methods are available which can identify the number of clusters present in the data and/or group the observations into their appropriate clusters based on feature characteristics, but there are only a few methods that can determine whether clusters are actually present in the data. We propose a novel method for testing the null hypothesis that no clusters are present in a given data set by comparing the number of features associated with the clusters to the expected number of features under an appropriate null distribution. We apply this method to a variety of simulated data sets and compare the results of our method to those of previously published methods. Overall, our method has comparable predictive accuracy and much shorter computing time, indicating that our method is a useful tool for determining if clusters are present in a data set.

email: **helgeson@live.unc.edu**

35h. VARIABLE SELECTION AND SUFFICIENT DIMENSION REDUCTION FOR HIGH DIMENSIONAL DATA

Yeonhee Park*, University of Florida

Zhihua Su, University of Florida

Contemporary data sets often involve large number of predictor variables, such a high-dimensional data brings the challenge for traditional data analysis methods. In this paper, we propose a two stages procedure: (1) to identify the relevant predictor variables and discard irrelevant variables with the help of screening method, and (2) to estimate the central subspace containing all useful information for relevant predictors. The two stages procedure is also applicable for multivariate response. This procedure can also handle the situation when the number of relevant predictors is still larger than the number of observations. Moreover, theoretical results are established for this procedure. We applied our methodology to simulated data and real data sets including leukemia data and prostate cancer data. They showed that our method works pretty well, in particular our method does not lose any important information in application to prostate cancer data.

email: **yeonhee@stat.ufl.edu**

35i. VARIABLE SELECTION FOR TREATMENT DECISIONS WITH SCALAR AND FUNCTIONAL COVARIATES

Adam Ciarleglio*, New York University School of Medicine

Eva Petkova, New York University School of Medicine and Nathan S. Kline Institute for Psychiatric Research

R. Todd Ogden, Columbia University

Thaddeus Tarpey, Wright State University

The amount and complexity of patient-level data being collected in randomized controlled trials offers both opportunities and challenges for developing personalized rules for assigning treatment for a given condition. For example, trials examining treatments for major depressive disorder (MDD) are not only collecting a large number of typical baseline data such as age, gender, or scores on various tests, but also data that measure the structure and function of the brain via magnetic resonance imaging (MRI), functional MRI (fMRI), or electroencephalography (EEG). These latter types of data have an inherent structure and may be considered as functional data. Unfortunately, there is often little clinical guidance about which, if any, of these many baseline covariates are prescriptive of treatment. We propose an approach that both selects important prescriptive covariates and estimates a treatment decision rule when there are many candidate covariates consisting of both scalar and functional data. We describe our method and how to implement it using existing software. Performance is evaluated on simulated data in a variety of settings and we apply our method to data arising from the study of patients suffering from MDD from which baseline scalar and functional data are available.

email: **Adam.Ciarleglio@nyumc.org**

35j. MOPM: MULTI-OPERATOR PREDICTION MODEL BASED ON HIGH-DIMENSIONAL FEATURES

Hojin Yang*, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

Joseph G. Ibrahim, University of North Carolina, Chapel Hill

We consider the problem of integrating and identifying important genomic, imaging, and biological markers to accurately predict low-dimensional outcome variables, such as disease status or behavioral scores. Such prediction problem can have a great impact in

public health from disease prevention, to detection, to treatment selection. The aim of this paper is to develop a multi-operator prediction modeling (MOPM) framework to perform a supervised dimension reduction and then build an accurate prediction model. We formulate the problem of supervised dimension reduction as a variable selection problem and propose an independent screening method to select a set of informative features, which may have a complex nonlinear relationship with outcome variables. Moreover, we develop a novel local projection method to use multiple linear operators to project all informative predictors into multiple local subspaces in order to capture reliable and informative covariate information. Theoretically, we systematically investigate some theoretical properties of MOPM. Our simulation results and real data analysis show that MOPM outperforms many state-of-theart methods in terms of prediction accuracy.

email: **hojiny0504@gmail.com**

35k. STRUCTURED SPARSE CCA FOR HIGH DIMENSIONAL DATA INTEGRATION

Sandra Safo*, Emory University

Qi Long, Emory University

Canonical Correlation Analysis (CCA) is a classical multivariate analysis tool that aims at studying association between two sets of variables by finding linear combinations of all available variables with maximum correlation. CCA has limitations in the high dimensional framework as it is usually of interest to select only a fraction of the variables in addition to studying association. Several methods for sparse CCA have been proposed in the literature. Although these methods have proven useful in various applications, their main drawbacks are failure to account for prior biological knowledge, and assumption of independence of the underlying covariance structures, the latter of which can be overly restrictive. In this paper, we propose a novel structured sparse CCA method that overcomes these limitations by incorporating biological information and making no assumptions on the underlying covariance structures. We compare our method to existing sparse CCA approaches via simulation studies and real data analysis using gene expression and metabolomics data from a cardiovascular disease study.

email: **seaddo@uga.edu**

35l. SPARC: OPTIMAL ESTIMATION AND ASYMPTOTIC INFERENCE UNDER SEMIPARAMETRIC SPARSITY

Yang Ning*, Princeton University

Han Liu, Princeton University

We propose a new inferential framework called semiparametric regression via chromatography (SPARC) to handle the challenges of complex data analysis featured by high dimensionality and heterogeneity. Under a semiparametric sparsity assumption, we develop a regularized statistical chromatographic estimation method, and establish the nearly optimal parameter estimation error bounds under Lq norms. Furthermore, we propose a unified framework for statistical inference including score, Wald and likelihood ratio tests under the SPARC model. The asymptotic properties of these tests are established in high dimensions. Our theory does not require the irrepresentable condition or any assumption on the minimal signal strength. We further examine the impact of model misspecification on parameter estimation and hypothesis tests. We also show that the proposed methods are applicable to the challenging settings with missing values or selection bias. Finally, we establish the corresponding theory for the multitask learning problem to handle the data with heterogeneity. In particular, under the semiparametric group sparsity assumption, we demonstrate that the resulting estimator can achieve an improvement in the estimation errors as compared to the L1-regularized estimator. These theoretical results are illustrated through simulation studies, and a real data example.

email: **y4ning@uwaterloo.ca**

35m. LOCAL-AGGREGATE MODELING FOR BIG-DATA VIA DISTRIBUTED OPTIMIZATION: APPLICATIONS TO NEUROIMAGING

Yue Hu*, Rice University

Genevera I. Allen, Rice University, Baylor College of Medicine and Texas Children's Hospital

Technological advances have led to a proliferation of structured big-data that is often collected and stored in a distributed manner. We are specifically motivated to build predictive models for multi-subject neuroimaging data based on each subject's brain imaging scans.

This is an ultra-high-dimensional problem that consists of a matrix of covariates (brain locations by time points) for each subject; few methods currently exist to fit supervised models directly to this tensor data. We propose a novel modeling and algorithmic strategy to apply generalized linear models (GLMs) to this massive tensor data in which one set of variables is associated with locations. Our method begins by fitting GLMs to each location separately, and then builds an ensemble by blending information across locations through regularization with what we term an aggregating penalty. Our so called, Local-Aggregate Model, can be fit in a completely distributed manner over the locations using an Alternating Direction Method of Multipliers (ADMM) strategy, and thus greatly reduces the computational burden. Furthermore, we propose to select the appropriate model through a novel sequence of faster algorithmic solutions that is similar to regularization paths. We will demonstrate both the computational and predictive modeling advantages of our methods via simulations and an EEG classification problem.

email: **yue.hu@rice.edu**

35n. RESIDUAL WEIGHTED LEARNING FOR ESTIMATING INDIVIDUALIZED TREATMENT RULES

Xin Zhou*, University of North Carolina, Chapel Hill

Michael R. Kosorok, University of North Carolina, Chapel Hill

Personalized medicine has received increasing attention among statisticians, computer scientists, and clinical practitioners. A major component of personalized medicine is to estimate individualized treatment rules. Recently, Zhao et al. (2012) proposed the outcome weighted learning (OWL) to construct individualized treatment rules that directly optimize the clinical outcome. However, the individualized treatment rule estimated by OWL would keep, if possible, the treatment assignments that the subjects actually received. This behavior of OWL weakens the finite sample performance. In this article, we propose a new method, called Residual Weighted Learning (RWL), to alleviate this problem, and to improve the finite sample performance. Not like OWL which weights the misclassification errors by the clinical outcomes, the RWL weights the errors by the residuals of outcome from a regression fit on clinical covariates other than the treatment assignment. We utilize the truncated hinge loss function in the RWL, and provide a difference of convex (d.c.) algorithm to solve the non-convex optimization problem. We show that the resulting estimator of the treatment rule is universally consistent. We further obtain a finite sample bound for the difference between the expected outcome using the estimated individualized treatment rule and that of the optimal treatment rule. The performance of our proposed RWL method is illustrated in simulation studies and an analysis of chronic depression data.

email: **xinzhou@live.unc.edu**

35o. INTEGRATIVE MULTI-OMICS CLUSTERING FOR DISEASE SUBTYPE DISCOVERY BY SPARSE OVERLAPPING GROUP LASSO AND TIGHT CLUSTERING

SungHwan Kim*, University of Pittsburgh

YongSeok Park, University of Pittsburgh

George Tseng, University of Pittsburgh

With the rapid advances in technologies of microarray and massively parallel sequencing, data of multiple omics sources from a large sample cohort are now frequently seen in many consortium studies. Effective multi-omics data integration has brought new statistical challenges. One of important biological objective of the data analysis is clustering patients in order to identify meaningful disease subtypes, which is the fundamental basis for tailored treatment and personalized medicine. Several methods have been proposed in the literature to accommodate this purpose, including the popular iCluster in many cancer applications. However, all of those methods fail to properly incorporate the information from inter-omics regulation flow and do not allow outlier samples scattering away from the tight clusters. In this paper, we propose a group structured iCluster method to incorporate a sparse overlapping group lasso technique and a tight clustering concept via regularization to circumvent the aforementioned pitfalls. We show by two real examples and simulated data that our proposed methods improve the original iCluster in clustering accuracy, biological interpretation, and are able to generate coherent tight clusters.

email: **suk73@pitt.edu**

35p. IDENTIFYING PREDICTIVE MARKERS FOR PERSONALIZED TREATMENT SELECTION

Yuanyuan Shen*, Harvard University

Tianxi Cai, Harvard University

Many illness show heterogeneous response to treatment, which motivates researchers to advocate the individualization of treatment to each patient. Many Individualized Treatment Rules (ITR) have been developed but not many approaches on identifying markers that can guide treatment selection have been studied. Traditional Wald test of interaction between treatment and markers has two major limitations: the validity of testing for interaction in terms of identifying important treatment selection markers is scale-dependent; and it doesn't consider potential non-linearity among the predictors. We propose a scale-independent score statistic to test and detect important baseline predictors that can guide treatment selection. Kernel machine framework is also incorporated to handle the non-linearity among predictors. Simulation studies show that our proposed kernel machine based score test is more powerful than the Wald test when there is non-linear effect among the predictors as well as when the outcome is binary and the link function is non-linear. Furthermore, when there is high-correlation among predictors and when the number of predictors is big, our method over-performs Wald test due to the limitations of Wald test under such scenarios.

email: **yshen@g.harvard.edu**

36. Recent Research in Adaptive Randomized Trials with the Goal of Addressing Challenges in Regulatory Science

ADAPTIVE ENRICHMENTWITH SUBPOPULATION SELECTIONAT INTERIM

Sue-Jane Wang*, U.S. Food and Drug Administration

Hsien-Ming James Hung, U.S. Food and Drug Administration

There is growing interest in pursuing adaptive enrichment for drug development because of its potential to achieve the goal of personalized medicine. There are many versions of adaptive enrichment proposed across many disease indications. Some are exploratory adaptive enrichment and others aim at confirmatory adaptive enrichment. In this paper presentation, we give a brief overview on adaptive enrichment and the methodologies that are growing in statistical literature. A case example that was planned to adapt two design elements, i.e., population adaptation and statistical information adaptation, will be given. We articulate the challenges in the implementation of a confirmatory adaptive enrichment trial. We also assess the consistency of treatment effect before and after adaptation. We also discuss and articulate design considerations for adaptive enrichment among a dual-composite null hypothesis, a flexible dual-independent null hypothesis and a rigorous dual-independent null hypothesis.

email: **suejane.wang@fda.hhs.gov**

POST-TRIAL SIMULATION OF TYPE I ERROR FOR DEMONSTRATION OF CONTROL OF TYPE I ERROR

Scott M. Berry*, Berry Consultants

The ability to demonstrate type I error of innovative adaptive trials through simulation allows a whole new world of innovative trial design. Simulation to demonstrate control can never explore the entire "null space"— but extensive pre-simulation can be done. Frequently the type I error can depend upon ancillary aspects of a trial — the rate under control, the shape of the sitribution, the accrual rate, and even the drop-out rate. We discuss the ability to do prospectively defined post-trial simulations to additionally demonstrate control of type I error with a bootstrapping type approach. We present several examples and discussions of the regulatory impact.

email: **scott@berryconsultants.com**

BAYESIAN COMMENSURATE PRIOR APPROACHES FOR PEDIATRIC AND RARE DISEASE CLINICAL TRIALS

Bradley P. Carlin*, University of Minnesota

Cynthia Basu, University of Minnesota

Brian Hobbs, University of Texas MD Anderson Cancer Center

Rare diseases are difficult to study, since the numbers of persons who can be enrolled in a traditional clinical trial is typically insufficient to demonstrate a statistically significant treatment effect.

Pediatric disease researchers face similar challenges. Here, drugs successfully tested on adults are sometimes available, but we still lack information on dosing, safety, and efficacy of these drugs in children. Full or partial extrapolation of existing adult data to the pediatric case is sometimes justified, but current methods are often ad hoc and depend crucially on knowing the appropriate amount of information to borrow from the adult data. This talk considers a collection of novel Bayesian statistical methods and software tools for more efficient and effective orphan and pediatric drug trials. Bayesian methods offer a formal statistical framework for incorporating all sources of knowledge (structural constraints, expert opinion, and both historical and experimental data), thus offering the possibility of substantially reduced trial sizes, thanks to their more efficient use of information. This in turn typically leads to increases in statistical power and reductions in cost and ethical hazard, the latter since fewer patients need be exposed to inferior treatments. Our methods use commensurate priors where possible to combine relevant auxiliary information, and we check our procedures to ensure adequate Type I error performance. We illustrate in the context of and using real data from ongoing clinical trials at the University of Minnesota and the University of Texas M.D. Anderson Cancer Center, and in disease areas such as adrenoleukodystrophy (ALD), Gaucher's Disease, epilepsy, Parkinson's Disease, and certain rare cancers or cancer subtypes.

email: **brad@biostat.umn.edu**

IDENTIFYING SUBPOPULATIONS WITH THE LARGEST TREATMENT EFFECT

Iván Díaz*, Johns Hopkins Bloomberg School of Public Health

Michael Rosenblum, Johns Hopkins Bloomberg School of Public Health

In the presence of effect modifiers, overall population effects often mask the presence of subpopulations with large and small treatment effects. Knowledge of such subpopulations is of high importance in personalized medicine as it allows physicians to assign the most beneficial treatment according to the patient's characteristics, potentially reducing costs, increasing efficacy, and improving the system overall. In this paper we present a method for classifying individuals according to their treatment effect, conditional on baseline variables. Existing methods rely on classification criteria that optimize the average treatment effect, but fail to account for the uncertainty in the estimates. We propose a classification criterion that optimizes the signal to noise ratio, ensuring optimal power of a hypothesis test of no effect. Our motivating application is the phase II MISTIE trial on minimally invasive surgery after Intracerebral Hemorrhage (ICH). We present the results of the analysis of the MISTIE II trial as well as simulations showing the properties of the method in finite samples.

email: **idiaz@jhu.edu**

37. Statistical Innovations in Functional Genomics and Population Health

QUALITY PRESERVING DATABASES: STATISTICALLY SOUND AND EFFICIENT USE OF PUBLIC DATABASES FOR AN INFINITE SEQUENCE OF TESTS

Saharon Rosset*, Tel Aviv University

Ehud Aharoni, IBM Research

Hani Neuvirth, IBM Research

Large databases whose usage is open to the scientific community to facilitate research are becoming commonplace, especially in Biology and Genetics. The emerging scenario in which a community of researchers sequentially conduct multiple statistical tests on one shared database gives rise to major multiple hypothesis testing issues. We suggest a scheme we term Quality Preserving Database (QPD) for controlling false discovery without any power loss by adding new samples for each use of the database and charging the user with the expenses. The crux of the scheme is a carefully crafted pricing system that fairly prices different user requests based on their demands while controlling false discovery. The statistical problem encountered is one of defining appropriate measures of false discovery that can be controlled sequentially, and designing methodologies that can control them in the context of QPD. We describe a simple QPD implementation based on controlling the family-wise error rate using a method called alpha-spending, and a more involved implementation based on controlling a measure called mFDR,

using an approach we term generalized alpha investing. We derive the favorable statistical properties of generalized alpha investing variants in general, and in the context of QPD in particular. The variant we implement can guarantee infinite use of a public database while preserving power, with very low costs, or even no costs under some realistic assumptions. We demonstrate this idea in simulations and describe its potential application to several real life setups.

email: **saharon@post.tau.ac.il**

FUSED LASSO ADDITIVE MODEL

Ashley Petersen*, University of Washington

Daniela Witten, University of Washington

Noah Simon, University of Washington

We consider the problem of predicting an outcome variable using covariates that are measured on independent observations, in the setting in which flexible and interpretable fits are desirable. We propose the fused lasso additive model (FLAM), in which each additive function is estimated to be piecewise constant with a small number of adaptively-chosen knots. FLAM is the solution to a convex optimization problem, for which a simple algorithm with guaranteed convergence to the global optimum is provided. FLAM is shown to be consistent in high dimensions, and an unbiased estimator of its degrees of freedom is proposed. We evaluate the performance of FLAM in a simulation study and on two data sets.

email: **dwitten@uw.edu**

IMPUTING TRANSCRIPTOME IN INACCESSIBLE TISSUES IN AND BEYOND THE GTEx PROJECT VIA RIMEE

Jiebiao Wang, University of Chicago

Dan Nicolae, University of Chicago

Nancy Cox, University of Chicago

Lin S. Chen*, University of Chicago

In order to synthesize new knowledge about the organization of gene expression across human tissues, the Genotype-Tissue Expression (GTEx) project collected the transcriptome data in a wide variety of tissues from organ donors. Some human tissues are hard to access and transcriptome information in those tissues have only limited availability. We show that those transcriptome data can be imputed by harnessing rich information in the GTEx data, and furthermore it is feasible to use GTEx data as reference and impute inaccessible tissues for future studies. Here we propose an approach -- Robust Imputation of Multi-tissue Expression incorporating EQTLs (RIMEE) to impute transcriptome in missing tissues by taking advantage of information in related tissues, related genes, and moreover, eQTLs. Based on cross-validation analyses of the nine tissues in the GTEx data, we evaluate the performance of imputing GTEx missing tissues, assess the contributions of cis- and trans-eQTLs in imputation, examine the feasibility of using GTEx as reference and imputing based on accessible tissues for future studies, and provide insights into the inter-tissue predictiveness and relatedness.

email: **lchen@health.bsd.uchicago.edu**

A BAYESIAN METHOD FOR THE DETECTION OF LONG-RANGE CHROMOSOMAL INTERACTIONS IN Hi-C DATA

Zheng Xu, University of North Carolina, Chapel Hill

Guosheng Zhang, University of North Carolina, Chapel Hill

Fulai Jin, Ludwig Institute for Cancer Research

Mengjie Chen, University of North Carolina, Chapel Hill

Patrick F. Sullivan, University of North Carolina, Chapel Hill

Zhaohui Qin, Emory University

Terrence S. Furey, University of North Carolina, Chapel Hill

Ming Hu, New York University

Yun Li*, University of North Carolina, Chapel Hill

Advances in chromosome conformation capture and next-generation sequencing technologies are enabling genome-wide investigation of dynamic chromatin interactions. For example, Hi-C experiments generate genome-wide contact frequencies between pairs of loci by sequencing DNA segments ligated from loci in close spatial proximity. One essential task in such studies is peak calling, that is, the identification of non-random interactions between loci from the two-dimensional contact frequency matrix. Successful fulfillment of this task has many important implications including identifying long-range interactions that assist in interpreting a sizable fraction of the results

from genome-wide association studies (GWAS). The task - distinguishing biologically meaningful chromatin interactions from massive numbers of random interactions - poses great challenges both statistically and computationally. Model based methods to address this challenge are still lacking. In particular, no statistical model exists that takes the underlying dependency structure into consideration. We propose a hidden Markov random field (HMRF) based Bayesian method to rigorously model interaction probabilities in the two-dimensional space based on the contact frequency matrix. By borrowing information from neighboring loci pairs, our method demonstrates superior reproducibility and statistical power in both simulations and real data.

email: **yunli@med.unc.edu**

FINE MAPPING OF COMPLEX TRAIT LOCI WITH COALESCENT METHODS IN LARGE CASE-CONTROL STUDIES

Ziqan Geng, University of Michigan

Paul Scheet, University of Texas MD Andersen Cancer Center

Sebastian Zöllner*, University of Michigan

Identifying the risk variants underlying association signals for complex diseases is challenging due to the complicated dependence structure of linkage disequilibrium. By modeling the hereditary process of a target region, coalescent-based approaches improve this identification and model the probability of carrying risk variants at all loci jointly. These probabilities provide Bayesian credible regions (CR) for the location of risk variants. However, existing coalescent-based methods are computationally very challenging and can only be applied to samples below 200 individuals. Here, we propose a novel approach to overcome this limitation. First, we infer a set of clusters from the sampled haplotypes. Then, we apply coalescent-based approaches to approximate the genealogy of the clusters. Hence, the dimension of external nodes in coalescent models is reduced from the total sample size to the number of clusters. We evaluate the cluster genealogy and their descendants phenotype distribution, to integrate over all positions and establish CRs where risk variants are most likely to occur. In simulation studies, our method correctly localizes short segments around true risk positions in datasets with thousands of individuals. Thus we have developed a novel approach to estimate the genealogy of sequenced regions that can be applied to very large case-control datasets.

email: **szoellne@umich.edu**

38. Big Data: Issues in Biosciences

BIG GENOMICS DATA ANALYTICS

Haiyan Huang*, University of California, Berkeley

Bin Yu, University of California, Berkeley

Explosive high-throughput technologies in the last decade demand developments of useful high dimensional statistical tools for systematic analyses of large genomics data such as gene expression, SNP, clinical and cellular data. An Integrative approach to the analysis of these data can push the statistical inference in and understanding of systems biology onto another level, beyond where traditional research stops. In this talk, we envision an integrative genomics data analysis that is built on the big data analytics platform such as SPARK, which respects computational and memory efficiencies, in addition to statistical power and robustness.

email: **hhuang@stat.berkeley.edu**

RECALCULATING THE RELATIVE RISKS OF AIR POLLUTIONTO ACCOUNT FOR PREFERENTIAL SITE SELECTION

James V. Zidek*, University of British Columbia

Gavin Shaddick, University of Bath

In the 1960s, over 2000 sites in the UK monitored black smoke (BS) air pollution due to concerns about its effect on public health that were clearly demonstrated by the famous London fog of 1952. Abatement measures led to a decline in the levels of BS and hence a reduction in the number of monitoring sites to less than 200 by 1996. A case study demonstrates that sites were removed preferentially, causing exaggerated estimates of pollution levels. This talk will describe methods for mitigating the effects of that overestimation and show through the case study that the relative risk of environmental health outcomes has been underestimated. The large number of monitoring sites and their associated high dimensional data vectors rule out naive use of classical geostatistical and Bayes-

ian hierarchical methods. Hence there is a need for novel approaches in analysis, which will be described. The work has important general implications for the setting of regulatory standards and the design of monitoring networks.

email: **jim@stat.ubc.ca**

FUNCTIONAL DATA ANALYSIS FOR QUANTIFYING BRAIN CONNECTIVITY

Hans-Georg Mueller*, University of California, Davis

Alexander Petersen, University of California, Davis

Owen Carmichael, Louisiana State University

Functional data analysis provides a toolbox for the analysis of data samples that can be viewed as being generated by repeated realizations of an underlying (and often latent) stochastic process. The application of this methodology to paired processes (X,Y) will be illustrated by the problem of quantifying connectivity for resting state fMRI data, where for each subject and each brain voxel, a BOLD time signal is recorded. The functional data analysis approach leads to various measures of functional correlation between X and Y. The resulting correlations between brain hubs provide a basis for the construction of subject-specific networks. A second application of functional data analysis is based on a novel construction of network functions that reflect inter- and intra-hub connectivity during resting state of the brain for each subject. A sample of networks is then represented by a sample of network functions, which may be represented by functional principal components. We apply these approaches to investigate the relations between brain connectivity and subject characteristics.

email: **hgmueller@ucdavis.edu**

39. Recent Advances in Statistical Ecology

EFFICIENT SPATIAL AND SPATIO-TEMPORAL FALSE DISCOVERY RATE CONTROL

Ali Arab*, Georgetown University

Analysis of spatial and spatio-temporal data often requires a large number of hypothesis tests. For example, one may test spatial data to detect activation areas or identify changes in the environment. Given the large number of hypothesis tests in these settings, type I error control cannot be effectively achieved using conventional multiple testing procedures such as Bonferroni. Alternatively, one may use the False Discovery Rate (FDR) control. However, FDR for applications with large number of tests may be inefficient due to low statistical power. For data with spatial or spatio-temporal structure, we may benefit from the spatial (and temporal) characteristics of the data in order to conduct tests in a multiresolution fashion (large scale to fine scale). In this paper, we provide an overview of existing methods and propose a hierarchical multiresolution approach to conduct FDR control for spatial and spatio-temporal signals. The proposed method results in higher efficiency (i.e., improved power) compared to existing FDR methods. The proposed method also allows flexibility for the researcher to focus the analysis in hypothesized activation areas rather than the whole domain.

email: **ali.arab@georgetown.edu**

MIXTURE OF INHOMOGENEOUS MATRIX MODELS FOR SPECIES-RICH ECOSYSTEMS

Frederic Mortier*, CIRAD - Tropical Forest Goods and Ecosystem Services Unit

Understanding how environmental factors could impact population dynamics is of primary importance for species conservation. Matrix population models are widely used to predict population dynamics. However, in species-rich ecosystems with many rare species, the small population sizes hinder a good fit of species-specific models. In addition, classical matrix models do not take into account environmental variability. We propose a mixture of regression models with variable selection allowing the simultaneous clustering of species into groups according to vital rate information (recruitment, growth, and mortality) and the identification of group-specific explicative environmental variables. We develop an inference method coupling the R packages flexmix and glmnet. We first highlight the effectiveness of the method on simulated datasets. Next, we apply it to data from a tropical rain forest in the Central African Republic. We demonstrate the accuracy of the inhomogeneous mixture matrix model in successfully reproducing stand dynamics and classifying tree species into well-differentiated groups with clear ecological interpretations.

email: **fmortier@cirad.fr**

SPATIO-TEMPORAL MODELING OF MULTIPLE SPECIES MIGRATION FLOW

Trevor J. Oswald*, University of Missouri

There are many complex interactions between species, including competition, predation, and mutualism, to name a few of the most commonly understood interactions. These interactions between species can lead to variations in their migratory patterns. Historically, much of the previous research has focused on migratory patterns of a single species between a small number of segregated regions (i.e., metapopulation analysis). Drawing concepts from metapopulation analysis, social demography, spatial econometrics, and dynamical spatio-temporal modeling, we extend the previous work by developing a methodology that can predict the migratory flows of several populations simultaneously. More generally, this model can be used for multivariate population flows in other areas of science. The model we propose makes use of spatio-temporal dynamics of the system, while accounting for uncertainty exhibited in the sampling, as well as the process underlying the migration flows within a Bayesian hierarachical modeling framework.

e-mail: **oswald.trev@hotmail.com**

STATISTICAL MODELING OF SPATIAL DISCRETE AND CONTINUOUS DATA IN ECOLOGY

Jun Zhu*, University of Wisconsin, Madison

Modeling spatial data in ecology and drawing statistical inference is challenging especially when response is possibly discrete and/or the sample size is big. The motivating examples include data derived from the land surveys in various regions of the United States, which require statistical methods for data analysis that balance modeling complexity, statistical efficiency, and computational feasibility. In this talk, some of the existing methodology for spatial discrete/continuous data is reviewed and new approaches are proposed. In particular, models for spatial count, ordinal, nominal, and proportional data are considered. Comparisons and connections will be drawn between different data types and modeling approaches. For illustration, the methods are applied to analyze land cover data for mapping and inferring about forest landscape structures.

email: **jzhu@stat.wisc.edu**

40. New Analytical Issues in Current Epidemiology Studies of HIV and other Sexually Transmitted Infections

A FRAMEWORK FOR QUANTIFYING RISK STRATIFICATION FROM DIAGNOSTIC TESTS: APPLICATION TO HPV TESTING IN CERVICAL CANCER SCREENING

Hormuzd Katki*, National Cancer Institute, National Institutes of Health

A test or biomarker that stratifies disease risks allows clinicians to only intervene on only those who have or will develop disease. We propose a general framework for risk stratification by introducing the risk stratification distribution, which is the distribution of the changes in disease risk indicated by each possible test result. The mean of this distribution, the mean risk stratification (MRS) is the average amount of extra disease (or deficit of disease) that a test reveals for an individual patient. The MRS is a function of not only the risk difference, but also marker positivity, demonstrating that a big risk difference does not imply good risk stratification for markers that are rarely positive. The MRS is also a function of Youden's index and disease prevalence, demonstrating that a large Youden's index does not imply good risk stratification if disease is too rare. We demonstrate that the net expected benefit of a diagnostic test is a function of test characteristics solely through the MRS. Reasoning based on MRS enforces rational decision-making based on the principle of "equal management of equal risks". We discuss examples from the presenter's experience serving on the guidelines committee for HPV testing in cervical cancer screening.

email: **katkih@mail.nih.gov**

COMBINING INFORMATION TO ESTIMATE ADHERENCE IN TRIALS OF PRE-EXPOSURE PROPHYLAXIS FOR HIV PREVENTION

James Hughes*, University of Washington

In trials of pre-exposure prophylaxis (PrEP) for HIV prevention understanding the effect of adherence on treatment efficacy is of great interest. However, even though most PrEP studies collect

multiple measures of adherence (e.g., self-report, pill counts, plasma drug levels), no rigorous approach for combining these various sources of information has been developed. We develop novel methods for combining these measures to estimate adherence using a latent variable model in a Bayesian framework. The approach is applied to data from a trial of intermittent PrEP use to understand variability in levels and patterns of adherence. We show how the methods can also provide insights about the utility of each measure for estimating adherence

email: **jphughes@u.washington.edu**

ANALYSIS OF LONGITUDINAL MULTIVARIATE OUTCOME DATA FROM COUPLES COHORT STUDIES: APPLICATION TO HPV TRANSMISSION DYNAMICS

Xiangrong Kong*, Johns Hopkins University

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 developed 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 was applied on the motivating HPV couple cohort data from the male circumcision (MC) trial in Uganda to assess the effect of MC on HPV transmission. Age stratified analysis was also conducted to understand the natural history of HPV infection and explain the mechanisms through which MC reduced HPV detections in men and women.

email: **xikong@jhsph.edu**

SAMPLE SIZE METHODS FOR ESTIMATING HIV INCIDENCE FROM CROSS-SECTIONAL SURVEYS

Jacob Moss Konikoff*, University of California, Los Angeles

Ron Brookmeyer, University of California, Los Angeles

Understanding HIV incidence, the rate at which new infections occur in populations, is critical for tracking and surveillance of the epidemic. In this paper we derive methods for determining sample sizes for cross-sectional surveys to estimate incidence with precision and to detect changes in incidence from two successive cross-sectional surveys with adequate power. In these surveys biomarkers such as CD4 cell count, viral load and recently developed serological assays are used to determine which individuals are in an early disease stage of infection. The total number of individuals in this stage, divided by the number of people who are uninfected, is used to approximate the incidence rate. Our methods account for uncertainty in the durations of time spent in the biomarker defined early disease stage. We find that failure to account for these uncertainties when designing surveys can lead to imprecise estimates of incidence and underpowered studies. We evaluated our sample size methods in simulations and found that they performed well in a variety of underlying epidemics. Code for implementing our methods in R is available from the authors upon request.

email: **jacob@konikoff.com**

DEVELOPMENT OF ACCURATE METHODS TO ESTIMATE HIV INCIDENCE IN CROSS-SECTIONAL SURVEYS

Oliver B. Laeyendecker*, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Accurate methods of estimating HIV incidence from cross-sectional surveys are critical to monitoring the epidemic and determining the population level impact of prevention efforts. Using 1000s of samples from individuals with known duration of infection we combined antibody titer and avidity assays with CD4 and viral load testing in a multi assay algorithm (MAA) where the mean duration that individuals appear recently infected was 159 days (95% CI 134, 186). We validated the method in three longitudinal cohorts, where the incidence estimated using the MAA was nearly identical to the observed incidence in the cohorts. For

a low incidence cohort of women at risk for HIV infection (HPTN064) the observed incidence was 0.24% (95% CI 0.07, 0.62) compared to 0.26% (95% CI 0.03, 0.95) using our MAA. In a moderate incidence population of individuals from a vaccine preparedness cohort (HIVNET001) the observed incidence was 1.04% (95% CI 0.70, 1.55) compared to 1.09% (95% CI 0.60, 1.84) using our MAA. In a high incidence cohort of African American MSM (HPTN061), the observed incidence was 3.02% (95% CI 2.01, 4.37) compared to 3.44% (95% CI 1.75, 6.20) using our MAA. A MAA provides a powerful tool to estimate population level HIV incidence.

email: **olaeyen1@jhmi.edu**

41. Statistical Advances and Challenges in Mobile Health

MICRO-RANDOMIZED TRIALS AND mHEALTH

Peng Liao, University of Michigan

Pedja Klasjna, University of Michigan

Ambuj Tewari, University of Michigan

Susan Murphy*, University of Michigan

Micro-randomized trials are trials in which individuals are randomized 100's or 1000's of times over the course of the study. The goal of these trials is to assess the impact of momentary interventions, e.g. interventions that are intended to impact behavior over small time intervals. A fast growing area of mHealth concerns the use of mobile devices for both collecting real-time data, for processing this data and for providing momentary interventions. We discuss the design and analysis of these types of trials.

email: **samurphy@umich.edu**

NOT EVERYBODY, BUT SOME PEOPLE MOVE LIKE YOU

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

Accelerometers are now used extensively in health studies, where they increasingly replace self-report questionnaires. The sudden success of accelerometers in these studies is due to the fact that they are cheap, easy to wear, collect millions of data points at high frequency (10-100Hz or more), store months worth of data, and can be paired with other devices, such as heart, gps, or skin temperature sensors. I will discuss the multi-resolution structure of the data and will introduce methods for movement recognition both for in-the-lab and in-the-wild data using second- and sub-second level data. I will introduce movelets, a powerful dictionary learning approach, designed for quick identification of movement patterns. At the minute level I will describe activity intensity measures (activity counts, vector magnitude, and activity intensity) and introduce functional data approaches for characterizing the circadian rhythm of activity and its association with health. The natural data structure induced by such observational studies is that of multilevel functional data (activity intensity measured at every minute for multiple days observed within each subject.) I will introduce fast functional data analysis approaches that can deal with the data complexity, describe its structure and its association with health outcomes. In particular, I will discuss results for two motivating studies: 1) the association between age and the circadian rhythm of activity; and 2) the association between mental health disorders and activity patterns.

email: **ccrainic@jhsph.edu**

SUPPORTING HEALTH MANAGEMENT IN EVERYDAY LIFE WITH MOBILE TECHNOLOGY

Predrag Klasnja*, University of Michigan

Susan A. Murphy, University of Michigan

Ambuj Tewari, University of Michigan

Mobile phones are becoming an increasingly important platform for the delivery of health interventions. Phones have been used to encourage physical activity and healthy diets, to monitor symptoms of asthma, heart disease, and chemotherapy side effects, and to send patients reminders to take medications and to attend appointments. In this talk, I will discuss why mobile phones are particularly well suited for creation of innovative and effective health interventions, I will review our work on using mobile phones to encourage physical activity, and I will highlight some of the challenges involved in developing and evaluating mHealth technologies.

email: **klasnja@umich.edu**

MEASURING STRESS AND ADDICTIVE BEHAVIORS FROM MOBILE PHYSIOLOGICAL SENSORS

Santosh Kumar*, University of Memphis

Emre Ertin, The Ohio State University

Mustafa al'Absi, University of Minnesota

David Epstein, National Institute on Drug Abuse, National Institutes of Health

Kenzie Preston, National Institute on Drug Abuse, National Institutes of Health

Annie Umbricht, Johns Hopkins University

Recent advances in the sensing and computational capacity of mobile devices have opened up enormous opportunities to improve patients' health and well-being. They can quantify dynamic changes in an individual's health state as well as key physical, biological, behavioral, social, and environmental factors that contribute to health and disease risk, anytime and anywhere. Such real-time monitoring can accelerate health research and optimize care delivery, e.g., via just-in-time personalized interventions. In this talk, I will present a computational model for automatically detecting stress, smoking, and cocaine use from mobile physiological sensors in the AutoSense suite.

email: **santosh.kumar@memphis.edu**

42. CONTRIBUTED PAPERS: Survey Research

ORDINAL BAYESIAN INSTRUMENT DEVELOPMENT: NEW KID ON THE PATIENT REPORTED OUTCOME MEASURES BLOCK

Lili Garrard*, University of Kansas Medical Center

Larry R. Price, Texas State University

Marjorie J. Bott, University of Kansas

Byron J. Gajewski, University of Kansas Medical Center

Traditional instrument development is often challenged by psychometric difficulties when the target audience represents small populations or rare diseases. We propose an innovative Ordinal Bayesian Instrument Development (OBID) method that seamlessly integrates expert and participant data in a Bayesian factor analysis framework, while utilizing fewer subjects than classical approaches and maintaining coherent validity evidence. When the instrument consists of all ordinal items, the ordinal factor analysis model is equivalent to a two parameter item response theory (IRT) model with a probit link. Prior distributions obtained from expert data are imposed on the IRT parameters and are updated with participants' data. Simulation data are used to demonstrate the efficiency of OBID by comparing its performance to classical instrument development with exact estimate procedures.

email: **lgarrard@kumc.edu**

QUANTIFYING PARENTAL HISTORY IN SURVEY DATA

Rengyi Xu*, University of Pennsylvania

Sara B. DeMauro, University of Pennsylvania

Rui Feng, University of Pennsylvania

Parental history has been identified as an important risk factor for the incidence of many diseases in their offspring. Most existing literatures use a binary indicator to quantify parental history. However, in some diseases, such as asthma, parent's age at onset of the disease might increase children's risk. Therefore, an estimator that incorporates parent's age at onset is desirable. When the data are collected from national household surveys, the complex survey sampling design needs to be taken into consideration. In this study, we develop a continuous standardized score, the so-called log-rank risk score, to quantify parental history that incorporates both the occurrence of disease and the age at onset for survey data. The proposed method is evaluated using the third National Health and Nutrition Examination Survey data to examine the separate effects of maternal and paternal history on the onset of asthma in children and to evaluate the relationship between age of asthma onset in parents and risk of asthma in their children. Using our new risk scores leads to smaller standard errors and thus more precise estimates than using a binary indicator. Our results also show children whose mother has an earlier age at onset have an increased risk.

email: **xurengyi@mail.med.upenn.edu**

BAYESIAN NONPARAMETRIC WEIGHTED SAMPLING INFERENCE

Yajuan Si*, University of Wisconsin, Madison

Natesh S. Pillai, Harvard University

Andrew Gelman, Columbia University

It has historically been a challenge to perform Bayesian inference in a design-based survey context. The present paper develops a Bayesian model for sampling inference in the presence of inverse-probability weights. We use a hierarchical approach in which we model the distribution of the weights of the nonsampled units in the population and simultaneously include them as predictors in a nonparametric Gaussian process regression. We use simulation studies to evaluate the performance of our procedure and compare it to the classical design-based estimator. We apply our method to the Fragile Family Child Wellbeing Study. Our studies find the Bayesian nonparametric finite population estimator to be more robust than the classical design-based estimator without loss in efficiency, which works because we induce regularization for small cells and thus this is a way of automatically smoothing the highly variable weights.

email: **ysi@biostat.wisc.edu**

HOW TO BEST COMPUTE PROPENSITY SCORES IN COMPLEX SAMPLES IN RELATION TO SURVEY WEIGHTS

Keith W. Zirkle*, Virginia Commonwealth University

Adam P. Sima, Virginia Commonwealth University

Survey data are usually not collected via simple random sampling so that inferences upon an intended population require the use of survey weights. However, similar to any observational study, confounding from a number of covariates can bias the results if not properly taken into account. Propensity score weighting is a popular method to address confounding in a sample. There is considerable interest in how to best combine survey weight and propensity score information when analyzing data. The literature suggests that when computing propensity scores within a complex sample, survey weights should be treated as an effect that contributes to the propensity score. Alternatively, survey weights could be used in their proper purpose when calculating the propensity scores. This study assesses the optimal use for survey weights in the calculation of propensity scores when the goal of the research is to draw inferences upon the intended population. A Monte Carlo simulation showed that the proper use of the survey weights never performed worse than the methods recommended by the literature and outperformed these methods when there was a strong level of confounding and a strong contribution of confounding variables to the survey weights. An application of this method was demonstrated using the NHANES dataset in assessing racial differences in prostate cancer screening.

email: **zirklekw@vcu.edu**

MULTIPLE IMPUTATION OF THE ACCELEROMETER DATA IN THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

Benmei Liu*, National Cancer Institute, National Institutes of Health

Mandi Yu, National Cancer Institute, National Institutes of Health

Barry I. Graubard, National Cancer Institute, National Institutes of Health

Richard Troiano, National Cancer Institute, National Institutes of Health

Nathaniel Schenker, National Center for Health Statistics, Centers for Disease Control and Prevention

The Physical Activity Monitor (PAM) component was introduced to the 2003-2004 National Health and Nutrition Examination Survey (NHANES) to collect objective information on physical activity including both movement intensity counts and steps. Due to an error in the accelerometer device initialization process, the steps data was missing for all participants in several primary sampling units (PSUs), typically a single county or group of contiguous counties, who had intensity count data from their accelerometers. To avoid potential bias and loss in efficiency in estimation and

inference involving the steps data, we considered methods to accurately impute the missing values for steps collected in the 2003-2004 NHANES. We proposed a multiple imputation approach based on a semi-parametric regression technique where the Alternating Conditional Expectation model is utilized to improve the imputation accuracy. This paper describes the approaches used in this imputation and evaluates the methods by comparing the distributions of the original and the imputed data. A simulation study using the observed data is also conducted as part of the model diagnostics. Finally some real data analyses are performed to compare the before and after imputation results.

email: **benmeiliu@hotmail.com**

SPLIT QUESTIONNAIRE SURVEY DESIGN IN THE LONGITUDINAL SETTING

Paul M. Imbriano*, University of Michigan

Trivellore E. Raghunathan, University of Michigan

Advancements in survey administration methodology and multiple imputation software now make it possible for planned missing data designs to be implemented for improving data quality through the reduction in survey length. Many papers have discussed implementing a cross sectional study with planned missing data using a split-questionnaire design, but the research in applying these methods to a longitudinal study has been limited. Using simulations and data from the Health and Retirement Study (HRS), we compared the performance of several methods for administering a split-questionnaire design in the longitudinal setting. Our findings suggest that the optimal design depends on the data structure, specifically on both the within-wave and between-wave variable correlations, and there exists a trade-off between the complexity and robustness of the design. These factors should be taken into account when constructing a longitudinal study with planned missing data.

email: **pimbri@umich.edu**

43. CONTRIBUTED PAPERS: Graphical Models

REGRESSION ANALYSIS OF NETWORKED DATA

Yan Zhou*, University of Michigan

Peter X. K. Song, University of Michigan

This paper concerns the development of a new regression analysis methodology to assess relationships between multi-dimensional response variables and covariates that are correlated through networks. To address analytic challenges pertaining to the integration of network topology into the regression analysis, we propose a method of hybrid quadratic inference functions (HQIF) that utilizes both prior and data-driven correlations among network nodes into statistical estimation and inference. Moreover, a Godambe information based tuning strategy is proposed to allocate weights between the prior and data-driven pieces of network knowledge, so that the resulting estimation achieves desirable efficiency. The proposed method is conceptually simple and computationally fast, and more importantly has appealing large-sample properties in both estimation and inference. This new methodology is evaluated through simulation studies and illustrated by a motivating example of neuroimaging data about an association study of iron deficiency on infant's auditory recognition memory.

email: **zhouyan@umich.edu**

INTEGRATIVE ANALYSIS OF GENETICAL GENOMICS DATA INCORPORATING NETWORK STRUCTURE

Bin Gao*, Michigan State University

Yuehua Cui, Michigan State University

Genetical genomics analysis provides promising opportunities to infer gene regulations and predict phenotypic variation by combining genetic variants, gene expressions, and phenotype data together. In this work, we use gene expressions to predict phenotypic response while considering the graphical structures on gene networks. Given that gene expressions are intermediate phenotypes between a trait and genetic variants, we follow the instrumental variable regression framework proposed by Lin et al. (2014) and treat genetic variants as instrumental variables to improve the prediction. In addition, we adopt a covariate-adjusted graph learning approach to improve the graphical structure of gene expression network. We propose a 2-step estimation procedure. In step 1, we apply the glasso algorithm to learn graphical structures on gene expressions while adjusting for the effects of genetic variants. In step 2, we use the predicted

expressions obtained from the first step as predictors while adopting a network constrained regularization method based on the updated graphical structures obtained from the first step to obtain better coefficients estimates. We establish the selection and estimation consistency of the 2-step estimation procedure. The utility of the method is demonstrated with extensive simulations and application to a mouse obesity dataset.

email: **gaobinmath@gmail.com**

ESTIMATING A GRAPHICAL INTRA-CLASS CORRELATION COEFFICIENT (GICC) USING MULTIVARIATE PROBIT-LINEAR MIXED MODELS

Chen Yue*, Johns Hopkins University

Shaojie Chen, Johns Hopkins University

Haris Sair, Johns Hopkins University

Raag Airan, Johns Hopkins University

Brian Caffo, Johns Hopkins University

Data reproducibility is a critical issue in all scientific experiments. In this manuscript, the problem of quantifying the reproducibility of graphical measurements is considered. The concept of image intra-class correlation coefficient (I2C2) is generalized and the concept of the graphical intra-class correlation coeffiecient (GICC) is proposed for such purpose. The concept of GICC is based on multivariate probit-linear mixed eect models. A Markov Chain EM (MCEM) algorithm is used for estimating the GICC. Simulations results with varied settings are demonstrated and our method is applied to the KIRBY21 test-retest dataset.

email: **cyue1@jhu.edu**

ESTIMATION OF DIRECTED SUBNETWORKS IN ULTRA HIGH DIMENSIONAL DATA FOR GENE NETWORK PROBLEM

Sung Won Han*, New York University

Hua (Judy) Zhong, New York University

Pathway analysis or gene-gene interaction study is very important to find the genome-wide mechanism of the cancer development. The directed acyclic graph is a commonly used model to estimate a network with gene regulatory. However, due to the advent and the diminishing cost of high-throughput data, ultra high dimensional datasets are currently generated. The estimation of DAGs is a NP-hard problem, and the computational time is large even in middle-sized data. The estimation of whole gene network based on gene expressions such as mRNAs is the problem of estimating DAGs with ultra high dimensional data. Thus, it is computationally infeasible with reasonable time to estimate causal relationship by using existing well-known methods. In most problem, they estimate the network with a few gene expressions, which known as a group with an important role. Thus, finding directed interactions within a group of similar functional genes is a reasonable approach if finding entire directed interactions is not feasible. In this presentation, we discuss how to estimate directed subnetworks in ultra high dimensional data by using some techniques used in social network problems.

email: **sungwonhan2@gmail.com**

LONGITUDINAL GRAPHICAL MODELS: OPTIMAL ESTIMATION AND ASYMPTOTIC INFERENCE

Quanquan Gu*, Princeton University

Yuan Cao, Princeton University

Yang Ning, Princeton University

Han Liu, Princeton University

In this paper, we propose a new semi-parametric graphical model, namely Longitudinal Graphical Model, for continuous longitudinal data with irregular followup. It models the joint distribution of d covariates at T time stamps, that are normally distributed with time varying mean and a common covariance matrix. Our goal is to estimate the common covariance structure shared by different time stamps. We present a novel parameter estimation algorithm based on conditioning and QR decomposition, which treats the time varying mean vectors as nuisance parameters. The proposed graphical model includes Gaussian graphical model as a special example. Under certain mild assumptions, we establish the parameter estimation error bounds in terms of different norms, which are of optimal rate. Furthermore, we study hypothesis test of the sparse precision matrix in the proposed graphical model. We propose score, Wald and likelihood ratio tests for the element of the sparse precision matrix of the graphical model. The asymptotic properties of these hypothesis tests are analyzed. In addition, we study multiple hypothesis test of the sparse precision matrix in the proposal graphical model. It is based on multiplier bootstrap technique and is able to test multiple

elements (i.e., a subgraph) of the precision matrix. The asymptotic property of the multiple hypothesis test are also established. We verify these appealing theoretical properties of the proposed graphical model through both simulations on synthetic datasets, and a real world microarray dataset.

email: **qgu@princeton.edu**

JOINTLY ESTIMATING GAUSSIAN GRAPHICAL MODELS FOR SPATIAL AND TEMPORAL DATA

Zhixiang Lin*, Yale University

Tao Wang, Yale University

Can Yang, Hong Kong Baptist University

Hongyu Zhao, Yale University

In this paper, we first propose a Bayesian neighborhood selection procedure to estimate Gaussian Graphical Models (GGMs). Our procedure is then extended to jointly estimate GGMs in multiple groups of data with complex structure, including spatial data, temporal data and data with both spatial and temporal structures. In our approach, Markov random field models are used to efficiently utilize the information embedded in the spatial and temporal structure. For the estimation of single GGM, we show the graph selection consistency of the proposed method in the sense that the posterior probability of the true model converges to one. We develop and implement an efficient algorithm for statistical inference. For 1,000 iterations of Gibbs sampling, the computational time is about 30 seconds for one graph with 100 nodes. Simulation studies suggest that our approach achieves better accuracy in network estimation compared with models not incorporating spatial and temporal dependency. Finally, we illustrate our method on the human brain gene expression microarray dataset, where the expression levels of genes are measured in different brain regions across multiple time periods.

email: **zhixiang.lin@yale.edu**

44. CONTRIBUTED PAPERS: Joint Models for Longitudinal and Survival Data

JOINT MODELING OF BIVARIATE LONGITUDINAL AND BIVARIATE SURVIVAL DATA IN SPOUSE PAIRS

Jia-Yuh Chen*, University of Pittsburgh

Stewart J. Anderson, University of Pittsburgh

Joint modeling of longitudinal and survival data has become increasingly useful for analyzing clinical trials data. Recent multivariate joint models relate one or more longitudinal outcomes to one or more failure times (e.g., competing risks) in the same subject. We consider a case where longitudinal and survival outcomes are measured in subject pairs (e.g., married couples). We propose a bivariate joint model incorporating within-pair correlations, both in the longitudinal and survival processes. We use a bivariate linear mixed-effects model for the longitudinal process where the random effects are used to model the temporal correlation among longitudinal outcomes and the correlation between different outcomes. For the survival process, we use a gamma frailty in a Weibull model to account for the correlation between survival times within pairs. The sub-models are then linked through shared, latent random effects, where the longitudinal and survival processes are conditionally independent given the random effects. Parameter estimates are obtained by maximizing the joint likelihood for the bivariate longitudinal and bivariate survival data using the EM algorithm.

email: **jic29@pitt.edu**

JOINT MODEL OF BIVARIATE SURVIVAL TIMES AND LONGITUDINAL DATA

Ke Liu*, University of Iowa

Ying Zhang, University of Iowa

Motivated by a study of muscular dystrophy in MD STARnet a joint model of bivariate survival times and longitudinal data is developed. We propose to analyze correlated bivariate survival responses associated with a longitudinal biomarker in the Frequentist paradigm. A Gamma frailty variable is used to account for the correlation between the two correlated survival outcomes in addition to the random variables that account for the correlation between the survival times and longitudinal maker. The EM algorithm is adopted to compute the maximum profile likelihood estimate. The bootstrap method is applied to estimate the standard error of estimated model parameters. The simulation study is conducted to demonstrate the validity of the proposed methodology. Finally the method is applied to the MD STAR net for illustration.

email: **ke-liu@uiowa.edu**

DYNAMIC PREDICTION OF ACUTE GRAFT-VERSUS-HOST DISEASE WITH TIME-DEPENDENT COVARIATES

Yumeng Li*, University of Michigan

Thomas M. Braun, University of Michigan

Acute Graft-versus-Host Disease (aGVHD) is a side-effect of hematopoietic cell transplantation (HCT) and is a leading cause of death in patients receiving HCTs. Thus, investigators would like to have models that accurately predict those most likely to suffer from aGVHD in order to minimize over-treatment of patients as well as reduce mortality. To this end, we propose using biomarkers (that are collected weekly) to predict future biomarker values and the time-to-aGVHD through both joint modeling and multi-state model methods. We consider settings in which the biomarkers are continuous or binary (above or below a threshold), and aGVHD is treated as binary or as a time-to-event (and possibly censored) outcome. We present simulation results for various models using settings based upon actual data collected at the University of Michigan Blood and Marrow Transplant Program.

email: **yumeng@umich.edu**

THE JOINT MODELLING OF RECURRENT EVENTS AND OTHER FAILURE TIME EVENTS

Luojun Wang*, The Pennsylvania State University

Vernon M. Chinchilli, The Pennsylvania State University

In many biomedical studies and clinical trials, recurrent events are commonly encountered, indicating progression in treatment or disease. When recurrent events are correlated with another failure event, such as death, we no longer should assume an independent censoring mechanism for the failure event. Huang and Wang (2004) proposed a joint modeling of a recurrent event process and a failure time in which a common, subject-specific, latent variable is used to model the association between the intensity of the recurrent event process and the hazard function of the failure time. However, in this setting, the correlation between the number of recurrent events occurring before the failure time or censoring time needs to be positive. Another model to consider is to construct a Farlie-Gumbel-Morgenstern (FGM) bivariate density function for the recurrent events and the failure time, in which the correlation between the recurrent events and the failure time or censoring time could be either positive or negative. The drawback to the FGM bivariate density is that it only can accommodate a weak level of correlation, i.e., the correlation cannot approach the boundaries of -1 or +1, as desired. In this work, we propose an alternative to the FGM bivariate density for the recurrent events and the failure time that can account for a stronger correlation. We illustrate the model and analysis using data in which the recurrent event is the occurrence of acute kidney injury (AKI) and the failure event is death.

email: **vicwong@psu.edu**

A BAYESIAN APPROACH FOR JOINT MODELING OF LONGITUDINAL MENSTRUAL CYCLE LENGTH AND FECUNDITY

Kirsten J. Lum*, Johns Hopkins University and Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Germaine M. Buck Louis, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Thomas A. Louis, Johns Hopkins University and U.S. Census Bureau

Female menstrual cycle length is thought to play an important role in couple fecundity, or the biologic capacity for reproduction irrespective of pregnancy intentions. A complete assessment of the association between menstrual cycle length and fecundity requires a model that accounts for multiple risk factors (both male and female) and the couple's intercourse pattern relative to ovulation. We develop a Bayesian joint model consisting of a mixed effects accelerated

failure time model for longitudinal menstrual cycle lengths and a hierarchical model for the conditional probability of pregnancy in a menstrual cycle given no pregnancy in previous cycles of trying, in which we include covariates for the male and the female and a flexible spline function of intercourse timing. Using our joint modeling approach to analyze data from the Longitudinal Investigation of Fertility and the Environment Study, a couple based prospective pregnancy study, we found a significant quadratic relation between menstrual cycle length and the probability of pregnancy even with adjustment for other risk factors, including male semen quality, age, and smoking status.

email: **kirsten.lum@gmail.com**

JOINT ANALYSIS OF MULTIPLE LONGITUDINAL PROCESSES AND SURVIVAL DATA MEASURED ON NESTED TIME-SCALES USING SHARED PARAMETER MODELS: AN APPLICATION TO FECUNDITY DATA

Rajeshwari Sundaram*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Somak Chatterjee, George Washington University

Fecundity is defined as the biologic potential of men and women for reproduction, and is often measured by estimating the probability of pregnancy in each menstrual cycle among couples having regular unprotected intercourse. Estimating fecundity is challenging, in part, given the effect that varying patterns of sexual intercourse may have on the length of pregnancy attempt. Clinical guidance is sometimes sought to aid couples in timing intercourse acts around ovulation to minimize the time needed to achieve pregnancy. Empirical evidence delineating the timing of intercourse relative to ovulation are few, resulting in a generalized clinical recommendation to have intercourse every other day (Practice Committee of the American Society for Reproductive Medicine, 2013). Understanding the relation between fecundity, intercourse behavior and other relevant covariates is increasingly relevant given population level changes in the sociodemographic characteristics of reproductive aged couples such as an increase in age at first pregnancy. This may be associated with reduced intercourse activity, longer time-to-pregnancy, an increased prevalence of infertility or a combination of all these factors. Our main objective is to jointly model intercourse behavior, a binary longitudinal process (measured on day level), menstrual cycle characteristic (measured on monthly level and TTP, a survival outcome (on monthly timescale), with a view towards prediction of both longitudinal processes on differing timescales and time to pregnancy. This is achieved using an empirical bayes approach of joint modeling of multivariate longitunidal processes and time to event.

email: **sundaramr2@mail.nih.gov**

45. CONTRIBUTED PAPERS: Functional Data Analysis

GENERALIZED MULTILEVEL FUNCTION-ON-SCALAR REGRESSION AND PRINCIPAL COMPONENT ANALYSIS

Jeff Goldsmith*, Columbia University

Vadim Zipunnikov, Johns Hopkins University

Jennifer Schrack, Johns Hopkins University

We consider regression models for generalized, multilevel functional responses: functions are generalized in that they follow an exponential family distribution and multilevel in that they are clustered within groups or subjects. This data structure is increasingly common across scientific domains and is exemplified by our motivating example, in which binary curves indicating physical activity or inactivity are observed for nearly six hundred subjects over five days. We use a generalized linear model to incorporate scalar covariates into the mean structure, and decompose subject-specific and subject-day-specific deviations using multilevel functional principal components analysis. Model parameters are estimated in a Bayesian framework using Stan, a programming language that implements a Hamiltonian Monte Carlo sampler. Simulations designed to mimic the application have good estimation and inferential properties with reasonable computation times for moderate datasets, in both cross-sectional and multilevel scenarios; code is publicly available. In

the application we identify effects of age and BMI on the time-specific change in probability of being active over a twenty-four hour period.

email: **jeff.goldsmith@columbia.edu**

INFERENCE ON FIXED EFFECTS IN COMPLEX FUNCTIONAL MIXED MODELS

So Young Park*, North Carolina State University

Ana-Maria Staicu, North Carolina State University

Luo Xiao, Johns Hopkins Bloomberg School of Public Health

Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

We discuss statistical inference in regression models involving complex-correlated functional responses and scalar or vector covariates. Current inferential procedures are developed for independently sampled functional responses and are not directly applicable to functional data that are correlated because of a longitudinal or spatial design, for example. We use bootstrap methodology to construct confidence bands for the covariates effect on the mean response. Additionally, we introduce a testing procedure for testing scalar covariate effect. Our methods are illustrated in a thorough simulation experiment and on the motivating application - the Baltimore Longitudinal Study on Aging (BLSA), where daily physical activity is recorded repeatedly for several hundreds of subjects of various ages.

email: **spark13@ncsu.edu**

GENERALIZED FUNCTION-ON-FUNCTION REGRESSION

Janet S. Kim*, North Carolina State University

Ana-Maria Staicu, North Carolina State University

Arnab Maity, North Carolina State University

We consider a non-linear regression models for functional responses and functional predictors observed on possible different domains. We introduce a flexible model that relates the value of response at a particular time point to the covariate over the entire domain as well as the time point of the response. There are two innovations in this paper. First, we develop an inferential procedure that reduces the dimension of model parameters by orthogonal projection of the functional response. Second, the proposed method accommodates realistic settings such as correlated error structure as well as sparse and/or irregular design. We investigate our methodology in finite sample size through simulations and real data applications.

email: **jskim3@ncsu.edu**

VARIABLE SELECTION IN FUNCTION-ON-SCALAR REGRESSION

Yakuan Chen*, Columbia University

Todd Ogden, Columbia University

Jeff Goldsmith, Columbia University

The problem of variable selection often arises in the context of models with functional responses and scalar predictors. In comparison with traditional regression models, this setting is complicated by the dimensionality of the response and coefficient curves and by the correlation structure of the residuals. By expanding the coefficient functions using a B-spline basis, we pose the function-on-scalar model as a multivariate multiple regression problem. Spline coefficients are grouped within coefficient function, and group-minimax concave penalty (MCP) is used for variable selection. We adapt techniques from generalized least squares to account for residual covariance by "pre-whitening" using an estimate of the covariance matrix and develop an iterative algorithm that alternately updates the spline coefficients and covariance. Simulation results indicate that this iterative algorithm often performs as well as pre-whitening using the true covariance. We apply our method to two-dimensional planar reaching motions in a study of the effects of stroke severity on motor control, and find that our method provides lower prediction errors than competing methods.

email: **chenyakuan@gmail.com**

BAYESIAN ADAPTIVE FUNCTIONAL MODELS WITH APPLICATIONS TO COPY NUMBER DATA

Bruce D. Bugbee*, University of Texas MD Anderson Cancer Center

Veera Baladandayuthapani, University of Texas MD Anderson Cancer Center

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

We present a Bayesian framework for the analysis of functional covariates in a regression context. This is done with both global and spatially-adaptive regularization schemes for the population level

weight function. To accomplish this we develop both MCMC and variational Bayes approximation methods. In addition to highlighting the need for these models, we showcase the usefulness of variational approximations for exploring complex, high-dimensional data. Finally, we present investigations of a motivating example based on exploring relationships between copy number aberrations and leukemia.

email: **brucebugbee@gmail.com**

FUNCTIONAL BILINEAR REGRESSION WITH MATRIX COVARIATES VIA REPRODUCING KERNEL HILBERT SPACE WITH APPLICATIONS IN NEUROIMAGING DATA ANALYSIS

Dong Wang, University of North Carolina, Chapel Hill

Dan Yang*, Rutgers University

Haipeng Shen, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

Traditional functional linear regression usually takes a one dimensional functional predictor as input and estimates the continuous coefficient function. Modern applications often generate two dimensional covariates, which when observed at grid points are matrices. To avoid inefficiency of the classical method involving estimation of a two dimensional coefficient function, we propose a bilinear regression model and obtain estimates via a smoothness regularization method. The proposed estimator exhibits minimax optimal property for prediction under the framework of Reproducing Kernel Hilbert Space. The merits of the method are further demonstrated by numerical experiments and an application on real MRI imaging data.

email: **dyang@stat.rutgers.edu**

SIMULTANEOUS CONFIDENCE BANDS FOR DERIVATIVES OF DEPENDENT FUNCTIONAL DATA

Guanqun Cao*, Auburn University

In this work, consistent estimators and simultaneous confidence bands for the derivatives of mean functions are proposed when curves are repeatedly recorded for each subject. The within-curve correlation of trajectories has been considered while the proposed novel confidence bands still enjoys semiparametric efficiency. The proposed methods lead to a straightforward extension of the two-sample case in which we compare the derivatives of mean functions from two populations. We demonstrate in simulations that the proposed confidence bands are superior to existing approaches which ignore the within-curve dependence. The proposed methods are applied to investigate the derivatives of mortality rates from period lifetables that are repeatedly collected over many years for various countries.

email: **gzc0009@auburn.edu**

46. CONTRIBUTED PAPERS: Methods in Causal Inference: Instrumental Variable, Propensity Scores and Matching

METHODS TO OVERCOME VIOLATIONS OF AN INSTRUMENTAL VARIABLE ASSUMPTION: CONVERTING A CONFOUNDER INTO AN INSTRUMENT

Michelle Shardell*, National Institute on Aging, National Institutes of Health

Instrumental variable (IV) methods are a powerful tool for consistently estimating causal effects in the presence of unmeasured confounding. However, the validity of instrumental variable (IV) methods relies on strong assumptions, some of which cannot be conclusively empirically verified. One such assumption is that the effect of the proposed instrument on the outcome is completely mediated by the exposure of interest. We consider the situation where this assumption is violated, but a weaker assumption holds in which the effect of the proposed instrument on the outcome is completely mediated by measured variables, including the exposure of interest. In this case, the proposed instrument is actually a confounder. We review some conventional IV methods and propose easy-to-use adaptations of these methods for use when the usual IV assumption is violated, but the weaker assumption holds. The proposed methods involve first "converting" the confounder into an IV, then applying conventional IV methods. Potential applications of the proposed methods in epidemiology include studies where

the exposure and outcome are known to exhibit seasonal variation and Mendelian randomization studies with genetic variants that are known to affect multiple phenotypes that may affect the outcome.

email: **mshardel@epi.umaryland.edu**

ASSESSING TREATMENT EFFECT OF THIOPURINES ON CROHN'S DISEASE FROM A UK POPULATION-BASED STUDY USING PROPENSITY SCORE MATCHING

Laura H. Gunn*, Stetson University

Sukhdev Chatu, St. George's University Hospital London

Sonia Saxena, Imperial College London

Azeem Majeed, Imperial College London

Richard Pollok, St. George's University Hospital London

Randomized controlled trials (RCTs) are a 'gold standard' for estimating minimally unbiased treatment effects on health outcomes; however, RCTs are not always feasible and population-based observational studies can be more appropriate. The Clinical Practice Research Datalink (CPRD) contains clinical and prescribing data for over 13 million patients in the United Kingdom; participating primary care practices are subject to regular audit to ensure data accuracy and completeness, allowing epidemiological studies of this data to be feasible. Since RCTs evaluating the impact of Thiopurine treatment on Crohn's disease patients is not practical, we used CPRD data to identify 5,640 patients with incident Crohn's cases diagnosed between 1989 and 2005, with at least an additional 5-year follow-up to 2010. Propensity score matching (PSM) is used to reduce bias obtained in estimates of treatment effects as a result of confounding between baseline factors and exposure group status. This presentation describes the PSM process, and applies optimal PSM, with a sensitivity analysis implementing additional matching techniques, using data collected from this nationally representative UK population-based study, where impact of duration and timing of Thiopurine treatment on the likelihood of surgery is assessed using a Cox proportional hazards model and PSM.

email: **lgunn@stetson.edu**

SEMIPARAMETRIC CAUSAL INFERENCE IN MATCHED COHORT STUDIES

Edward H. Kennedy*, University of Pennsylvania

Dylan S. Small, University of Pennsylvania

Famously, odds ratios can be estimated in case-control studies using standard logistic regression, ignoring the outcome-dependent sampling. In this paper we prove an analogous result for treatment effects on the treated in cohort studies. Specifically, in studies where a sample of treated subjects is observed along with a separate sample of possibly matched controls, we show that efficient and doubly robust estimators of effects on the treated are computationally equivalent to standard estimators, which ignore the matching and exposure-based sampling. This is not the case for general average effects. With respect to issues of efficiency and study design, we show that matched cohort studies are often more efficient than random sampling for estimating effects on the treated, and we derive the optimal number of matches in such studies for a given set of matching variables. We illustrate our results via simulation and in an evaluation of the National Supported Work training program.

email: **edwardh.kennedy@gmail.com**

REVISITING THE COMPARISON OF COVARIATE ADJUSTED LOGISTIC REGRESSION VERSUS PROPENSITY SCORE METHODS WITH FEW EVENTS PER COVARIATE

Fang Xia*, Duke University School of Medicine

Phillip J. Schulte, Duke University School of Medicine

Laine Thomas, Duke University School of Medicine

When treatments are compared in observational data sources, adjustment for measured confounding is often achieved by covariate-adjustment or propensity score methods including inverse propensity weighting (IPW) and stratification by quintiles of the propensity score distribution. With binary outcomes, over-fitting may be a serious concern when there are fewer than 10 events per covariate. This is often cited as a reason to prefer propensity methods to logistic regression adjustment in the case of a rare outcome but common treatment. The recommendation is based on the median bias observed in a single simulation study of propensity stratified versus covariate adjusted logistic regression under conditions where the total number of events

was typically less than 16. It is unclear whether this result would generalize to well-powered studies with a large number of covariates, or to IPW as opposed to stratification. In order to clarify the relative performance of these methods, we conducted a simulation study across a range of conditions with at least 20 total events. All three methods demonstrated minimal bias and similar performance, even with as few as 2 events per confounder. Our results suggest that all of these techniques remain an option for many adjustment applications.

email: **fang.katrina.xia@gmail.com**

BAYESIAN LATENT PROPENSITY SCORE APPROACH FOR AVERAGE CAUSAL EFFECT ESTIMATION ALLOWING COVARIATE MEASUREMENT ERROR

Elande Baro*, University of Maryland Baltimore County

Yi Huang, University of Maryland Baltimore County

Anindya Roy, University of Maryland Baltimore County

In observational studies, it is often the case that covariates are measured with error. The naive approach is to ignore the error and use naive propensity score methods with observed covariates to estimate the average causal effect (ACE). It has been shown that the naïve approach might bias the ACE inference. Dr. Yi Huang developed a set of causal assumptions allowing covariate measurement error and extend the standard propensity scoring theory (without measurement error) by proving the consistency of ACE estimation using the proposed latent propensity scores. She proposed a joint likelihood approach in finite mixture model format for ACE estimation under continuous outcome. In Huang and etc paper, EM algorithm is used, where the numerical performance is not ideal due to the large dimensions of unknown parameters. We extend this work and use Bayesian estimation method under the latent propensity score model in finite mixture model format. The method captures the uncertainty in propensity score subclassification arising from the unobserved measurement error. Simulations studies are presented to show the performance of this newly developed Bayesian approach compared to the existing EM algorithm and naive approach ignoring the error. It shows that Bayesian method provides more stable inference with good standard error estimates than EM. In addition, we investigate the case of ACE estimation under binary outcome and discuss its identifiability. This is a joint work with Dr Yi Huang and Dr Anindya Roy.

email: **baroel1@umbc.edu**

COMPARATIVE PERFORMANCE OF MULTIVARIATE MATCHING METHODS THAT SELECT A SUBSET OF OBSERVATIONS

Maria de los Angeles Resa*, Columbia University

Jose R. Zubizarreta, Columbia University

This paper presents a Monte Carlo simulation study of the comparative performance of multivariate matching methods that select a subset of observations from typically larger samples of treated and controls. The methods considered are the widespread method of greedy nearest neighbor matching with propensity score calipers, optimal matching of an optimally chosen subset, and the recent method of optimal cardinality matching. The main findings are: (i) covariate balance, as measured by differences in means, variance ratio, Kolmogorov-Smirnov distance, and the cross-match test statis- tic, is better with cardinality matching as by design it satisfies balance requirements; (ii) for a given level of balance, the resulting sample sizes are larger with cardinal- ity matching than with the other methods; (iii) in terms of distances, optimal subset matching performs best; (iv) estimates from cardinality matching have lower RMSEs, provided tight requirements for balance; (v) specifically, matching with fine balance for all the covariates plus strong requirements for mean balance have the lowest RM- SEs. In statistical practice, an extensively used rule of thumb is to balance covariates so that their absolute standardized differences in means are not greater than 0.1. The simulation results suggest that stronger forms of balance should be pursued in practice.

email: **maria@stat.columbia.edu**

IMPROVING TREATMENT EFFECT ESTIMATION IN THE PRESENCE OF TREATMENT DELAY THROUGH TRIPLET MATCHING

Erinn M. Hade*, The Ohio State University

Bo Lu, The Ohio State University

Hong Zhu, University of Texas Southwestern Medical Center

In health related studies with a longitudinal cohort, it is common that patients initiate treatment at different time points. In observational studies, multiple factors may contribute to why treatment is not administered at the desired time. Patients who are delayed in receiving treatment are often substantially different from those who receive timely treatment. Therefore, ignoring information on treatment delay may lead to biased estimation of treatment effects. To take advantage of this information, we propose to estimate the intended effect of having treatment on time, versus delayed treatment, versus never being treated. Balancing scores are first created to summarize the covariates information related to treatment initiation. Using these estimated scores, we will create matched groups of three observations (triplets with one observation from each of the treatment groups) and compare the treatment effect between groups. Further, we compare different matching algorithms to evaluate the matching quality. We apply these methods to data investigating the timing of adjuvant surgery for breast cancer.

email: **hade.2@osu.edu**

47. CONTRIBUTED PAPERS: Covariates Measured with Error

LOCALLY EFFICIENT SEMIPARAMETRIC ESTIMATORS FOR PROPORTIONAL HAZARDS MODELS WITH MEASUREMENT ERROR

Yuhang Xu*, Iowa State University

Yehua Li, Iowa State University

Xiao Song, University of Georgia

We propose a new class of semiparametric estimators for proportional hazards models in presence of measurement error in the covariates, where the baseline hazard function, the hazard function for the censoring time, and the distribution of the true covariates are deemed as unknown infinite dimensional parameters. We estimate the model components by solving a system of estimating equations based on the semiparametric efficient scores under a sequence of restricted models where the logarithm of the hazard functions are approximated by reduced rank regression splines. By slowly increasing the rank of the restricted model with the sample size, we show that the proposed estimators are locally efficient in the sense that the estimators are semiparametrically efficient if the distribution of the error-prone covariates is specified correctly and are still consistent and asymptotic normal if this distribution is misspecified. Our simulation studies show that the proposed estimators have smaller variances than the competing methods. We further illustrate the new method with a real application in an HIV clinical trial.

email: **yuhangxu@iastate.edu**

SEPARATING VARIABILITY IN PRACTICE PATTERNS FROM STATISTICAL ERROR; AN OPPORTUNITY FOR QUALITY IMPROVEMENT

Laine Thomas*, Duke University

Phillip J. Schulte, Duke University

Quality improvement studies seek to establish the degree of variability in practice patterns and outcomes across different providers. Wide variation suggests that institutional factors play a role in affecting outcomes, and high performing institutions should be studied and emulated. Therefore, the magnitude of variation is a key parameter of interest. Despite the extensive literature on methods for hospital monitoring and profiling, variability across providers is usually displayed in figures and histograms using techniques that either over-estimate or under-estimate the actual degree of variation. As a result, conclusions regarding the extent of variation based on these figures may be wrong. A simple correction can be derived from the hierarchical model, but is rarely used in medical literature. Although beneficial in many cases, this relies on the hierarchical model assumptions, such as normally distributed variation across providers. When features of the distribution, such as bi-modality or skewness, are of particular interest this approach is not adequate. To achieve greater flexibility, we cast this as a measurement error problem and apply recently developed methods for density estimation in the presence of measurement error. We compare the alternative approaches by simulation and interpret the results in the context of a motivating example.

email: **leellio2@gmail.com**

GOODNESS-OF-FIT TESTING OF ERROR DISTRIBUTION IN LINEAR ERRORS-IN-VARIABLES MODEL

Xiaoqing Zhu*, Michigan State University

The paper discusses a goodness-of-fit test for the error density function in linear error-in-variables regression models using the deconvolution kernel density estimator. The test statistic is an analog of the Bickel and Rosenblatt type of statistics, which is the square integrated error of the deconvolution kernel estimator and a smoothed version of the parametric fit of the density. Under the null hypothesis, the asymptotic distribution of the proposed test statistic is derived for the ordinary smooth and supersmooth deconvolution problems. A simulation study also shows the efficiency of this test.

email: **zhuxiaoq@msu.edu**

ESTIMATING RECURRENCE AND INCIDENCE OF PRETERM BIRTH IN CONSECUTIVE PREGNANCIES SUBJECT TO MEASUREMENT ERROR IN GESTATION: A NOVEL APPLICATION OF HIDDEN MARKOV MODELS

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

Prediction of preterm birth as well as characterizing the etiological factors affecting both the recurrence and incidence of preterm birth are important problems in obstetrics. The NICHD consecutive pregnancy study (CPS) recently examined this question by collecting data on a cohort of women with at least two pregnancies over a fixed time interval of eight years. Unfortunately, measurement error due to the dating required for calculating gestational age may misclassify preterm births and bias results obtained with standard longitudinal techniques. This article proposes a flexible approach that accounts for measurement error in gestational age when making inference. We propose a hidden Markov modeling approach that allows for measurement error in gestational age by exploiting the relationship between gestational age and birthweight. We use this novel methodology to estimate the effect of important covariates on the risk of pre-term birth in repeated pregnancies, focusing on the incidence and recurrence of preterm birth in the CPS cohort.

email: **albertp@mail.nih.gov**

MULTI-STATE MODEL WITH MISSING CONTINUOUS COVARIATE

Wenjie Lou*, University of Kentucky

Richard J. Kryscio, University of Kentucky

Erin Abner, University of Kentucky

Multi-state models are very useful tools to model chronic disease processes in which patients might go through several difference states (e.g., preclinical, mild disease, severe disease, death). For example, in studies of dementia patients might experience an intermittent state called mild cognitive impairment (MCI) before they become demented or die. The common way to account for patient characteristics in the disease process in the multi-state model is to add covariates in the transition intensity functions. In most applications, covariates have to be fully observed; yet clinical data are almost always incomplete in practice. In this paper, we propose a maximum simulated likelihood method to handle the missing continuous covariates in multi-state models. Our simulation study shows that the proposed method works quite well in most MAR cases. We also apply the method to a real dataset, a longitudinal dementia study cohort of 5,404 subjects.

email: **wenjie.lou@uky.edu**

WEIGHTED L1-PENALIZED CORRECTED QUANTILE REGRESSION FOR HIGH DIMENSIONAL MEASUREMENT ERROR MODELS

Abhishek Kaul*, Michigan State University

Hira L. Koul, Michigan State University

Standard formulations of prediction problems in high dimension regression models assume the availability of fully observed covariates and sub-Gaussian and homogenous model errors. This makes these methods inapplicable to measurement errors models where covariates are unobservable and observations are possibly non sub-Gaussian and heterogeneous. We propose a weighted penalized corrected quantile estimator for regression parameters in linear regression models with additive measurement errors, where unobservable covariate is nonrandom. The proposed estimators forgo the need for the above mentioned model assumptions. We study

these estimators in a high dimensional sparse setup where the dimensionality can grow exponentially with the sample size. We provide bounds for the statistical error associated with the estimation, that hold with asymptotic probability 1, thereby providing the ?1-consistency of the proposed estimator. We also establish the model selection consistency in terms of the correctly estimated zero components of the parameter vector. A simulation study that investigates the finite sample accuracy of the proposed estimator is also included in the paper.

email: **kaulabhi@stt.msu.edu**

48. ORAL POSTERS: Clinical Trials

48a. SPLIT-SAMPLE BASED AND MULTIPLE IMPUTATION ESTIMATION AND COMPUTATION METHODS FOR META-ANALYSIS OF CLINICAL TRIAL DATA AND OTHERWISE HIERARCHICAL DATA

Geert Molenbergs*, Universiteit Hasselt

Geert Verbeke, Katholieke Universiteit Leuven

Michael G. Kenward, London School of Hygiene and Tropical Medicine

Wim Van der Elst, Universiteit Hasselt

Lisa Hermans, Universiteit Hasselt

Vahid Nassiri, Katholieke Universiteit Leuven

Analyzing hierarchical data can be challenging. Such data occur in longitudinal and/or multi-centric trials, regular meta-analyses, surrogate endpoint evaluation (Burzykowski, Molenberghs, and Buyse 2005), etc. Further, these issues occur for a wide variety of settings: small samples (e.g., orphan diseases) on the one hand and big data on the other. In between, we have meta-analyses based on just a few large trials. Similar issues occur in small-area epidemiology. It has been found empirically that lack of balance is an important, though not the only, contributor to the difficulties. Another aspect is the lack of so-called complete sufficient statistics (Molenberghs et al 2014). A broad paradigm is to use the following three-step approach: (a) apply a method to render the data balanced; (b) analyzed the so-obtained data; (c) apply combination rule to render a single, valid inference. Two viable candidates for step (a) are: (a1) multiple imputation (Little and Rubin 2002, Carpenter and Kenward 2013) and (a2) so-called split sampling, a pseudo-likelihood based approach. Method (a2) provides a formal basis for such methods as a two-stage approach for linear mixed models, a fixed-effects approach for meta-analyses, etc. Under (a1), step (c) is the classical combination rules; under (a2) an information-sandwich estimator is used.

email: **geert.molenberghs@uhasselt.be**

48b. OVER-PARAMETERIZATION IN ADAPTIVE DOSE-FINDING STUDIES

John O'Quigley, Universite Pierre et Marie Curie

Nolan A. Wages, University of Virginia

Mark R. Conaway, University of Virginia

Ken Cheung, Columbia University

Ying Yuan, University of Texas MD Anderson Cancer Center

Alexia Iasonos*, Memorial Sloan Kettering Cancer Center

Adaptive, model-based, dose-finding methods, such as the continual reassessment method, have been shown to have good operating characteristics. One school of thought argues in favor of the use of parsimonious models using a strict minimum number of parameters. In particular, for the standard situation of a single homogeneous group, the usual approach is to appeal to a one-parameter model. Other authors argue that richer models lead to improved performance. Here, we show that increasing the dimension of the parameter space, in the context of adaptive dose-finding studies, is usually counter-productive and, rather than leading to improvements in operating characteristics, the added dimensionality is likely to result in problems. Among these are inconsistencies of sample estimates, lack of coherency of escalation or de-escalation, erratic behavior, getting stuck at the wrong level and, in general, poorer performance in terms of correct identification of the targeted dose. Our conclusions are based on both theoretical results and simulations.

email: **john.oquigley@upmc.fr**

48c. IMPROVING SOME CLINICAL TRIALS INFERENCE BY USING RANKED AXILLARY COVARIATE

Hani Samawi*, Georgia Southern University

Rajai Jabrah, Georgia Southern University

Robert Vogel, Georgia Southern University

Daniel Linder, Georgia Southern University

The main objective in a randomized clinical trial of studies such as in cancer, AIDS, etc. is to compare the outcome of interest between two or more groups. Clinical trials are considered the "gold standard" of biomedical research and of its strengths are the ability to measure changes and/or evaluate of treatments over time with maximizing power of statistics and validity. Clinical trials are expensive, and the cost of clinical trials on medical treatments and devices, public health investigators are increasing with each phase and continue to escalate, especially in phase III. The idea proposed in this project is to use auxiliary covariates by adopting Ranked Set Sampling (RSS) technique to select the subjects for each treatment-arms, to utilize inexpensive auxiliary covariates information into a randomized clinical trials. The goal is to provide a more precise estimator of the population mean of the outcome of interest to recover the difficult to obtain information, without making any additional assumptions other than those already necessary for (RSS) and the ordinary least square estimators from a regression model to hold.

email: **samawi.hani2@gmail.com**

48d. DIRECT ESTIMATION OF THE MEAN OUTCOME ON TREATMENT WHEN TREATMENT ASSIGNMENT AND DISCONTINUATION COMPETE

Xin Lu*, Emory University

Brent A. Johnson, University of Rochester

Several authors have investigated the challenges of statistical analyses and inference amidst early treatment termination, including a loss of efficiency in randomized controlled trials and its connection to dynamic regimes in observational studies. Popular estimation strategies for causal estimands in dynamic regimes lend themselves to studies where treatment is assigned at a finite number of points; the extension to continuous treatment assignment is non-trivial and introduces other caveats. We re-examine this particular problem from a different perspective and propose a new direct estimator for the mean outcome of a target treatment length policy that does not model the propensity score. Because this strategy does not include a model for treatment selection, the estimator works well in both discrete and continuous time and avoids finite sample bias associated with squeezing continuous time data into intervals. We show how the competition of treatment assignment and terminating event through time leads to an intriguing type of competing risks problem. We exemplify the direct estimator through small sample numerical studies and the analyses of two real datasets. When all model assumptions are correct, our simulation studies show that the direct estimator is more precise than inverse weighted estimator.

email: **xlu28@emory.edu**

48e. BAYESIAN INTERIM ANALYSIS METHODS FOR PHASE IB EXPANSION TRIALS ENABLE EARLIER GO/NO-GO DECISIONS IN ONCOLOGY DRUG DEVELOPMENT

James Lymp*, Genentech

Jane Fridlyand, Genentech

Hsin-Ju Hsieh, Genentech

Daniel Sabanes Bove, F. Hoffmann-La Roche

Somnath Sarkar, F. Hoffmann-La Roche

Phase Ib expansion trials are often used in oncology drug development to obtain preliminary safety and efficacy information for making decisions in the next phase of development. Such trials are typically single arm and depend on comparison to reliable external information which can be challenging to obtain, especially in a combination setting. We describe two Bayesian approaches to interim analysis that enable earlier decisions based on a binary primary efficacy endpoint. The posterior probability approach is based on the probability that the new therapy is effective using the current evidence at the time of interim analysis. The predictive probability approach is based on the probability that the trial would conclude that the new therapy is effective if carried out to completion. We motivate each of these methods, illustrate their use and interpretation, and demonstrate the impact on decisions of various parameters including the prior distribution. Simulations show that operating characteristics, such as the probability of making a correct

decision, can be effectively controlled. Bayesian methods for interim analysis can help earlier decision making for a drug development program because they are flexible with the number and timing of interim analyses and naturally incorporate historical and concurrent controls.

email: **lymp.james@gene.com**

48f. UNIFIED ADDITIONAL REQUIREMENT IN CONSIDERATION OF REGIONAL APPROVAL FOR MULTI-REGIONAL CLINICAL TRIALS

Zhaoyang Teng*, Boston University

Yeh-Fong Chen, The George Washington University

Mark Chang, AMAG Pharmaceuticals and Boston University

To speed up the process of bringing a new drug to the market, more and more clinical trials are conducted simultaneously in multiple regions. After demonstrating the overall drug's efficacy across regions, the regulatory and drug sponsor may also want to assess the drug's effect in specific region(s). Most of the recent approaches imposed a common criterion to assess the consistency of treatment effects between the interested region(s) and the entire study population regardless the number of regions included in a Multi-Regional Clinical Trials (MRCT). As a result, the needed sample size to achieve the desired probability of satisfying the regional requirement could be huge and implausible for the trial sponsors to implement. In this paper, we propose a unified additional requirement for regional approval by considering the parameters in the additional requirement depending on the number of regions. In particular, the values of parameters are determined by considering a reasonable sample size increase with the desired probability satisfying the additional requirement. Considering the practicality of the global trial or sample size increase, we suggest the values of the parameters for different number of regions. We also introduce the assurance probability curve to evaluate the performance of different regional requirements.

email: **tzy0614@bu.edu**

48g. EFFICIENCIES OF BAYESIAN ADAPTIVE PLATFORM CLINICAL TRIALS

Ben Saville*, Berry Consultants

Scott Berry, Berry Consultants

A "platform trial" is a clinical trial in which multiple treatments for the same indication are tested simultaneously. Bayesian adaptive platform designs offer attractive features such as dropping treatments for futility, declaring one or more treatments superior, or adding new treatments to be tested during the course of a trial. Such designs can be more efficient at finding beneficial treatments relative to traditional two group designs. We quantify these efficiencies via simulation to show that platform trials on average can find beneficial treatments with fewer patients, fewer deaths or poor outcomes, less time, and with greater probabilities of success than traditional designs.

email: **ben@berryconsultants.com**

48h. A BAYESIAN SEMIPARAMETRIC MODEL FOR INTERVAL CENSORED DATA WITH MONOTONE SPLINES

Bin Zhang, Cincinnati Children's Hospital Medical Center

Yue Zhang*, University of Cincinnati

Generalized odds ratio hazard (GOPH) models are general enough to include several commonly used models, such as proportional hazards model and proportional odds model. However, its application is much undeveloped to interval censored data in which the partial likelihood does not exist. In this paper, we propose a novel Bayesian approach to analyze the interval censored data with GOPH model. The baseline cumulative hazard function was modelled with finite dimensional monotone splines. Gibbs sampler is easy to implement and the performance of the proposed method was evaluated by extensive simulation studies. A real life data set was analyzed by using the aforementioned method as an illustration.

email: **zhang3ye@mail.uc.edu**

48i. COMPREHENSIVE EVALUATION OF ADAPTIVE DESIGNS FOR PHASE I ONCOLOGY CLINICAL TRIALS

Sheau-Chiann Chen*, Vanderbilt University

Yunchan Chi, National Cheng Kung University

Yu Shyr, Vanderbilt University

To provide valuable recommendations for selecting a phase I trial design, a comprehensive evaluation measure is proposed.

This measure is used to evaluate the performance of the 3+3 design and three Bayesian adaptive designs, i.e., the continual reassessment method design (CRM), Bayesian continual reassessment method design (B-CRM), and modified toxicity probability intervals design (mTPI). A simulation study determined that none of the designs are uniformly better than the others based on overall scores. We use the gap size, which is the distance between the true toxicity probability of MTD and the true toxicity probability of the next higher dose, to provide some insight into comparison results. When both target toxicity rate and gap size increase, the performance of the 3+3 and mTPI designs tend to be better than the CRM design. In practice, since the investigators may have some knowledge about target toxicity rate, the CRM design is best applied for a very small target toxicity rate, while the mTPI design is recommended for use for a target toxicity rate between 0.1 and 0.35. When the target toxicity rate and toxicity probability patterns are unknown, the mTPI design is recommended.

email:
sheau-chiann.chen@vanderbilt.edu

48j. STATISTICAL INFERENCE FOR COMPOSITE OUTCOMES BASED ON PRIORITIZED COMPONENTS

Ionut Bebu*, The George Washington University

John M. Lachin, The George Washington University

Composite endpoints are common in cardiovascular (CV) trials, and the time-to-first-event analysis is the standard approach for testing the treatment effect. This approach treats all individual components as of equal relevance, although they may not be of equal importance clinically or to the patients. To address this issue, several authors have proposed to rank the individual outcomes based on their importance, and then to combine them based on their ranks. Two such approaches include the proportion in favor of treatment parameter and the win ratio parameter. This talk will describe tests and confidence intervals for composite outcomes based on prioritized components using the large sample distribution of certain multivariate multi-sample U-statistics. This nonparametric approach provides a general solution for both the proportion in favor of treatment parameter and the win ratio parameter, and it can be extended to stratified studies, and the comparison of more than two groups. The proposed results are illustrated using data from the Prevention of Events with Angiotensin Converting Enzyme Inhibition (PEACE) Trial.

email: **iobebu@gmail.com**

48k. THE IMPACT OF COVARIATE MISCLASSIFICATION USING GENERALIZED LINEAR REGRESSION UNDER COVARIATE-ADAPTIVE RANDOMIZATION

Liqiong Fan*, Medical University of South Carolina

Sharon D. Yeatts, Medical University of South Carolina

Background: Covariate misclassification may lead to bias in treatment effect estimate, impact the power and type I error, and affect trial validity and reliability. When covariate adaptive randomization is used to control the imbalance between treatment arms, misclassification may also impact the intended treatment assignment. It is unclear whether the appropriate analysis strategy should adjust for the misclassified covariate or the corrected covariate. We provide computational simulation results and asymptotic result to explore the impact of such misclassification on the statistical operating characteristics. Methods: Binary/frequency outcome were simulated with treatment and one error-prone dichotomized prognostic covariate. Randomization schemes were stratified within the covariate. Simulation scenarios were created based on the misclassification rate and the covariate effect on the outcome. Models including unadjusted, adjusted by the misclassified covariate, adjusted by the corrected covariate were compared. Result: Under covariate-adaptive randomization with logistic regression, type I error can be maintained in the adjusted model either with the misclassified covariate or the corrected covariate. Randomization procedure does not have additional impact on power loss and bias caused by covariate misclassification. The magnitude of power loss and bias depends on the covariate effect on the outcome and the misclassification rate. With poisson log linear model, type I error is inflated with misclassified model. Conclusion: Correction for covariate misclassification should be taken into consideration during trial design and later analysis.

email: **fanliq@musc.edu**

48l. NON-INFERIORITY TEST BASED ON TRANSFORMATIONS

Santu Ghosh*, Wayne State University

Arpita Chatterjee, Georgia Southern University

Samiran Ghosh, Wayne State University

Non-inferiority trials are becoming very popular for comparative effectiveness research. These trials are required to show that an experimental drug is not inferior than a known reference drug by a small pre-specified amount. Hence, non-inferiority trials are of great importance to pharmaceutical companies, when superiority can not be claimed. In this paper we consider a three-arm non-inferiority trial consists of a placebo, areference treatment, and an experimental treatment. However unlike the traditional choices, we assume that the distributions of the end points corresponding to these treatments are unknown and suggest a test procedure for a three non-inferiority trial based on transformations in conjunction with a normal approximation. Theoretical properties of our method are investigated. An alternative test procedure based the bootstrap percentile-t method is discussed. We compare the performance of these test procedures in simulated data sets. These methods are further illustrated in a study on mildly asthmatic patients.

email: **saghos@med.wayne.edu**

48m. METHODS ACCOUNTING FOR MORTALITY AND MISSING DATA IN RANDOMIZED TRIALS WITH LONGITUDINAL OUTCOMES

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

Chenguang Wang, Johns Hopkins School of Medicine

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

Randomized trials are the standard for establishing evidence favoring a treatment over standard care in clinical settings. We will consider randomized trials where the outcome is a clinical characteristic of the patient and is measured at a single time or multiple fixed times after randomization. Defining and estimating the treatment effect is complicated when the population being studied is expected to have high mortality rates during the trial due to either the age of the patients or the underlying conditions for which the patients are being treated. In such trials, clinical outcomes that would be measured are “truncated due to death”. In addition, among survivors, there is often missing data. Using data from a randomized trial among patients with acute respiratory distress syndrome (ARDS), we describe and demonstrate methods for accounting for mortality and missing data within randomized trials with special attention to sensitivity analysis to the untestable assumptions required in the analyses.

email: **elizabethcolantuoni@gmail.com**

48n. A SEMIPARAMETRIC BAYESIAN APPROACH USING HISTORICAL CONTROL DATA FOR ASSESSING NON-INFERIORITY IN THREE ARM TRIALS

Arpita Chatterjee*, Georgia Southern University

Santu Ghosh, Wayne State University

Samiran Ghosh, Wayne State University

Historical information is always relevant for designing clinical trials. The incorporation of historical information in the new trial can be very beneficial. Some of these benefits include reduction of effective sample size, increase the statistical power, and reduction of cost and ethical hazard. However, if current data conflicts with historical data, borrowing information from historical data can give misleading results. In this project we consider a semiparametric Bayesian approach based on Dirichlet Process prior that can borrow relevant information from historical data for assessing non-inferiority in three-arm trials. The scale parameter of Dirichlet Process prior can be treated as a tuning parameter which can control the dependencies between the historical and current data. A simulation study is developed to demonstrate that our suggested method can be successfully applied. Finally we apply the proposed methodology to a real data set.

email:
achatterjee@georgiasouthern.edu

48o. DESIGN PARAMETERS AND EFFECT OF THE DELAYED-START DESIGN IN ALZHEIMER'S DISEASE

Guoqiao Wang*, University of Alabama, Birmingham

Richard E. Kennedy, University of Alabama, Birmingham

Lon S. Schneider, University of Southern California

Gary R. Cutter, University of Alabama, Birmingham

Purpose and methods. The delayed-start design, in which patients are randomly assigned to placebo or treatment for a pre-specified frame of time and then those (or a randomized portion of those) in the placebo group are also given the treatment, was recommended for AD. Critical design parameters such as sample size and the timing of treatment switch have been proposed, the purpose of this study is to extend the existing theory and to verify those design parameters through simulation based on a meta-database of previous trials in AD. Conclusion. When only a randomized portion of the patients originally on placebo late will receive the treatment, the optimal sample size allocation ratio between the treatment group, the continuing placebo group, and the delayed-start treatment group is 1:1:1. The weight on estimators from the 3 groups depends on the correlation among the slopes in the delayed-start group; however, the correlation is relatively small. We also verified that the optimal time of treatment switch is in the middle of the study for evenly spaced measurements. We will also provide power estimation for given sample sizes.

email: **guoqiao@uab.edu**

49. CENS Invited Session – Careers in Statistics: Skills for Success

HOW TO BE SUCCESSFUL IN ORAL AND WRITTEN COMMUNICATIONS AS A BIOSTATISTICIAN

Peter Grant Mesenbrink*, Novartis Pharmaceuticals Corporation

As statisticians, often it is required that technical information needs to be explained to non-statisticians both as oral and written communications. The variability of knowledge of biostatistics for the audience receiving the information is often quite large. Thus, it is often required for biostatisticians to be able to adapt their style of communication while still providing key scientific and strategic input to projects in which they are involved. Best practices for excelling in oral and written communicators across different disciplines will be discussed when giving formal presentations as well as how to address questions and their answers when communicating with audiences who were not formally trained as biostatisticians will be discussed.

email: **peter.mesenbrink@novartis.com**

NAVIGATING THE ACADEMIC JUNGLE WITHOUT GOING BANANAS

Amy H. Herring*, University of North Carolina, Chapel Hill

We consider survival strategies for the academic jungle. Some individuals appear to exibit immunity from hazards, occupational or otherwise. However, we argue the feeling that one has escaped the dreaded jaguar only to disappear into an unanticipated sinkhole while dealing with a pesky bot fly, all the while entertaining a group of howler monkeys, is much more normative. We'll discuss risks, perceived and real, and share strategies employed by successful species navigating jungles of all types. As with any journey, a map is often critical to successful navigation, and we will conclude by discussing useful features of academic cartography.

email: **aherring@bios.unc.edu**

WHAT AM I GOINGTOBE WHEN I GROW UP? EVOLVING AS A STATISTICIAN

Nancy L. Geller*, National Heart, Lung and Blood Institute, National Institutes of Health

The speaker will describe her own professional evolution, both intellectually (to a clinical trials biostatistician) and in terms of leadership (to director of a biostatistics group and former president of the American Statistical Association). Recognizing that we all will have setbacks and developing the resiliance to continue pursuing your goals despite setbacks is important. Finding the right mentor(s) is very helpful, especially when your own path is not

crystal clear. Developing a "yes I can" attitude, especially when opportunities arise to do something you have never done before, is a big plus. With many opportunities for intellectual pursuits and leadership, individuals in our field can have a rich and satisfying career.

email: **ng@helix.nih.gov**

50. Analysis Methods for Data Obtained from Electronic Health Records

IMPROVING THE POWER OF GENETIC ASSOCIATION TESTS WITH IMPERFECT PHENOTYPE DERIVED FROM ELECTRONIC MEDICAL RECORDS

Jennifer A. Sinnott*, Harvard School of Public Health

Wei Dai, Harvard School of Public Health

Katherine P. Liao, Brigham and Women's Hospital

Elizabeth W. Karlson, Brigham and Women's Hospital

Isaac Kohane, Harvard Medical School

Robert Plenge, Merck Research Laboratories

Tianxi Cai, Harvard School of Public Health

To reduce costs and improve clinical relevance of genetic studies, such studies could be performed in hospital-based cohorts by linking phenotypes extracted from electronic medical records (EMRs) to genotypes assessed in routinely collected medical samples. It can be difficult to extract accurate information about disease outcomes from large numbers of EMRs, but recently numerous algorithms have been developed to infer phenotypes. Although these algorithms are quite accurate, they typically do not provide perfect classification due to the difficulty in inferring meaning from the text. Some algorithms can produce for each patient a probability that the patient is a disease case, which can be thresholded to define case-control status, and this estimated case-control status has been used to replicate known genetic associations in EMR-based studies. However, using the estimated disease status in place of true disease status results in outcome misclassification, which can diminish test power and bias odds ratio estimates. We propose to instead directly model the algorithm-derived probability of being a case. We demonstrate how our approach improves test power and effect estimation. Our work provides an easily implemented solution to a major practical challenge that arises in the use of EMR data.

email: **jas953@mail.harvard.edu**

NONPARAMETRIC ESTIMATION OF PATIENT PROGNOSIS WITH APPLICATION TO ELECTRONIC HEALTH RECORDS

Patrick J. Heagerty*, University of Washington

Alison E. Kosel, University of Washington

We develop an algorithm and associated inference for creating local patient outcome predictions. The intended goal of the methods is to provide an estimate of the full outcome distribution for a given subject by providing summary data for a specific axis-parallel neighborhood with a fixed subset size. We develop inference for the local predictions and implement the methods using a dynamic computational interface. We illustrate the methods with a large electronic health records based back pain cohort, and comment on extensions of the methods to comparative estimation.

email: **heagerty@uw.edu**

MINING EHR NARRATIVES FOR CLINICAL RESEARCH

Enedia Mendonca*, University of Wisconsin, Madison

email: **emendonca@biostat.wisc.edu**

51. Statistical Challenges of Survey and Surveillance Data in US Government

USING VENUE-BASED SAMPLING TO RECRUIT HARD-TO-REACH POPULATIONS

Maria Corazon B. Mendoza*, Centers for Disease Control and Prevention

Chris Johnson, Centers for Disease Control and Prevention

Brooke Hoots, Centers for Disease Control and Prevention

Teresa Finlayson, Centers for Disease Control and Prevention

Courses in the design of sample surveys typically cover several types of sampling designs that are used to survey a population of interest. Examples in these courses assume a sampling frame is readily identifiable. But what happens if no easily identifiable sampling frame exists? This talk explores this question by introducing one of the sampling designs that the Centers for Disease Control and Prevention uses to survey hard-to-reach populations: venue-based sampling (VBS). Using the National HIV Behavioral Surveillance Study System as an example, VBS will be described and the logistics of creating the sampling frame and of sampling individuals will be discussed. In addition, issues such as obtaining accurate counts (estimates of individuals attending an event) and multiplicity (an individual belonging to more than one sampling unit), will be introduced and examined because of their impact on point and variance estimation and with respect to how they can be used to study the effectiveness of the sampling design. Participant counts at different stages of the sampling procedure will be shown to illustrate how the final sample is achieved and to give a sense of the potential yield of individuals in a VBS design.

email: **wyx1@cdc.gov**

DEVELOPMENT OF GUIDELINES FOR THE PRESENTATION OF DATA FROM THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

Margaret Devers Carroll*, National Health and Nutrition Examination Survey, Centers for Disease Control and Prevention

Using objectively measured National Health and Nutrition Examination Surveys (NHANES) data, the prevalence of important health characteristics can be estimated, e.g. the percent of obese adults ages 20 years and over. Since 1999 data from NHANES has been released in two year cycles. Similar to previous NHANES surveys the sampling plan for each 2 year survey follows a highly stratified, multistage probability design which involves the selection of primary sampling units (PSUs, counties or groups of contiguous counties), segments (groups of dwelling units) within PSUs, dwelling units within segments and sample persons within dwelling units. Although the sample size of each 2 year cycle is approximately 10,000, the number of first stage units is much smaller (at most 17). Consequently design based estimates of variance can be extremely unstable. Furthermore, the complex sample design can impact variance estimates substantially. Therefore, there is a need to develop guidelines for the presentation of NHANES data. In this talk, statistics used in formulating the most recent set of proposed guidelines, including degrees of freedom, design effect and relative confidence interval width, will be discussed.

email: **mdc3@cdc.gov**

DATA SWAPPING METHODS FOR STATISTICAL DISCLOSURE LIMITATION

Guangyu Zhang*, National Center for Health Statistics, Centers for Disease Control and Prevention

Joe Fred Gonzalez, National Center for Health Statistics, Centers for Disease Control and Prevention

Anna Oganyan, National Center for Health Statistics, Centers for Disease Control and Prevention

Alena Maze, National Center for Health Statistics, Centers for Disease Control and Prevention

Protection of confidentiality and privacy of survey participants' (individuals or establishments) data is of primary importance to federal agencies when releasing microdata to the public. Records that have unique combinations of key variables are particularly vulnerable to disclosure. Statistical disclosure limitation techniques, such as random data swapping and recoding, have been used as disclosure protection strategies. To protect confidentiality and to maintain the accuracy of statistical inferences, in this study we

create clusters of subjects by Euclidean distances of observed variables and apply random data swapping methods within homogenous clusters. We conduct simulations and apply our methods to a National Health Interview Survey (NHIS) public-use data set, available from the National Center for Health Statistics.

email: **VHA1@cdc.gov**

PRACTICAL APPROACHES TO DESIGN AND INFERENCE THROUGH THE INTEGRATION OF COMPLEX SURVEY DATA AND NON-SURVEY INFORMATION SOURCES

John L. Eltinge*, U.S. Bureau of Labor Statistics

Rachel M. Harter, RTI International

Practical sample survey work is currently encountering important opportunities and challenges arising from the increased availability of data from alternative (non-survey) sources. These sources potentially can provide very rich information that could be integrated with traditional sample survey processes through, e.g., targeting of subpopulations; enhancement of sample frames; improvement of unit contact and other dimensions of fieldwork; direct replacement of survey items that are especially burdensome, expensive or error-prone; and improvement of editing, imputation or weight construction. However, these non-survey data are subject to important quality issues, including unit coverage; item missingness; definitional and aggregation issues; use of proxy variables; imputation errors; and recording errors. This paper reviews these quality issues and suggests a unified framework for evaluation of error characteristics for both traditional survey data sources, and the abovementioned non-survey sources. This framework in turn suggests some practical methods for integration of these data sources. The primary concepts are illustrated with applications to two large-scale surveys.

email: **Eltinge.John@bls.gov**

52. Reconstructing the Genomic Landscape from High-Throughput Data

COPY NUMBERS IN CIRCULATING TUMOR CELLS (CTCs) USING DNA-Seq

Henrik Bengtsson*, University of California, San Francisco

Detecting and characterizing circulating tumor cells (CTCs) in the blood of cancer patients could bring additional and crucial understanding on how the cancers spread, how they circumvent therapy, and how to better target them. It has been shown that CTCs resemble genomic aberrations of the primary tumor, also when long time has passed, meaning they could also be used in patient follow ups. Different techniques have been proposed to isolate, capture and enrich CTCs, to remove contamination from non-malignant epithelial cells or leukocytes, and to amplify DNA obtained from these small counts of cells. Originally DNA copy-number (CN) microarrays was used for genomic profiling of CTCs but recently high-throughput DNA sequencing (DNA-Seq) has entered the arena. Due to the very limited amount of DNA available from each capture, which results in very low signal-to-noise ratios (SNR), it is critical to optimize the assay and the analytical procedures in order to succeed. We propose statistical methods for quantitatively assessing the performance of these different approaches. In essences these methods capture the effective SNR for a particular setup and its power to detect aberrations, which makes it possible to objectively decide which methods are better than others. They allow us to compare whether DNA-Seq or DNA microarrays are better suited for the challenge or not, what DNA-Seq coverage is needed for detecting events of allelic imbalance such as copy-neutral loss of heterozygosity (LOH), and more.

email: **hb@biostat.ucsf.edu**

DNA COPY NUMBER ANALYSES FOR FAMILY BASED DESIGNS

Ingo Ruczinski*, Johns Hopkins University

We present novel methods and software for DNA copy number analyses in family based designs with sequencing or array data. In the first example, we consider the evidence that a rare copy number variant may be causal when only a few affected subjects in a multiplex extended pedigree are sequenced or typed, by quantifying the probability of allele sharing by all affected relatives given it was seen in any one family member under the null hypothesis of complete absence of linkage and association. In the second example, we present a new method to infer de novo copy number variants in trios by defining “minimum distance” statistic to capture

differences in copy numbers between offspring and parents which reduces technical variation from probe effects and genomic waves, a major source of false positive identifications in copy number analyses. Following segmentation of the minimum distance by circular binary segmentation, final inference regarding de novo copy number events is based on a posterior calling step.

email: **ingo@jhu.edu**

RECONSTRUCTING 3-D GENOME CONFIGURATIONS: HOW AND WHY

Mark Robert Segal*, University of California, San Francisco

The three-dimensional (3D) configuration of chromosomes within the eukaryote nucleus is consequential for several cellular functions including gene expression regulation and is also strongly associated with cancer-causing translocation events. While visualization of such architecture remains limited to low resolutions (due to compaction, dynamics and scale), the ability to infer structures at high resolution has been enabled by recently-devised chromosome conformation capture (3C) techniques. In particular, when coupled with next generation sequencing, such methods yield an unbiased inventory of genome-wide chromatin interactions. Various algorithms have been advanced to operate on such data to produce reconstructed 3D configurations. Several studies have shown that such reconstructions provide added value over raw interaction data with respect to downstream biological insights. However, such added value has yet to be realized for higher eukaryotes since no genome-wide reconstructions have been inferred for these organisms because of computational bottlenecks and organismal complexity. Here we propose a two-stage algorithm, deploying multi-dimensional scaling, that overcomes these barriers. After showcasing 3D architectures for mouse embryonic stem cells and human lymphoblastoid cells we discuss methods for evaluating these solutions and, time permitting, downstream applications thereof.

email: **Mark.Segal@ucsf.edu**

A LATENT VARIABLE APPROACH FOR INTEGRATIVE CLUSTERING OF MULTIPLE GENOMIC DATA TYPES

Ronglai Shen*, Memorial Sloan-Kettering Cancer Center

Large-scale integrated cancer genome characterizationefforts including the cancer genome atlas have created unprecedented opportunities to study cancer biology in the context of knowing the entire catalog of genetic alterations. A clinically important challenge is to discover cancer subtypes and their molecular drivers in a comprehensive genetic context. I will present a latent variable framework for joint modeling of discrete and continuous data types that arise from integrated genomic, epigenomic, and transcriptomic profiling. We show application of the method to the TCGA pan-cancer cohort with whole-exome DNA sequencing, SNP6.0 array, mRNA sequencing data in 3,000 patient samples spanning 12 cancer types. In addition, I will introduce a topic on intratumor heterogeneity which is characterized by the presence of genetically and phenotypically distinct subclones of tumor cells. Genetic diversity within a tumor is increasingly recognized as a driver of rapid disease progression, resistance to targeted therapies, and poor survival outcome. I will present a statistical approach to reconstruct clonal composition from high-throughput sequencing of bulk tumor samples.

email: **rlshen@umich.edu**

53. Statistical Methods for Single Molecule Experiments

WALKING, SLIDING, AND DETACHING: TIME SERIES ANALYSIS FOR CELLULAR TRANSPORT IN AXONS

John Fricks*, The Pennsylvania State University

Jason Bernstein, The Pennsylvania State University

William Hancock, The Pennsylvania State University

Kinesin is a molecular motor that, along with dynein, moves cargo such as organelles and vesicles along microtubules through axons. Studying these transport process is vital, since non-functioning kinesin has been implicated in a number of neurodegenerative diseases, such as Alzheimer's disease. Over the last twenty years, these motors have been extensively studied through in vitro experiments of single molecular motors using laser traps and fluorescence techniques. However, an open challenge has been to explain in vivo behavior of these systems when incorporating the data from in vitro

experiments into straightforward models. In this talk, I will discuss recent work with my experimental collaborator, Will Hancock (Penn State), to understand more subtle behavior of a single kinesin than has previously been studied, such as sliding and detachment and how such behavior can contribute to our understanding of in vivo transport. Data from these experiments include time series taken from fluorescence experiments for kinesin. In particular, we will use novel applications of switching time series models to explain the shifts between different modes of transport.

email: **fricks@gmail.com**

ANALYZING SINGLE-MOLECULE PROTEIN-TARGETING EXPERIMENTS VIA HIERARCHICAL MODELS

Samuel Kou*, Harvard University

Yang Chen, Harvard University

Recent technological advances allow scientists to follow a biological process on a single-molecule basis. These advances also raise many interesting data-analysis problems. In this talk we will focus on recent single-molecule experiments on protein targeting. To maintain proper cellular function, proteins often need to be transported inside or out of a cell. The detailed molecular mechanism behind such a process (often referred to as protein targeting) is not well understood. Single-molecule experiments are designed to unveil the detailed mechanism and reveal the functions of different organelles involved in the process. The experimental data consist of hundreds of stochastic time traces (from the fluorescence recording of the experimental system). We introduce a Bayesian hierarchical model on top of a hidden Markov model (HMM) to analyze these data and use the statistical results to answer the biological questions. We will discuss model selection, the construction of the hierarchical model, their biological meaning as well as our new understanding of the detailed mechanism behind protein transportation.

email: **kou@stat.harvard.edu**

BIMOLECULAR REACTION, DATA TYPES, AND AN ALTERNATIVE MODEL TO THE SMOLUCHOWSKI THEORY

Hong Qian*, University of Washington

Bimolecular reaction is fundamental in biology and biochemistry; it is crucial in all forms of life. The classical description of a bimolecular reaction assumes that the reaction is largely driven by the Brownian motion of the two reactant molecules. The rate of bimolecular reaction, fundamental to the understanding of biological processes, is approximated by a classical diffusion theory, due to M. Smoluchowski (1917). We recently identify that, the key issue, unlike in the classical theory, is intimately related to what type of data is collected for quantifying a bimolecular reaction. An alternative model, based on coupled diffusion with Markov switching, will be provided. We will discuss its connection with an experimental data collection technique, fluorescence correlation spectroscopy (FCS), and the application on nonlinear reversible bimolecular reaction

e-mail: **hqian@u.washington.edu**

HIDDEN MARKOV MODELS WITH APPLICATIONS IN CELL ADHESION EXPERIMENTS

Jeff C. F. Wu, Georgia Institute of Technology

Ying Hung*, Rutgers University

Cell adhesion experiments refer to biomechanical experiments that study protein, DNA, and RNA at the level of single molecules. The study of cell adhesion plays a key role in many physiological and pathological processes, especially in tumor metastasis in cancer research. Motivated by the analysis of a specific type of cell adhesion experiments, a new framework based on hidden Markov model is proposed. A double penalized order selection procedure is introduced and shown to be consistent in estimating the number of hidden states in hidden Markov models. Simulations show that the proposed framework outperforms existing methods. Applications of the proposed methodology to real data demonstrate the accuracy of estimating receptor-ligand bond lifetimes and waiting times, which are essential in kinetic parameter estimation.

email: **yhung@stat.rutgers.edu**

54. Subgroup Analysis and Adaptive Trials

A BAYES RULE FOR SUBGROUP REPORTING – BAYESIAN ADAPTIVE ENRICHMENT DESIGNS

Peter Mueller*, University of Texas, Austin

We discuss Bayesian inference for subgroups in clinical trials. We start with a decision theoretic approach, based on a straightforward extension of a 0/c utility function and a probability model across all possible subgroup models. We show that the resulting rule is essentially determined by the odds of subgroup models relative to the overall null hypothesis M0 of no treatment effects and relative to the overall alternative M1 of a common treatment effect in the entire patient population. This greatly simplifies posterior inference. We then generalize the approach to allow for subgroups that are characterized by arbitrary interactions of covariates. The two key elements of the generalization are a flexible nonparametric Bayesian response function and a separate description of the subgroup report that is not linked to the parametrization of the response model. We discuss an application to an adaptive enrichment design for targeted therapy.

email: **pmueller@math.utexas.edu**

SUBGROUP-BASED ADAPTIVE (SUBA) DESIGNS FOR MULTI-ARM BIOMARKER TRIALS

Yanxun Xu, University of Texas, Austin

Lorenzo Trippa, Harvard University

Peter Mueller, University of Texas, Austin

Yuan Ji*, NorthShore University HealthSystem and University of Chicago

Targeted therapies based on biomarker profiling are becoming a mainstream direction of cancer research and treatment. Depending on the expression of specific prognostic biomarkers, targeted therapies assign different cancer drugs to subgroups of patients even if they are diagnosed with the same type of cancer by traditional means, such as tumor location. For example, Herceptin is only indicated for the subgroup of patients with HER2+ breast cancer, but not other types of breast cancer. However, subgroups like HER2+ breast cancer with effective targeted therapies are rare and most cancer drugs are still being applied to large patient populations that include many patients who might not respond or benefit. Also, the response to targeted agents in human is usually unpredictable. To address these issues, we propose SUBA, subgroup-based adaptive designs that simultaneously search for prognostic subgroups and allocate patients adaptively to the best subgroup-specific treatments throughout the course of the trial. The main features of SUBA include the continuous reclassification of patient subgroups based on a random partition model and the adaptive allocation of patients to the best treatment arm based on posterior predictive probabilities. We compare the SUBA design with three alternative designs including equal randomization, outcome-adaptive randomization and a design based on a probit regression. In simulation studies we find that SUBA compares favorably against the alternatives.

email: **jiyuan@uchicago.edu**

DETECTION OF CANCER SUBGROUP ASSOCIATED ALTERNATIVE SPLICING

Jianhua Hu*, University of Texas MD Anderson Cancer Center

Xuming He, University of Michigan

Alternative splicing is known to be a critical factor in cancer formation and progression. In real experiments, high heterogeneity is often observed among cancer patients. Specifically, alternative splicing variants may show different degrees among or only occur to subgroups of cancer patients. We propose a penalized mixture statistical model integrated with dimension reduction of the interaction space based on ANOVA-type model and a sequential testing procedure to detect genes with such cancer subgroup structure.

email: **jhu@mdanderson.org**

55. CONTRIBUTED PAPERS: Methods to Assess Agreement

KAPPA STATISTICS FOR CORRELATED MATCHED-PAIR CATEGORICAL DATA

Zhao Yang*, University of Tennessee Health Science Center

Ming Zhou, Bristol-Myers Squibb

Kappa statistic is widely used to assess the agreement in the independent matched-pair data encountered in psychometrics, educational measurement, epidemiology, diagnostic imaging, and etc. However, the correlated matched-pair (clustered matched-pair and physician-patients) data which are more commonly expected from the medical practice, like the dental and ophthalmological care, the shared-decision making between general practice physician and patients. The traditional method, ignoring the dependence within a cluster, is generally inappropriate to handle the correlated matched-pair data. For clustered matched-pair data, a non-parametric variance estimator is proposed for the kappa statistic without within-cluster correlation structure or distributional assumptions. For the physician-patients data, relying on a plausible assumption of conditional independence (responses from patients of the same physician are conditionally independent given their physician's responses), a semi-parametric variance estimator is developed for the kappa statistic. The proposed estimators provide convenient tools for efficient computations and non-simulation-based alternatives to the existing bootstrap-based methods. The results from extensive Monte Carlo simulation suggest the proposed methods perform reasonably well for at least a moderately large number of clusters. Real collections of data are analyzed to illustrate the application.

email: **zyang20@uthsc.edu**

SAMPLE SIZE METHODS FOR CONSTRUCTING CONFIDENCE INTERVALS FOR THE INTRA-CLASS CORRELATION COEFFICIENT

Kevin K. Dobbin*, University of Georgia

Alexei C. Ionan, University of Georgia

The intraclass correlation coefficient (ICC) in a two-way analysis of variance is a ratio involving three variance components. Two recently developed methods for constructing confidence intervals (CI's) for the ICC are the Generalized Confidence Interval (GCI) and Modified Large Sample (MLS) methods. The resulting intervals have been shown to maintain nominal coverage. But methods for determining sample size for GCI and MLS intervals are lacking. This paper presents sample size methods that guarantee control of the mean width for GCI and MLS intervals. In the process, two variance reduction methods are employed, which we term dependent conditioning and inverse Rao-Blackwellization. Asymptotic results provide lower bounds for mean CI widths, and show that MLS and GCI widths are asymptotically equivalent. Simulation studies are used to investigate the new methods. It is shown that the new methods result in adequate sample size estimates, that the asymptotic estimates are accurate, and that the variance reduction techniques are effective. Software is made available that implements the methods in R. Future extensions of these results are discussed.

email: **dobbinke@uga.edu**

STATISTICAL METHODS FOR ASSESSING REPRODUCIBILITY IN MULTICENTER NEUROIMAGING STUDIES

Tian Dai*, Emory University

Ying Guo, Emory University

Recently in the neuroimaging community, there is an increasing trend of conducting multi-center studies. One major challenge arises when combining data from different centers is that the properties of brain images of the same person can vary considerably across centers since they are acquired using different scanners and protocols. Thus, it is crucial to effectively measure the reproducibility of images acquired from various sites. However simply assessing reproducibility based on raw brain images suffers from high dimensionality of the data and can be inefficient. In this work, we propose a two-stage network-based agreement method for fMRI data. In the first stage, we use a blind signal separation method to extract functional networks from fMRI data and estimate the temporal dynamics of these networks under experimental conditions. In the second stage, we propose agreement indices for functional data to assess the agreement of the brain network temporal dynamics estimated from the same subjects brain images acquired across different centers. We develop

nonparametric estimation methods for the proposed indices and establish their asymptotic properties. The proposed methods are applied to the Functional Bioinformatics Research Network (fBIRN) Phase I Traveling Subject study to investigate the reproducibility of fMRI images in multi-center studies.

email: **tian.dai88@gmail.com**

NONPARAMETRIC REGRESSION OF AGREEMENT MEASURE BETWEEN ORDINAL AND CONTINUOUS OUTCOMES

AKM F. Rahman*, Emory University

Limin Peng, Emory University

Ying Guo, Emory University

Amita Manatunga, Emory University

The effect of covariates on the agreement measure between ordinal and continuous outcomes is considered in a nonparametric framework. Peng et al. (2011) introduced a nonparametric broad sense agreement (BSA) measure between ordinal and continuous outcomes without considering covariates inormation. The inhomogeneity in BSA of heterogeneous population explained by covaraites is of considerable interest in many settings including mental health study. We propose a nonparametric kernel type estimator accommodating the effect of covariates on the agreement measure. Simulation studies demonstrates the effectiveness of the proposed method. We illustrate our methodologies via an application to a mental health study.

email: **afrahma@emory.edu**

INTER-OBSERVER AGREEMENT FOR A MIXTURE OF DATA TYPES

Shasha Bai*, University of Arkansas for Medical Sciences

Marcelo A. Lopetegui, The Ohio State University

Assessment of observer agreement has numerous applications in medical studies. A great amount of effort has been applied to further the advancement of measuring observer agreement for continuous and categorical data, in both univariate and multivariate cases. However, there has been a lack of research in this area for data containing a mixture of different data types. We present a scenario in clinical workflow studies where the assessment of inter-observer agreement is needed for a set of data containing a mixture of different types. We review the problems and limitations of using a single agreement statistic for these data, and provide a solution to these issues by way of a composite method of agreement statistics. This composite method offers meaningful and comprehensive evaluation of inter-observer agreement to clinicians.

email: **SBai@uams.edu**

ASSESSING REPRODUCIBILITY OF DISCRETE AND TRUNCATED RANK LISTS IN HIGH-THROUGHPUT STUDIES

Qunhua Li*, The Pennsylvania State University

Reproducibility is essential for reliable scientific discovery in high-throughput studies. Assessment of reproducibility often involves characterizing the concordance of ranked candidate lists from replicate experiments. Recently Li et al (2011) developed a copula mixture model, called IDR, to assess the reproducibility of candidates on two rank lists. This method allows one to select candidates according to a reproducibility-based criterion, and is particularly convenient when the selection thresholds are difficult to determine on the original lists. However, it is not applicable when a large number of ties ranks are present or when some candidates are unobserved in one replicate, for example, being truncated by a significance threshold. Here we generalize this method to handle discreteness and incompleteness in the rank lists using a latent variable approach. The generalized approach not only allows ties and partially replicated candidates, but also maintains the interpretability of the original model. Using simulation studies, we showed that, when discreteness or truncation is present, our method is able to identify substantially more real signals than the original model and produce better calibrated error rates than existing methods. We illustrated this method using a ChIP-seq dataset and a radiology dataset with highly discrete diagnosis ratings. Our method shows superior performance over existing methods in both cases.

email: **qunhua.li@psu.edu**

EXPONENTIATED LINDLEY POISSON DISTRIBUTION

Mavis Pararai*, Indiana University of Pennsylvania

Gayan Liyanag, Indiana University of Pennsylvania

Broderick Oluyede, Georgia Southern University

A new lifetime distribution called the exponentiated power lindley Poisson distribution is proposed. All it's properties will be explored including momentgenerating function, order statistics and entropy measures. Real data is applied to the proposed model.

email: **pararaim@iup.edu**

56. CONTRIBUTED PAPERS: Methylation and RNA Data Analysis

IDENTIFY DIFFERENTIAL ALTERNATIVE SPLICING EVENTS FROM PAIRED RNA-Seq DATA

Cheng Jia*, University of Pennsylvania

Mingyao Li, University of Pennsylvania

RNA-Seq has become indispensable for whole-transcriptome profiling due to its superiority over predecessor technologies in dynamic range and overall resolution. One of the areas in which RNA-Seq shines is detecting compositional differences among multiple isoforms transcribed from the same genetic locus between different conditions, i.e., differential alternative splicing (DAS). Current tools to discover and quantify DAS events from RNA-Seq data suffer from a spectrum of issues. The Bayesian method with MCMC simulation is inefficient and lacks power as a result of considering only junction reads. Moreover, existing likelihood-based methods rely on independent assumptions between samples, which limits their applicability in experiments with matched samples or repeated measurements. We have devised a novel approach to model the changes in the exon-inclusion levels using Hotelling's T-squared distribution, taking consideration of possible correlations between the paired samples. In addition, inspired by Fisher's method, we have implemented a p-value aggregation algorithm to generate gene-level p-values, which greatly simplifies the ensuing steps of data analyses.

email: **jiacheng@mail.med.upenn.edu**

FUNCTIONAL NORMALIZATION OF 450K METHYLATION ARRAY DATA IMPROVES REPLICATION IN LARGE CANCER STUDIES

Jean-Philippe Fortin*, Johns Hopkins Bloomberg School of Public Health

Aurelie Labbe, McGill University

Mathieu Lemire, Ontario Institute of Cancer Research

Brent W. Zanke, Ottawa Hospital Research Institute

Thomas J. Hudson, Ontario Institute of Cancer Research

Elana J. Fertig, Johns Hopkins School of Medicine

Celia MT Greenwood, Jewish General Hospital Montreal

Kasper D. Hansen, Johns Hopkins Bloomberg School of Public Health

We propose an extension to quantile normalization which removes unwanted technical variation using control probes. We adapt our algorithm, functional normalization, to the Illumina 450k methylation array and address the open problem of normalizing methylation data with global epigenetic changes, such as human cancers. Using datasets from The Cancer Genome Atlas and a large case-control study, we show that our algorithm outperforms all existing normalization methods with respect to replication of results between experiments, and yields robust results even in the presence of batch effects. Functional normalization can be applied to any microarray platform,provided suitable control probes are available.

email: **fortin946@gmail.com**

DETECTING DIFFERENTIALLY METHYLATED REGIONS (DMRs) BY MIXED-EFFECT LOGISTIC MODEL

Fengjiao Hu*, Georgia Regents University

Hongyan Xu, Georgia Regents University

Cancer is among the leading causes of death worldwide, and DNA methylation at CpG loci in genomic regions has important implications in cancer. A lot of statistical methods have been proposed by using proportion of methylated allele (methylation rate) to detect the association of cancer and methylation at single CpG locus. However considering the correlation among the methylation rates

of close-by CpG sites, we propose mixed-effect logistic regression model in this study to detect differentially methylated regions (DMRs), while treating methylation rates from each subject as a cluster, and proportions of methylated molecules for each site as observations. Age and gender are included in the model as the covariates. Simulations were performed to show that the mixed-effect logistic regression was robust to detect DMRs after adjusting covariates, with Type I error well-controlled and good power. The results indicating that this mixed-effect logistic regression method is a promising approach for detecting DMRs with methylation data from next-generation **sequencing.**

email: fhu@gru.edu

PENALIZED MODELING FOR VARIABLE SELECTION AND ASSOCIATION STUDY OF HIGH-DIMENSIONAL MicroRNA DATA WITH REPEATED MEASURES

Zhe Fei*, University of Michigan

Yinan Zheng, Northwestern University

Wei Zhang, University of Illinois, Chicago

Justin B. Starren, Northwestern University

Lei Liu, Northwestern University

Andrea A. Baccarelli, Harvard School of Public Health

Yi Li, University of Michigan

Lifang Hou, Northwestern University

Motivation: MicroRNAs (miRNAs) are short single-stranded non-coding molecules that usually function as negative regulators to silence or suppress gene expression. Owning to the dynamic nature of miRNA and reduced microarray and sequencing costs, a growing number of researchers are now measuring high-dimensional miRNA expression data using repeated or multiple measures in which each individual has more than one sample collected and measured over time. However, the commonly used univariate association testing or the site-by-site (SBS) testing may underutilize the longitudinal feature of the data, leading to underpowered results and less biologically meaningful results. Results: We propose a penalized regression model incorporated with grid search method (PGS), for analyzing associations of high-dimensional miRNA expression data with repeated measures as well as variable selection of significant miRNAs. The development of this analytical framework was motivated by a real-world miRNA dataset. Comparisons between PGS and the SBS testing revealed that PGS provided smaller phenotype prediction errors and higher enrichment of phenotype-related biological pathways than the SBS testing. Our extensive simulations showed that PGS provided more accurate estimates and higher sensitivity than the SBS testing with comparable specificities.

email: **feiz@umich.edu**

COMPARISON OF PAIRED TUMOR-NORMAL METHODS FOR DIFFERENTIAL EXPRESSION ANALYSIS OF RNA-Seq DATA

Janelle R. Noel*, University of Kansas Medical Center

Alice Wang, University of Kansas Medical Center

Rama Raghavan, University of Kansas Medical Center

Prabhakar Chalise, University of Kansas Medical Center

Byunggil Yoo, Childrens Mercy Hospital Kansas City

Sumedha Gunewardena, Kansas Intellectual and Developmental Disabilities Research Center

Jeremy Chien, University of Kansas Medical Center

Brooke L. Fridley, University of Kansas Medical Center

Discovery of differentially expressed (DE) genes is imperative for the understanding of the genomic basis of complex diseases and phenotypes. Concurrently, there is a lack of RNA-seq methods that can account for dependency in paired designs. In this study, we applied 8 DE analysis methods to an RNA-seq study involving paired ovarian tumor samples pre- and post- treatment with carboplatin taken from 11 ovarian cancer patients. Supplementary to the empirical comparison with real data, we simulated 1,000 paired and unpaired datasets under numerous scenarios (i.e. varying level of dependency, number of subjects, and effect size). To assess the type I error rates for the paired and unpaired datasets, paired t-tests and two-sample t-tests were conducted. Results showed 35 common genes detected by six DE analysis methods ($p<0.05$). Under the null hypothesis, the type I error rate was conservative for paired designs that were analyzed incorrectly using two-sample t-tests ($0.001<p<0.016$). We also found that the power to detect average mean

differences in DE was not affected by either paired (rho=0.3, 0.5) or unpaired designs (rho=0). This study demonstrates the differences in DE analysis methods when selecting "associated" genes, and the importance of using proper statistical tests for RNA-seq data.

email: **jnoel@kumc.edu**

DETECTING DIFFERENTIAL ALTERNATIVE SPLICING WITH BIOLOGICAL REPLICATES BETWEEN TWO GROUPS FROM RNA-Seq DATA

Yu Hu*, University of Pennsylvania

Cheng Jia, University of Pennsylvania

Dwight Stambolian, University of Pennsylvania

Mingyao Li, University of Pennsylvania

Alternative splicing, a post-transcriptional process that allows multiple messenger RNA (mRNA) isoforms to be produced by a single gene, is a regulated process and a major mechanism for generating protein diversity. Detecting differential alternative splicing (DAS) between two groups of samples (e.g., cases vs. controls) could provide an effective way to discover disease susceptibility genes. To detect DAS from RNA-Seq data, we make use of information on known gene structures and pre-estimated isoform relative abundances. For each alternatively spliced exon of a gene, we divide isoforms into two categories depending on whether the exon is included or not. The inclusion level of the alternatively spliced exon is then estimated as the total relative abundances of all isoforms with the exon included. Our estimation utilizes all available reads of the gene and extracts more information on alternative splicing than methods based on junction reads alone. Additionally, alternatively spliced exons reflecting the same isoform(s) can be aggregated together, thus reducing the number of subsequent tests. To detect DAS, we assume the exon-inclusion levels follow a beta or an inflated-beta distribution, and test DAS by comparing the parameters of the beta or inflated-beta distributions between the two groups through the use of a likelihood ratio test. Results based on simulated data and the analysis of a real RNA-seq dataset on human eyes demonstrate the superior performance of our method as compared to several existing methods, including Cuffdiff, DEXSeq, MATS, and DSGSeq.

email: **huyu1@mail.med.upenn.edu**

FUNCTIONAL REGION-BASED TEST FOR DNA METHYLATION

Kuan-Chieh Huang*, University of North Carolina, Chapel Hill

Yun Li, University of North Carolina, Chapel Hill

Recent technological advances have allowed us to conduct large-scale epigenome-wide association studies (EWASs). DNA methylation (DNAm) is of particular interest because it is highly dynamic and has been shown to be associated with many complex human traits. Typically, DNAm level at hundreds of thousands of sites is measured and each of these sites is examined separately (i.e., single-site analysis). However, because of the correlation structure among the sites and because many of them fall in naturally defined regions (e.g., gene or regulatory), it is conceptually straightforward to imagine achieving enhanced statistical power by performing region-based test especially when there are small or moderate signals in the region. Here, we propose FunMethyl, a functional regression framework to perform association testing between multiple DNAm sites in a region and a quantitative outcome. Instead of collapsing DNAm variants or building a kernel matrix, we consider every individual's DNAm levels in a region as a stochastic process and further estimate the DNAm function in that region using the proposed smoothing techniques. Our results from both real data based simulations and real data analysis clearly show that FunMethyl outperforms single-site analysis across a wide spectrum of realistic scenarios.

email: **kchuang@live.unc.edu**

57. CONTRIBUTED PAPERS: New Developments in Imaging

ESTIMATING DYNAMICS OF WHOLE-BRAIN FUNCTIONAL CONNECTIVITY IN RESTING-STATE fMRI BY FACTOR STOCHASTIC VOLATILITY MODEL

Chee-Ming Ting*, Universiti Teknologi Malaysia, Malaysia

Hernando Ombao, University of California, Irvine

Sh-Hussain Salleh, Universiti Teknologi Malaysia, Malaysia

Most studies of resting-state fMRI assume temporal stationarity of functional connectivity (FC) between distinct brain

regions, identified using a constant covariance model fitted on the entire time course. However, emerging evidence suggests that FC may exhibit dynamic changes over time, arguably even more prominent during rest when mental activities are unconstrained. We consider the problem of quantifying these changes which may provide insights into the fundamental properties of brain networks. Recent studies employed the conventional sliding-window technique assuming locally stationary covariances over short-time segments. In this work, we use multivariate stochastic volatility (MSV) model that allows a time-varying covariance process to better capture the non-stationary dynamics of FC in resting-state fMRI. We further incorporate a latent factor model to achieve a reliable and computationally efficient estimation of large-dimensional covariance matrices for analyzing evolving full-brain networks with a large number of nodes. The stochastic volatility analysis is performed on a lower-dimensional common factor series instead of on the observations directly. We propose a robust two-step estimation procedure by first estimating the factor model using the principal component (PC) methods followed by the MSV model with quasi maximum likelihood (QML) using Kalman filter and expectation maximization (EM) algorithm. The proposed method was evaluated on a resting-state fMRI dataset of 25 healthy subjects.

email: **cmting1818@yahoo.com**

KERNEL SMOOTHING GEE FOR LONGITUDINAL fMRI STUDIES

Yu Chen*, University of Michigan

Min Zhang, University of Michigan

Timothy D. Johnson, University of Michigan

Longitudinal fMRI studies are beginning to play an important role in understanding the development of the human brain. In this setting random effects models have convergence issues and, typically, generalized estimating equations (GEE) are employed. However, due to the large number of multiple comparisons, GEE methods suffer from a lack of statistical power. To increase power, we propose a kernel smoothing generalized estimating equation (KernGEE) method with a locally adaptive bandwidth to study the temporal trend of fMRI measurements for each brain voxel. In order to address the spatial correlation among voxels and to increase power, we use a kernel function that borrows information across neighboring voxels, spatially smoothing parameter estimates. The kernel bandwidth at each voxel is determined by leave-one-out cross validation. Therefore, our method can provide a set of spatially smoothed estimators for each voxel with increased efficiency. Meanwhile, correction for multiple comparisons is obtained using Efron's empirical null distribution method. We apply our KernGEE method to a longitudinal dataset studying brain mechanisms of risk for alcoholism and other substance abuse. We will also investigate the relationships between activated brain regions and several covariates including IQ, age, gender, behavioral and personality variables.

email: **cheyu@umich.edu**

A HIERARCHICAL BAYESIAN MODEL FOR STUDYING THE IMPACT OF STROKE ON BRAIN MOTOR FUNCTION

Zhe Yu*, University of California, Irvine

Raquel Prado, University of California, Santa Cruz

Erin Burke Quinlan, University of California, Irvine

Steven C. Cramer, University of California, Irvine

Hernando Ombao, University of California, Irvine

Stroke is a disturbance in the blood supply to the brain which results in the loss of brain functions, in particular motor function. A study was conducted by neuroscientists to investigate the impact of stroke on the motor-related brain regions. In the study, functional MRI (fMRI) data were collected from stroke patients and healthy controls while the subjects performed a simple hand motor task. To explore the changes in the brain due to stroke, we developed a hierarchical Bayesian approach for modeling the multi-subject fMRI data. Our approach simultaneously estimates activation and connectivity at the group level, and provides estimates for region/subject-specific hemodynamic response function (HRF) and condition-specific connectivity. Moreover, the use of spike and slab priors allows for direct posterior inference on the connectivity network. Using our model, we observed several potential effects of stroke on the motor system function: executing the simple motor task

requires more involvement of the higher level motor control regions in both the stroke affected and unaffected hemisphere compared to healthy subjects. We also noted increased communication within and between these secondary motor regions. These findings provide insight into different neural correlates of movement after stroke versus healthy individuals.

email: **zhey@uci.edu**

SOURCE ESTIMATION FOR MULTI-TRIAL MULTI-CHANNEL EEG SIGNALS: A STATISTICAL APPROACH

Yuxiao Wang*, University of California, Irvine

Hernando Ombao, University of California, Irvine

Raquel Prado, University of California, Santa Cruz

Electroencephalography (EEG) has been widely used in studying the dynamics in human brains due to its relatively high temporal resolution (in millisecond). EEGs are indirect measurements of neuronal sources. Estimation of the underlying sources is challenging due to the ill-posed inverse problem. EEGs are typically modeled as a linear mixing of the underlying sources. Here, we consider source modeling and estimation for multi-channel EEG data recorded over multiple trials. We propose parametric models to characterize the latent source signals and develop methods for estimating the processes that drive the source -- instead of merely recovering the source signals. Moreover, we develop metrics for connectivity between channels through latent sources by studying the properties of the estimated mixing matrix. Our estimation procedure pulls information from all trials using a two-stage approach: first, we apply the second order blind identification (SOBI) method to estimate the mixing matrix and second, we estimate the parameters for latent sources using maximum likelihood. Our methods will also impose regularization to ensure sparsity. Our proposed methods have been evaluated on both simulated data and EEG data obtained from a motor learning study. This project is in collaboration with the Space-Time Modeling group at UC Irvine.

e-mail: **yxwang87@gmail.com**

AN EXPLORATORY DATA ANALYSIS OF EEGs TIME SERIES: A FUNCTIONAL BOXPLOTS APPROACH

Duy Ngo*, University of California, Irvine

Hernando Ombao, University of California, Irvine

Marc G. Genton, University of Science and Technology

Ying Sun, King Abdullah University of Science and Technology

We conduct exploratory data analysis on electroencephalograms (EEG) data to study the brain's electrical activity during resting state. The standard approaches to analyzing EEG are classified either into the time domain (ARIMA modeling) or the frequency domain (via periodograms). Our goal here is to develop a systematic procedure for analyzing periodograms collected across many trials (which consists of 1 second traces) during the entire resting state period. In particular, we use functional boxplots to extract information from the many trials [1]. First, we formed consistent estimators for the spectrum by smoothing the periodograms using a bandwidth selected using the generalized cross-validation of the Gamma deviance. We then obtained descriptive statistics from the smoothed periodograms using functional box plots which provide the median and outlying curves. The performance of functional boxplot is compared with the classical point-wise boxplots and functional mixed effects models in a simulation study and the EEG data. Moreover, we explored the spatial variation of the spectral power for the alpha and beta frequency bands by applying the surface boxplot method on periodograms computed from the many resting-state EEG traces. This work is in collaboration with the Space-Time Group at UC Irvine.

[1] Sun, Y., and Genton, M.G. (2011), "Functional Boxplots," Journal of Computational and Graphical Statistics, 20, 316-334.

e-mail: **dngo5@uci.edu**

A BAYESIAN FUNCTIONAL LINEAR COX REGRESSION MODEL (BFLCRM) FOR PREDICTING TIME TO CONVERSION TO ALZHEIMER'S DISEASE

Eunjee Lee*, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

Dehan Kong, University of North Carolina, Chapel Hill

Yalin Wang, Arizona State University

Kelly Sullivan Giovanello, University of North Carolina, Chapel Hill

Joseph Ibrahim, University of North Carolina, Chapel Hill

The aim of this paper is to develop a Bayesian functional linear Cox regression model (BFLCRM) with both functional and scalar covariates. This new development is motivated by establishing the likelihood of conversion to Alzheimer's disease (AD) in 346 patients with mild cognitive impairment (MCI) enrolled in the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI1) and the optimal early markers of conversion. These 346 MCI patients were followed over 48 months, with 161 MCI participants progressing to AD at 48 months. The functional linear Cox regression model was used to establish that the conversion time to AD can be accurately predicted by functional covariates including hippocampus surface morphology and scalar covariates including brain MRI volumes, cognitive performance (ADAS-Cog), and APOE status. Posterior computation proceeds via an efficient Markov chain Monte Carlo algorithm. A simulation study is performed to evaluate the finite sample performance of BFLCRM.

e-mail: **eunjee2@email.unc.edu**

58. CONTRIBUTED PAPERS: Latent Variable and Principal Component Models

A LATENT VARIABLE MODEL FOR ANALYZING CORRELATED ORDERED CATEGORICAL DATA

Ali Reza Fotouhi*, University of The Fraser Valley

In many statistical studies in medicine, clinical trials, and agriculture the responses are recorded on an ordinal scale. The ordered responses may be clustered and the subjects within the clusters may be positively correlated. A commonly used method to accommodate this correlation is to add a random component to the linear predictor of each clustered response. Moreover, some unobservable characteristics may have significant effect on the categorization of the responses. In this article we introduce a latent variable model for analyzing ordered categorical data in which the random effects are not a nuisance but are of explanatory interest. We introduce two correlated random effects in the latent variable model to control for cluster and category variability. Four commonly used models probit, logistic, complementary log-log, and log-log models for correlated ordered categorical data are special cases of this latent-variable model. We validate the proposed model through a simulation study and use it to analyze the data obtained from twelve populations of strawberries in a randomized block experiment.

email: **ali.fotouhi@ufv.ca**

ESTIMATION OF BRANCHING CURVES IN THE PRESENCE OF SUBJECT SPECIFIC RANDOM EFFECTS

Angelo Elmi*, The George Washington University

Sarah J. Ratcliffe, University of Pennsylvania

Wensheng Guo, University of Pennsylvania

Branching curves are a technique for modeling curves that change trajectory at a change (branching) point. Currently, the estimation framework is limited to independent data, and smoothing splines are used for estimation. Here, extension of the branching curve framework to the longitudinal setting, where the branching point varies by subject, will be discussed. If the branching point is modeled as a random effect, then the longitudinal branching curve framework is a Semiparametric Nonlinear Mixed Effects Model. Given existing issues with using random effects within a smoothing spline, we express the model as a B-spline Based Semiparametric Nonlinear Mixed Effects Model. Simple, clever smoothness constraints are enforced on the B-splines at the change point. The method is applied to Women's Health data where we model the shape of the labor curve (cervical dilation measured longitudinally) before and after treatment with oxytocin (a labor stimulant).

email: afelmi@gwu.edu

COMPOSITE LARGE MARGIN CLASSIFIERS WITH LATENT SUBCLASSES FOR HETEROGENEOUS BIOMEDICAL DATA

Guanhua Chen*, Vanderbilt University

Yufeng Liu, University of North Carolina, Chapel Hill

Michael R. Kosorok, University of North Carolina, Chapel Hill

High dimensional classification problems are prevalent in a wide range of modern scientific applications. Despite a large number of candidate classification techniques available to use, practitioners often face a dilemma of the choice between linear and general nonlinear classifiers. Specifically, simple linear classifiers have good interpretability, but may have limitations in handling data with complex structures. In contrast, general nonlinear kernel classifiers are more flexible but may lose interpretability and have higher tendency for overfitting. In this paper, we consider data with potential latent subgroups in the classes of interest. We propose a new method, namely the Composite Large Margin Classifier (CLM) to address the issue of classification with latent subclasses. The CLM aims to find three linear functions simultaneously: one linear function to split the data into two parts, with each part being classified by a different linear classifier. Our method has comparable prediction accuracy to a general nonlinear kernel classifier and it maintains the interpretability of traditional linear classifiers. We demonstrate the competitive performance of the CLM through comparisons with several existing linear and nonlinear classifiers by Monte Carlo experiments. Analyzing Alzheimer's disease classification problem using CLM not only provides lower classification error in discriminating cases and controls, but also identifies subclasses in controls which are more likely to develop into disease in the future.

email: **g.chen@vanderbilt.edu**

EVALUATION OF COVARIATE-SPECIFIC ACCURACY OF BIOMARKERS WITHOUT A GOLD STANDARD

Zheyu Wang*, Johns Hopkins University

Xiao-Hua Zhou, University of Washington

In recent years, advances in biomarker discovery have re-energized the field of diagnostic medicine, as researchers continuously strive to obtain more convenient, economical, accurate, and/or timely diagnoses, by adding or combining various biomarkers to create novel diagnostic procedures. Accordingly, it is important to assess the accuracy of biomarkers. There are three major issues in biomarker evaluation: 1) The underlying medical condition, or the gold standard, can unknown due to time and cost constraints, lack of biotechnology, or concerns over the invasive nature of a diagnostic procedure. This issue is becoming more common and pressing with the growing interest and emphasis on preclinical diagnosis and prevention. 2) Compared with traditional diagnostic tests, biomarker levels are more easily affected by patients' characteristics. Therefore diagnosis based on biomarkers need to be personallized. 3)With the improvement in clinical practice, there is a need to go beyond traditional binary disease status and incorporate ordinal gold standard. In this talk, we will propose a finite mixture model approach to address these issues.

email: **wangzy@jhu.edu**

LINEAR MIXED MODEL WITH UNOBSERVED INFORMATIVE CLUSTER SIZE: APPLICATION TO A REPEATED PREGNANCY STUDY

Ashok K. Chaurasia*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

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

Modeling with informative cluster size is a common issue in analyzing clustered data, where the outcome is associated with cluster size. This paper addresses the informative cluster size problem in linear mixed models when the cluster size is censored on all subjects. This problem is motivated by the NICHD Consecutive Pregnancy Study, where the objective is to study the relationship between birthweight and parity. It is hypothesized that the birthweight profile is associated with the number of births over a woman's lifetime, resulting in an informative cluster size. However, in this study, a woman's lifetime number of births is not observed (censored at the end of the study window). In this paper we develop a pattern mixture model to account for informative

cluster size by treating the unobserved cluster size (lifetime number of births) as a latent variable. We compare this approach with the simple alternatives where we use the observed number of births at the end of the study as cluster size. For estimating the population mean trajectory, we show theoretically, with simulations and in the real data application that the simple approach can serve as reliable approximation for the latent variable approach.

email: **achaurasia.uconn@gmail.com**

A SEMIPARAMETRIC MODEL OF ESTIMATING NON-CONSTANT FACTOR LOADINGS

Zhenzhen Zhang*, University of Michigan

Brisa Sanchez, University of Michigan

Factor analysis is a commonly used method in modeling multivariate exposure data. Typically, the measurement model is assumed to have constant factor loadings. We propose models that relax this assumption by using penalized splines to estimate factor loadings that change with other covariates. We implement two different approaches of penalizing the smoothing splines, the generalized cross validation criterion and the random effects, and incorporate them into the EM algorithm through Newton-Raphson and Monte-Carlo methods. The likelihood ratio test is used to test whether a factor loading is constant.

email: **zhzh@umich.edu**

NESTED PARTIALLY-LATENT CLASS MODELS (npLCM) FOR ESTIMATING DISEASE ETIOLOGY IN CASE-CONTROL STUDIES

Zhenke Wu*, Johns Hopkins University

Scott L. Zeger, Johns Hopkins University

The Pneumonia Etiology Research for Child Health (PERCH) study attempts to infer the distribution of pneumonia-causing bacterial or viral pathogens in developing countries from measurements outside of the lung. Recent developments in test standardization make it possible to collect multiple specimens to detect a large number of pathogens at once with varying degrees of etiologic relevance and measurement precision. With this data, researchers seek to estimate the population fraction of cases caused by each pathogen, and to develop algorithms to assist clinical diagnosis when presented with complex data on an individual case. We describe a latent variable model to address these two analytic goals using data from a case-control design. We assume each observation is a draw from a mixture model for which each component represents one pathogen. Conditional dependence among multivariate binary measurements on a single subject is induced by nesting subclasses within each disease class. Measurement precision can be estimated using the control sample for whom the etiologic class is known. We use stick-breaking priors on the subclass weights to estimate the population and individual etiologic distributions that are averaged across models indexed by different numbers of subclasses. Assessment of model fit and individual diagnosis is done using posterior samples drawn by Gibbs Sampling. We demonstrate the method's operating characteristics via a simulation study tailored to the motivating scientific problem and illustrate the model with a detailed analysis of PERCH study data.

email: **zhwu@jhu.edu**

59. CONTRIBUTED PAPERS: Developments and Applications of Clustering, Classification, and Dimension Reduction Methods

SEPARABLE SPATIO-TEMPORAL PRINCIPAL COMPONENT ANALYSIS

Lei Huang*, Johns Hopkins University

Philip T. Reiss, New York University School of Medicine

Luo Xiao, Johns Hopkins University

Vadim Zipunnikov, Johns Hopkins University

Martin A. Lindquist, Johns Hopkins University

Ciprian Crainiceanu, Johns Hopkins University

Current brain imaging studies often acquire large images that are observed over time. Examples of such studies include BOLD fMRI, DCE-MRI and dynamic PET. To model such data we introduce a class of separable spatio-temporal process using explicit latent process modeling. To account for the size and spatio-temporal structure of the data, we extend principal component analysis to achieve dimensionality reduction at the individual process level. We introduce necessary identifiability conditions for each model and develop scalable estima-

tion procedures. The method is motivated by and applied to an fMRI study designed to analyze the relationship between pain and brain activity.

email: **huangracer@gmail.com**

PENALIZED CLUSTERING USING A HIDDEN MARKOV RANDOM FIELD MODEL: DETECTING STATE-RELATED CHANGES IN BRAIN CONNECTIVITY

Yuting Xu*, Johns Hopkins University

Martin Lindquist, Johns Hopkins University

In the statistical analyses of task-based fMRI time series, the activity in a set of regions of interest (ROIs) change with the experimental process. We assume that the activity in the ROIs can be modeled as Gaussian random vectors, with a mean and correlation matrix that changes between different states as the task progresses. In this work, we introduce a penalized Gaussian Hidden Markov random field model, to detect changes in brain connectivity and simultaneously achieve shrinkage parameter estimation. The clustering assignment, or the hidden state, is obtained via MRF-MAP estimation, which takes into account the time-dependent structure within and across subjects. We incorporate several popular sparse precision matrix estimation algorithms to achieve better variable selection and parameter estimation. The method is applied to various simulation data as well as fMRI data from an anxiety study, illustrating the efficacy of the proposed method compared to alternative methods.

email: **xuyuting@jhu.edu**

CLUSTERING OF BRAIN SIGNALS USING THE TOTAL VARIATION DISTANCE

Carolina Euán*, Centro de Investigación en Matemáticas (CIMAT), A.C.

Hernando Ombao, University of California, Irvine

Joaquin Ortega, Centro de Investigación en Matemáticas (CIMAT), A.C.

Pedro Alvarez-Esteban, Universidad de Valladolid, Spain

We are interested in studying the spatial structure of brain signals during a learning motor task. Our research is based on the spectral analysis of the electroencephalograms (EEG) traces recorded by our neurologist collaborator. The EEG data was collected across three different phases of rest and practice. Our goal is to develop a procedure for detecting and characterizing differences in spatial variation of the EEG power between the different learning phases. At each cannel we estimate the spectrum using the smoothed periodograms. These indicate the distribution of power across different frequency bands in each channel. Our principal tool is the Total Variation (TV) Distance which is a similarity measure in the clustering algorithm. Our procedure esssentially clusters time series at different channels that share similar spectral structures (i.e., similar smoothed periodograms). Using our proposed procedure we will be able to cluster channels that behave similarly within each learning phase of the experiment. This work has been in collaboration with the Space-Time Modeling group at UC Irvine.

email: **ceuan@uci.net**

IMPACT OF DATA REDUCTION ON ACCELEROMETER DATA IN CHILDREN

Daniela Sotres-Alvarez*, University of North Carolina, Chapel Hill

Yu Deng, University of North Carolina, Chapel Hill

Guadalupe X. Ayala, San Diego State University

Mercedes Carnethon, Northwestern University

Alan M. Delamater, University of Miami

Carmen R. Isasi, Albert Einstein College of Medicine

Sonia Davis, University of North Carolina, Chapel Hill

Kelly R. Evenson, University of North Carolina, Chapel Hill

Accelerometry, typically worn by participants for one week, provides an indicator of physical activity and sedentary behavior through measured accelerations. A challenge with accelerometer data is the difficulty to distinguish non-wear from sedentary behavior, since theoretically they both can register 0 counts per epoch (e.g. 15 seconds). Current practice defines non-wear (which becomes missing data) with a certain number of consecutive zero counts, and summarizes accelerometer data only for individuals with a minimum number of days each with a minimum number of hours of wear (e.g. at least 3 days of 10 hours/day). This approach does not make full use of the rich information contained in the data and might not effectively handle the missing data on incomplete days. We compared various methods to account for missing data using data

from the SOL Youth Study (1,400 children ages 8-16 y). The percentage of complete data from standard approach varied from 70.6% children with 4+ wear days (10hrs/day) including a weekend day to 89.8% with 3+ wear days (8hr/day) without including necessarily a weekend day. We compared the bias and precision of estimates for physical activity and sedentary behavior between the standard approach, imputation methods and Wavelet-based functional mixed models.

email: **dsotres@unc.edu**

LEARNING LOGIC RULES FOR DISEASE CLASSIFICATION: WITH AN APPLICATION TO DEVELOPING CRITERIA SETS FOR THE DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS

Christine M. Mauro*, Columbia University

Donglin Zeng, University of North Carolina, Chapel Hill

M. Katherine Shear, Columbia University

Yuanjia Wang, Columbia University

In Psychiatry, clinicians rely on criteria sets from the Diagnostic and Statistical Manual of Mental Disorders to make diagnoses. Each criteria set has several symptom domains. In order to be diagnosed, an individual must meet the minimum number of symptoms required for each domain. Several approaches to determine these minimum values are proposed. In simple scenarios, an exhaustive search is feasible. For more complicated scenarios, another approach is necessary. Given disease status and the count of symptoms present in each domain, a linear discriminate function is fit within each domain. Since one must meet the criteria for all domains, a positive diagnosis is only issued if the prediction in each domain is positive. The overall decision rule is therefore the intersection of all the domain specific rules. We propose two algorithms, SVM Iterative and Logistic Iterative, to fit this model. The proposed methods are flexible enough to be adapted to complicated settings including using high-dimensional data, other logic structures, or non-linear discriminant functions. In simulations, the Exhaustive Search (when applicable), SVM Iterative and Logistic Iterative perform well when compared with the oracle rule. The methods are then applied to construct a criteria set for Complicated Grief, a new psychiatric disorder.

email: **cmm2212@cumc.columbia.edu**

CHARACTERIZING TYPES OF PHYSICAL ACTIVITY: AN UNSUPERVISED WAY

Jiawei Bai*, Johns Hopkins University

Luo Xiao, Johns Hopkins University

Vadim Zipunnikov, Johns Hopkins University

Ciprian M. Crainiceanu, Johns Hopkins University

Predicting the type of activity performed by human subjects using accelerometry data is crucial to many different areas of research. Currently supervised learning methods dominate this field since they provide high prediction accuracy when the activity types of interest are known. However, in free-living circumstances the activity types of interest are often unclear. We proposed an unsupervised learning method to extract the key basic components (movelets) of the acceleration time series. These key movelets acted like building blocks which constructed the whole signal of activity. We further investigated the interpretation of these key movelets and found most of them having the signal patterns very close to some important basic activity types, such as standing, sitting, lying and walking. Using this method, we could avoid manually defining types or categories of activity, and build subject-specific dictionaries of key components for the subjects. This allows for better and deeper comparison of physical activity status of different subjects.

email: **javybai@gmail.com**

SIMULTANEOUS MODEL-BASED CLUSTERING AND VARIABLE SELECTION: EXTENSION TO MIXED-DISTRIBUTION DATA

Katie Evans, Dupont

Tanzy M. T. Love*, University of Rochester

Sally W. Thurston, University of Rochester

Current model-based clustering methods, such as LatentGold (Vermunt & Magidson, 2005) and MultiMix (Hunt & Jorgensen, 1999) can accommodate data with variables of mixed-distributional forms. In these methods, statistical criteria guide the manual selection of relationships between clustering variables, but not the selection of variables important to clustering. Clustering variable selection procedures, such as Raftery & Dean (2006) and Maugis et al (2009), are limited to data consisting of normally distributed variables. Our new frame-

work for model-based clustering on data with continuous and discrete variables extends the cluster variance structure framework set forth by Fraley and Raftery (1999). In modeling how each variable contributes to cluster determination, we allow for relations within and between the continuous and discrete variables (termed mixClust.) We also modify and extend existing likelihood-based variable selection procedures to accommodate data with variables of mixed-distributional forms (ESR) and only require at least one continuous variable. Simulation study results show desirable properties of our method when applied to data with variables of mixed-distributional forms and improved performance over existing methods when applied to only normally distributed data. Applying mixClust and ESR to prostate cancer data generates subgroups with different responses to treatment.

email: **tanzy_love@urmc.rochester.edu**

60. CONTRIBUTED PAPERS: Survival Analysis: Methods Development and Applications

PREDICTIVE MODEL AND DYNAMIC PREDICTION FOR RECURRENT EVENTS WITH DEPENDENT TERMINATION

Li-An Lin*, University of Texas Health Sciences Center at Houston

Sheng Luo, University of Texas Health Sciences Center at Houston

Barry Davis, University of Texas Health Sciences Center at Houston

In clinical trials of hypertension medications, cardiovascular disease events frequently recur over the study follow-up times. Recently, predictive models have been routinely used to assess risk in clinical trials. Patients are interested in knowing their risk of disease recurrence and death as their conditions change, such as risk factors, event history. However, modeling event history that would facilitate prediction is still underdeveloped. In this article, we propose a predictive model based on generalized renewal processes and joint frailty model where the impact of past events on further events and individual heterogeneity have been accounted. The proposed model is assessed by the receiver operating characteristic curve, which is derived using Monte Carlo approach. Simulation studies demonstrate that the proposed methods perform well in practical situations. After the model has been fitted to the training dataset, one can estimate new patients' future survival probability for next disease events and death conditional on current information. Finally, the proposed tools are applied to the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT).

email: **li.an.lin@uth.tmc.edu**

AN EXTENDED SELF-TRIGGERING MODEL FOR RECURRENT EVENT DATA

Jung In Kim*, University of North Carolina, Chapel Hill

Feng-Chang Lin, University of North Carolina, Chapel Hill

Jason Fine, University of North Carolina, Chapel Hill

Recurrent event data frequently appear in longitudinal studies when study subjects experience more than one event during the observation period. In reality, one may observe the subsequent events influenced by previous events; hence, the triggering scheme of event occurrence shall be considered when modeling such data. In this paper, we extend the Cox proportional hazard model with time-varying information of previous events to enhance the model fitness and prediction. Parameter estimation and statistical inferences can be easily achieved via a partial likelihood function. A jointed statistical test is provided to assess the existence of the effects from previous events. We demonstrate our approach via comprehensive simulation studies and cystic fibrosis registry data in chronic pseudomonas infections. Significantly, our model provides a better prediction amongst currently existing ones.

email: **jikim@live.unc.edu**

A PAIRWISE-LIKELIHOOD AUGMENTED ESTIMATOR FOR THE COX MODEL UNDER LEFT-TRUNCATION

Fan Wu*, University of Michigan

Sehee Kim, University of Michigan

Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Yi Li, University of Michigan

Survival data collected from prevalent cohorts are subject to left-truncation. The conventional conditional approach using Cox model disregards the information in the marginal likelihood of truncation

time thus can be inefficient. On the other hand, the stationary assumption under length-biased sampling (LBS) methods to incorporate the marginal information can lead to biased estimation when it is violated. In this paper, we propose a semiparametric estimation method by augmenting the Cox partial likelihood with a pairwise likelihood, by which we eliminate the unspecified truncation distribution in the marginal likelihood, yet retain the information about regression coefficients and the baseline hazard. Exploring self-consistency of the estimator, we give a fast algorithm to solve for the regression coefficients and the cumulative hazard simultaneously. The proposed estimator is shown to be consistent and asymptotically normal with a sandwich-type consistent variance estimator. Simulation studies show a substantial efficiency gain in both the regression coefficients and the cumulative hazard over Cox model estimators, and that the gain is comparable to LBS methods when the stationary assumption holds. For illustration, we apply the proposed method to the RRI-CKD data.

email: **fannwu@umich.edu**

RANK-BASED TESTING BASED ON CROSS-SECTIONAL SURVIVAL DATA WITH OR WITHOUT PROSPECTIVE FOLLOW-UP

Kwun Chuen Gary Chan*, University of Washington

Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Existing linear rank statistics cannot be applied to cross-sectional survival data without follow-up since all subjects are essentially censored. However, partial survival information is available from backward recurrence times, and is frequently collected from health surveys without prospective follow-up. Under length-biased sampling, a class of linear rank statistics is proposed based only on backward recurrence times without any prospective follow-up. When follow-up data are available, the proposed rank statistic and a conventional rank statistic that utilizes follow-up information from the same sample are shown to be asymptotically independent. We discuss four ways to combine these two statistics when follow-up is present. Simulations show that all combined statistics have substantially improved power compared to conventional rank statistics, and a Mantel-Haenszel test performed the best among the proposal statistics. The method is applied to a cross-sectional health survey without follow-up and a study of Alzheimer's disease with prospective follow-up.

email: **kcgchan@u.washington.edu**

COMPUTATION EFFICIENT MODELS FOR FITTING LARGE-SCALE SURVIVAL DATA

Kevin He*, University of Michigan

Yanming Li, University of Michigan

Ji Zhu, University of Michigan

Yi Li, University of Michigan

Time-varying effects model is a flexible and powerful tool for modeling the dynamic changes of covariate effects. In survival analysis, however, time-varying effects is often difficult to model since the the modifying variable, time, is part of response instead just being a covariant. The computational burden increases quickly as the number of sample size grows and analysis with relatively large sample size out-powers existing statistical methods and softwares. We propose a novel application of Quasi-Newton method with inexact line search procedure to model the dynamic changes of regression coefficients in survival analysis. The algorithm converges superlinearly and is computationally efficient. Numerical examples show that the computational cost of our algorithm remains low for even large data sets. Thus, the proposed methods are applicable to large-scale data for which the application of existing methods is impractical or fails completely. The methods are applied to national kidney transplant data and study the impact of potential risk factors on post-transplant survival.

email: **kevinhe@umich.edu**

MULTIPLE IMPUTATION FOR INTERVAL CENSORED DATA WITH TIME-DEPENDENT AUXILIARY VARIABLES USING INCIDENT AND PREVALENT COHORT DATA

Wen Ye*, University of Michigan

Douglas Schaubel, University of Michigan

Due to the rarity of the disease and high transplant rate in biliary atresia patients, the true incidence of portal hypertension (PHT) related complications in the absence of liver transplantation is unknown. Motivated by the need to understand the true clinical burden of PHT in these patients, we developed

a risk score based multiple imputation method to combine data from incident and prevalent cohort studies to overcome data scarcity; and to recover interval-censored and dependently right censored time of clinical PHT onset. The risk scores used for imputation are calculated using BLUP estimates of the random effects, derived from prognostic factors and based on longitudinal models for the time-varying prognostic factors. To incorporate the full uncertainty in the imputes, we included a bootstrap procedure and by replacing the BLUP estimates with draws from their posterior distributions. In addition, we used weighted Kaplan-Meier estimator to adjust for survival-selection in the prevalent component of the sample.

email: **wye@umich.edu**

MODEL FLEXIBILITY FOR REGRESSION ANALYSIS OF SURVIVAL DATA WITH INFORMATIVE INTERVAL CENSORING

Tyler Cook*, University of Missouri, Columbia

Jianguo Sun, University of Missouri, Columbia

One problem that researchers face when analyzing survival data is how to handle the censoring distribution. It is often assumed that the observation process generating the censoring is independent of the event time of interest and can then effectively be ignored, but this assumption is clearly not always realistic. Unfortunately one cannot generally test for independent censoring without additional assumptions or information. Therefore, the researcher is faced with a choice between using methods designed for informative or noninformative censoring. This project investigates the effectiveness of two methods developed for the analysis of informative case I and case II interval censored data under both types of censoring. Extensive simulation studies indicate that the methods produce unbiased results in the presence of both informative and noninformative censoring. The efficiency of the informative censoring methods is then compared with approaches created to handle noninformative censoring. The results of these simulation studies can provide guidelines for deciding between models when facing a practical problem where one is unsure about the dependence of the censoring distribution.

email: **tlcm89@mail.missouri.edu**

61. ORAL POSTERS: GWAS and META Analysis of Genetic Studies

61a. HYPOTHESIS TESTING FOR SPARSE SIGNALS IN GENETIC ASSOCIATION STUDIES

Xihong Lin*, Harvard University

email: **xlin@hsph.harvard.edu**

61b. META-ANALYSIS OF GENE-ENVIRONMENT INTERACTION IN CASE-CONTROL STUDIES BY ADAPTIVELY USING GENE-ENVIRONMENT CORRELATION

Bhramar Mukherjee*, University of Michigan

Shi Li, University of Michigan

John D. Rice, University of Michigan

Jeremy M. G. Taylor, University of Michigan

Heather Stringham, University of Michigan

Michael L. Boehnke, University of Michigan

There has been a significant volume of literature on using gene-environment (G-E) independence to enhance power for testing gene- environment interaction (GEI) in case-control studies. However, there is little work thus far to study the role of G-E independence in a meta-analysis setting where the assumption could vary across studies. In this paper, we propose an appropriate adaptation of the empirical-Bayes (EB) type shrinkage estimator previously proposed by Mukherjee and Chatterjee (2008) to a meta-analysis context. The retrospective likelihood framework for inference is used to derive an adaptive combination of estimators obtained under the constrained model (assuming G-E independence) and unconstrained model (without any assumptions) with weights determined by using information on G-E association parameters derived from multiple studies/cohorts. Our simulation studies indicate that this newly proposed estimator has improved mean-squared-error (MSE) properties than the standard alternative

of using the inverse variance weighted estimator that combines study-specific constrained, unconstrained or EB estimators. The results were illustrated by analyzing data from a study of Type 2 diabetes, with six different case-control studies contributing to the meta-analysis. We considered the interaction between genetic markers on the obesity related FTO gene and environmental factors with Type 2 diabetes as the case-control outcome of interest

email: **bhramar@umich.edu**

61c. PARTIAL LINEAR VARYING INDEX COEFFICIENT MODEL FOR GENE-ENVIRONMENT INTERACTIONS

Xu Liu*, Michigan State University

Yuehua Cui, Michigan State University

Gene-environment interactions play key roles in many complex diseases. In this paper, we propose a partial linear varying index coefficient model (PLVICM) to assess how multiple environmental factors acting jointly to modify individual genetic risk on complex disease. Our model is generalized from varying index coefficient model while discrete variables are admitted as the linear part. Therefore, PLVICM allows us to study the nonlinear interaction between grouped continues environments and genes as well as the interaction between the linear form of discrete environments and genes simultaneously. We derive a profile method to estimate parametric parameters and a B-spline backfitted kernel method to estimate nonlinear functions. The consistency and asymptotic normality properties of parametric and nonparametric estimates are established under some regularity conditions. Hypothesis testing for the parametric coefficients and nonparametric functions are conducted. Results show that the statistics of testing parametric coefficients are asymptotically Chi-squared distributed, and the statistics of testing nonparametric functions approximately follow a Chi-squared distribution. The utility of the method is demonstrated through extensive simulations and a case study.

email: **xuliu@stt.msu.edu**

61d. TREE-BASED MODEL AVERAGING APPROACHES FOR MODELING RARE VARIANT ASSOCIATION IN CASE-CONTROL STUDIES

Brandon J. Coombes*, University of Minnesota

Saonli Basu, University of Minnesota

Sharmistha Guha, Fair Isaac Corporation

Nicholas Schork, J. Craig Venter Institute

Multi-locus effect modeling is a powerful approach for detection of genes influencing a complex disease. Especially for rare-variants, we need to analyze multiple variants together to achieve adequate power for detection. In this paper, we propose a parsimonious tree model and several branching model mechanisms to assess the joint effect of a group of rare variants on a binary trait in a case-control study. The tree model implements a data reduction strategy within a likelihood framework and all approaches use a weighted score test to assess the statistical significance of the effect of the group of variants on the disease. The primary advantage of the proposed model averaging approaches is that it performs model averaging over a substantially smaller set of models supported by the data and thus gains power to detect multi-locus effects. We illustrate the proposed model on simulated and real data, study the performance of these model-averaging approaches compared to the model selection method proposed by Basu and Pan (2011). Extensive simulations and real data application demonstrate the advantage the proposed approach in presence of moderate number of null variants and presence of linkage equilibrium among the variants.

email: **coom0054@umn.edu**

61e. A FUNCTIONAL APPROACH TO ASSOCIATION TESTING OF MULTIPLE PHENOTYPES IN SEQUENCING STUDIES

Sneha Jadhav*, Michigan State University

Qing Lu, Michigan State University

Sequencing-based association studies are proving to be increasingly useful in genetic research of complex diseases. In many of these studies, multiple phenotypes are collected. These phenotypes can be different measurements of an underlying disease, or measurements characterizing multiple diseases for studying common genetic mechanism (e.g., pleiotropic effects). Multiple phenotypes and high-dimensionality of the sequencing data pose challenges for their association studies. To address these challenges, we propose a non-

parametric method, which first constructs smooth functions from individuals sequencing data, and then uses these to construct a U statistic for testing of association. The proposed method has the advantages of providing a general framework of analyzing various types of phenotypes, considering linkage disequilibrium between genetic markers, and allowing for different directions and magnitude of effects. Through preliminary simulation study, we found it had comparable performance to existing methods when the distribution assumptions of existing methods hold. Nevertheless, it outperformed the existing methods when their distribution assumptions were violated.

email: **jadhavsn@stt.msu.edu**

61f. 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

Dan-Yu Lin, University of North Carolina, Chapel Hill

High-throughput DNA sequencing allows for the genotyping of common and rare variants for genetic association studies. At the present time and for the foreseeable future, it is not economically feasible to sequence all individuals in a large cohort. A cost-effective strategy is to sequence those individuals with extreme values of a quantitative trait. We consider the design under which the sampling depends on multiple quantitative traits. Under such trait-dependent sampling, standard linear regression analysis can result in bias of parameter estimation, inflation of type 1 error, and loss of power. We construct a likelihood function that properly reflects the sampling mechanism and utilizes all available data. We implement a computationally efficient EM algorithm and establish the theoretical properties of the resulting maximum likelihood estimators. Our methods can be used to perform separate inference on each trait or simultaneous inference on multiple traits. We pay special attention to gene-level association tests for rare variants. We demonstrate the superiority of the proposed methods over standard linear regression through extensive simulation studies. We provide applications to the Cohorts for Heart and Aging Research in Genomic Epidemiology Targeted Sequencing Study and the National Heart, Lung, and Blood Institute Exome Sequencing Project.

email: **dragontaoran@gmail.com**

61g. META-ANALYSIS OF COMPLEX DISEASES AT GENE LEVEL BY GENERALIZED FUNCTIONAL LINEAR MODELS

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

Haobo Ren, Regeneron Pharmaceuticals, Inc.

Yun Li, University of North Carolina, Chapel Hill

Christopher Amos, Dartmouth Medical School

Wei Chen, University of Pittsburgh

Momiao Xiong, University of Texas, Houston

Jason Moore, Dartmouth Medical School

Generalized functional linear models (GFLMs) are developed to perform a meta-analysis of multiple case-control studies to connect genetic data to dichotomous traits adjusting for covariates. Based on the GFLMs, $\chi ^ 2$-distributed Rao's efficient score test and likelihood ratio test (LRT) statistics are introduced to test for an association between a complex trait and multiple genetic variants in one genetic region. Extensive simulations are performed to evaluate empirical type I error rates and power performance of the proposed models and tests. The proposed Rao's efficient score test statistics control the type I error very well and have higher power than the existing methods of

MetaSKAT when the causal variants are both rare and common. When the causal variants are all rare (i.e., minor allele frequencies less than 0.03), the Rao's efficient score test statistics have similar or slightly lower power than MetaSKAT. The LRT statistics generate accurate type I error rates for homogeneous genetic effect models and may inflate type I error rates for heterogeneous genetic models due to big degrees of freedom, and have similar or slightly higher power than the Rao's efficient score test statistics. The proposed methods were applied to analyze type 2 diabetes data from a meta-analysis of eight European studies, and detected significant association for genes APB, APOE, FTO, and LPL, while MetaSKAT detected none. The models and related test statistics can analyze rare variants or common variants or a combination of the two, and can be useful in the whole genome-wide and whole exome association studies.

email: **fanr@mail.nih.gov**

61h. GENE LEVEL META-ANALYSIS OF QUANTITATIVE TRAITS BY FUNCTIONAL LINEAR MODELS

Yifan Wang*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Ruzong Fan, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Michael Boehnke, University of Michigan

Wei Chen, University of Pittsburgh

Yun Li, University of North Carolina, Chapel Hill

Momiao Xiong, University of Texas, Houston

Functional linear models are developed for meta-analysis of multiple studies to connect genetic data to quantitative traits adjusting for covariates. The models can analyze rare variants or common variants or the combinations of the two. Both likelihood ratio test (LRT) and F-distributed statistics are introduced to test association between quantitative traits and multiple genetic variants in one genetic region. Extensive simulations are performed to evaluate empirical type I error rates and power performance of the proposed models and tests. We show that the proposed LRT and F-distributed statistics control the type I error very well and have higher power than the existing methods of MetaSKAT. The proposed methods were applied to analyze four blood lipid levels in data from meta-analysis of eight European studies. It was found that the proposed methods detect more significant association than MetaSKAT and the p-values of the proposed LRT and F-distributed statistics are usually much smaller than those of MetaSKAT. The functional linear models and related test statistics can be useful in the whole genome-wide and whole exome association studies.

email: **yifan.wang@nih.gov**

61i. A NEW ESTIMATING EQUATION APPROACH FOR SECONDARY TRAIT ANALYSES IN GENETIC CASE-CONTROL STUDIES

Xiaoyu Song*, Columbia University

Iuliana Ionita-Laza, Columbia University

Ying Wei, Columbia University

In this manuscript, we propose a new estimating equation based approach that provides unbiased secondary traits analysis in genetic case-control studies. In genetic studies, analysis on secondary traits is an important way to discover potential disease pathways. When data are collected from case-control designs, direct analyses are often biased. Several methods have been proposed to address this issue, including the inverse-probability-of-sampling-weighted (IPW) approach, the maximum likelihood (ML) approach, the adaptive weighted approach and the bias correction approach. Comparing to the existing ones, the proposed estimating equation based approach enjoys the following properties. One, it creates a general framework that is applicable to a wide range of genetic models. It could be used to model various types phenotypes (continuous or binary) and SNPs (additive or dominate, single or multiple), and is also easy to incorporate covariates. Second, it is computationally simple and straightforward. We compared our method with the existing ones in both numerical studies and a stroke GWAS data. The proposed method was shown to be less sensitive to the sampling scheme and underlying disease model. For these reasons, we believe that our new methods complement the existing approaches, and are useful to analyze secondary traits.

email: **xs2148@columbia.edu**

61j. NOVEL STATISTICAL MODEL FOR GWAS META-ANALYSIS AND ITS APPLICATION TO TRANS-ETHIC META-ANALYSIS

Jingchunzi Shi*, University of Michigan

Seunggeun Lee, University of Michigan

Trans-ethnic Genome-wide association studies (GWAS) meta-analysis has proven to be a practical and profitable approach for identifying loci which contribute to the missing heritability of complex traits. However, the expected genetic effects heterogeneity cannot be easily accommodated through existing approaches. In response, we propose a novel trans-ethnic meta-analysis methodology with flexibly modeling of the expected genetic effects heterogeneity across diverse populations. Specifically, we consider a modified random effect model in which genetic effect coefficients are random variables whose correlation structure across ancestry groups reflects the expected heterogeneity (or homogeneity) among ancestry groups. To test for associations, we derive the data-adaptive variance component test with adaptive selection of the correlation structure to increase the power. Simulations demonstrate that our proposed method performs with substantial improvements in comparison to the traditional meta-analysis methods. Furthermore, our proposed method provides scalable computing time for genome-wide data. For real data analysis, we re-analyzed the published type 2 diabetes GWAS meta-analyses from Consortium et al. (2014), and successfully identified one additional SNP which clearly exhibits genetic effects heterogeneity among different ancestry groups but could not be detected by the traditional meta-analysis methods.

email: **shijingc@umich.edu**

61k. MULTIPLE PHENOTYPE ASSOCIATION TESTING BASED ON SUMMARY STATISTICS IN GENOME-WIDE ASSOCIATION STUDIES

Zhonghua Liu*, Harvard School of Public Health

Xihong Lin, Harvard School of Public Health

Multiple correlated phenotypes might share a common genetic basis, referred to as pleiotropy in genetics. However, current available methods for identifying genetic variants with pleiotropic effects on multiple phenotypes are limited. In this paper, we present a toolkit of statistical methods that harness the correlation structures among the multiple phenotypes to boost statistical powers to detect such genetic variants based on summary statistics. We conduct extensive simulation studies to show that our methods maintain correct type I error rates, and their statistical powers are compared in a wide range of situations. We further apply these methods to a genome-wide association study of plasma lipids levels and identify hundreds of novel genetic variants that conventional single-trait analysis approaches failed to discover. We also develop an R package MPAT available for public uses.

email: **zliu@mail.harvard.edu**

61l. A NEW APPROACH FOR DETECTING GENE-BY-GENE INTERACTIONS THROUGH META-ANALYSES

Yulun Liu*, University of Texas, Health Science Center at Houston

Paul Scheet, University of Texas MD Anderson Cancer Center

Yong Chen, University of Texas, Health Science Center at Houston

There is increasing interest in detecting gene-by-gene interactions for complex traits, with varying, but substantial, proportions of heritability remaining unexplained by surveys of single-SNP genetic variation. The major challenges from traditional regression-based methods are the large number of possible pairs under investigation, with a requisite need to correct for multiple testing, and the restrictive assumptions of large marginal effects to reduce the search space and limit the number of tests. Both of these challenges may limit power, especially when the marginal effects are in fact modest. In this talk, we propose a new procedure for detecting gene-by-gene interactions through meta-analyses. Our approach is pragmatic when data-sharing limitations restrict mega-analyses. It is also computationally efficient in that it applies a dimension reduction procedure and thus may scale for higher-order interactions as well. We compare the type I error and power of our proposed procedure relative to existing methods and evaluate their strengths and limitations.

email: **Yulun.Liu@uth.tmc.edu**

61m. GENOME-WIDE ASSOCIATION STUDIES FOR FUNCTIONAL VALUED TRAITS

Han Hao*, The Pennsylvania State University

Rongling Wu, The Pennsylvania State University

Genome-wide association studies (GWAS) has been widely used to detect the association between genetic variations and phenotypic variations. A great number of GWAS approaches have been developed in the past decade, but few of them were designed for functional trait values. Functional trait values are widely seen in biological shape analysis, dynamic progressions and clinical trials, and it is crucial to integrate the functional feature with GWAS and receive high statistical power. We here propose a model-free approach to address a GWAS problem with functional trait values. There is no assumption for the functional form, but a parametric form can be involved to account for specific biological mechanism. The method is applied on a real dataset and verified to be quite powerful.

email: **haohan421@gmail.com**

61n. KERNEL-BASED TESTING FOR NONLINEAR EFFECT OF A SNP-SET UNDER MULTIPLE CANDIDATE KERNELS

Tao He*, Michigan State University

Ping-Shou Zhong, Michigan State University

Yuehua Cui, Michigan State University

Kernel-based testing framework has been proved very powerful in SNP-set association analysis by measuring the similarity between genotypes through a fixed kernel function and comparing it to the phenotype similarity. However, given a set of kernel candidates, there is no general criterion to construct weighted kernel, which has more flexibility than single kernel. Based on the asymptotic results, we proposed a weighted kernel strategy where the weights were optimized to maximize the signal-to-noise ratio of the weighted kernel. The proposed method was demonstrated through simulations and real data applications.

email: **hetao@stt.msu.edu**

61o. A GENERAL FRAMEWORK OF GENE-BASED ASSOCIATION TESTS FOR CORRELATED CASE-CONTROL SAMPLES

Han Chen*, Harvard School of Public Health

Chaolong Wang, Harvard School of Public Health

Xihong Lin, Harvard School of Public Health

In genetic association studies, gene-based tests have been widely used to test association with a set of genetic variants, genes or pathways. However, existing gene-based tests such as burden tests and the sequence kernel association test require the critical assumption that observations are independent, which is violated in the presence of population stratification and cryptic relatedness. We observe inflated type I error rates when using these tests to analyze correlated samples. Here we propose a general framework of gene-based tests for correlated case-control samples, which degenerates into corresponding tests for independent samples in the absence of population structure or relatedness. We fit a generalized linear mixed model under the null hypothesis and derive the test statistics and their asymptotic distributions. We show in simulation studies that our tests have correct type I error rates in correlated samples, in contrast to those tests assuming independence. We compare the power of our tests in various scenarios and illustrate how they could be used to test different scientific hypotheses. We also apply our tests to a real data example.

email: hanchen@hsph.harvard.edu

61p. ALGORITHM TO COMPUTE THE IDENTITY COEFFICIENTS AT A PARTICULAR LOCUS GIVEN THE MARKER INFORMATION

J. Concepcion Loredo-Osti*, Memorial University

Haiyan Yang, Memorial University

There are some problems in modern genetics where the inferring the identity coefficients or a linear combination of them a particular locus given a the data on a set of markers may play an important role. For example, if the identity coefficients at a given chromosomal location using the marker information spanning a the region of interest have already been estimated, the identity by descent status used in gene mapping problems can be easily obtained. It would also be possible to compute these identity by descent coefficients conditional on a particular model which be useful in addressing gene genealogy problems,

modelling jointly linkage and linkage disequilibrium, genetic counselling and other forensic applications. On this presentation, an extension to Karigl (1981), Abney(2009) or Cheng-Ozsoyoglu (2014) algorithms for computing the identity coefficients that incorporates the marker information is introduced. A comparison with other procedures in the context of identity by descent estimation is also discussed.

email: **jcloredoosti@mun.ca**

61q. ESTIMATING THE EMPIRICAL NULL DISTRIBUTION OF MAXMEAN STATISTICS IN GENE SET ANALYSIS

Xing Ren*, University at Buffalo, SUNY

Jeffrey Miecznikowski, University at Buffalo, SUNY

Song Liu, Roswell Park Cancer Institute

Jianmin Wang, Roswell Park Cancer Institute

Gene set analysis is a widely-used framework for testing enrichment of differentially expressed genes in a set of genes. The method involves computing a maxmean statistic and estimating the null distribution of the maxmean statistics via a restandardization procedure. We derive an asymptotic null distribution of the maxmean statistic and propose an empirical method to estimate the empirical null distribution. We show that our method is more accurate in controlling the type 1 error when testing a large number of gene sets.

email: **xingren@buffalo.edu**

61r. USAT: A UNIFIED SCORE-BASED ASSOCIATION TEST FOR MULTIPLE PHENOTYPE-GENOTYPE ANALYSIS

Debashree Ray*, University of Minnesota

Saonli Basu, University of Minnesota

Genome-wide Association Studies (GWASs) for complex diseases often collect data on multiple correlated endophenotypes. Multivariate analysis of these correlated phenotypes can improve the power to detect genetic variants. Multivariate analysis of variance (MANOVA) can perform such association analysis at a GWAS level, but the behavior of MANOVA under different trait models have not been carefully investigated. In this paper, we show that MANOVA is generally very powerful for detecting association but there are situations where MANOVA may not have any detection power. We investigate the behavior of MANOVA, both theoretically and using simulations, and derive conditions where MANOVA loses power. Based on our findings, we propose a unified score-based test USAT that can perform better than MANOVA in such situations and do almost as good as MANOVA elsewhere. USAT reports an approximate asymptotic p-value for association and is computationally efficient at GWAS level. We have studied through extensive simulations the performance of USAT, MANOVA and other existing approaches and demonstrated the advantage of using USAT in detecting association between a genetic variant and multivariate phenotypes. We applied USAT on ARIC type 2 diabetes data with five correlated traits on 5,819 Caucasians and detected some significantly associated novel genetic variants.

email: **rayxx267@umn.edu**

62. Statistical Inference with Random Forests and Related Ensemble Methods

CONSISTENCY OF RANDOM FORESTS

Gerard Biau*, Pierre and Marie Curie University

Erwan Scornet, Pierre and Marie Curie University

Jean-Philippe Vert, Pierre and Marie Curie University

Random forests are a learning algorithm proposed by L. Breiman in 2001 which combines several randomized decision trees and aggregates their predictions by averaging. Despite its wide usage and outstanding practical performance, little is known about the mathematical properties of the procedure. This disparity between theory and practice originates in the difficulty to simultaneously analyze both the randomization process and the highly data-dependent tree structure. In this talk, we take a step forward in forest exploration by proving a consistency result for Breiman's original algorithm in the context of additive models. Our analysis also sheds an interesting light on how random forests can nicely adapt to sparsity in high-dimensional settings.

email: **gerard.biau@upmc.fr**

ASYMPTOTIC THEORY FOR RANDOM FORESTS

Stefan Wager*, Stanford University

Random forests have proven themselves to be reliable predictive algorithms in many application areas. Not much is known, however, about the statistical properties of random forests. Several authors have established conditions under which their predictions are consistent, but these results do not provide practical estimates of the scale of random forest errors. In this paper, we analyze a random forest model based subsampling, and show that random forest predictions are asymptotically normal provided that the subsample size s scales as s(n)/n = o(log(n) ^ {-d}), where n is the number of training examples and d is the number of features. Moreover, we show that the asymptotic variance can consistently be estimated using an infinitesimal jackknife for bagged ensembles recently proposed by Efron (2013). In other words, our results let us both characterize and estimate the error-distribution of random forest predictions. Thus, random forests need not only be treated as black-box predictive algorithms, and can also be used for statistical inference.

email: **swager@stanford.edu**

DETECTING FEATURE INTERACTIONS IN BAGGED TREES AND RANDOM FORESTS

Lucas K. Mentch*, Cornell University

Giles Hooker, Cornell University

Additive models remain popular statistical tools due to their ease of interpretation and as a result, hypothesis tests for additivity have been developed to determine the appropriateness of such models. However, as data grows in size and complexity, practitioners are relying more heavily on learning algorithms because of their predictive superiority. Due to the black-box nature of these learning methods, the increase in predictive power is assumed to come at the cost of interpretability and understanding. In this talk, we discuss our recent work that demonstrates that many popular learning algorithms, such as bagged trees and random forests, have desirable asymptotic properties. In particular, we produce a central limit theorem for predictions when base learners are built with subsamples, thereby allowing for statistical inference. In addition to producing confidence intervals and hypothesis tests for feature significance, we show that by enforcing a grid structure on the test set, we can formally test the plausibility of various additive structures. We develop notions of total and partial additivity and demonstrate that both tests can be carried out at no additional computational cost to the original ensemble.

email: **lucasmentch@gmail.com**

VARIABLE SELECTION WITH BAYESIAN ADDITIVE REGRESSION TREES

Shane T. Jensen*, University of Pennsylvania

Justin Bleich, University of Pennsylvania

Adam Kapelner, University of Pennsylvania

Edward I. George, University of Pennsylvania

There is a crucial need for effective variable selection procedures in high dimensional data, where it is difficult to detect subtle individual effects and interactions between factors. Bayesian Additive Regression Trees are a promising alternative to more parametric regression approaches, such as the lasso or Bayesian latent indicator models. BART constructs an ensemble of decision trees from the set of possible predictors of an outcome variable. We develop principled methodology that adapts BART to variable selection as well as incorporating additional data as prior information. We evaluate the performance of our BART-based approach in simulation settings as well as an application to the gene regulatory network in yeast.

email: **stjensen@wharton.upenn.edu**

63. Mediation and Interaction: Theory, Practice and Future Directions

A UNIFICATION OF MEDIATION AND INTERACTION: A FOUR-WAY DECOMPOSITION

Tyler J. VanderWeele*, Harvard University

It is shown that the overall effect of an exposure on an outcome, in the presence of a mediator with which the exposure may interact, can be decomposed into four components: (i) the effect of the exposure in the absence of the mediator, (ii) the interactive effect when the mediator is left to what it would be in the absence of exposure, (iii) a mediated interaction, and (iv) a pure mediated

effect. These four components, respectively, correspond to the portion of the effect that is due to neither mediation nor interaction, to just interaction (but not mediation), to both mediation and interaction, and to just mediation (but not interaction). This four-way decomposition unites methods that attribute effects to interactions and methods that assess mediation. Certain combinations of these four components correspond to measures for mediation, while other combinations correspond to measures of interaction previously proposed in the literature. Prior decompositions in the literature are in essence special cases of this four-way decomposition. The four-way decomposition can be carried out using standard statistical models, and software is provided to estimate each of the four components. The four-way decomposition provides maximum insight into how much of an effect is mediated, how much is due to interaction, how much is due to both mediation and interaction together, and how much is due to neither.

email: **tvanderw@hsph.harvard.edu**

PARTIAL IDENTIFICATION OF THE PURE DIRECT EFFECT UNDER EXPOSURE-INDUCED CONFOUNDING

Caleb Miles*, Harvard University

Eric Tchetgen Tchetgen, Harvard University

In causal mediation analysis, nonparametric identification of the pure (natural) direct effect typically relies on fundamental assumptions of (i) so-called "cross-world-counterfactuals" independence and (ii) no exposure-induced confounding. When the mediator is binary, bounds for partial identification have been given when neither assumption is made, or alternatively when assuming only (ii). We extend these bounds to the case of a polytomous mediator, and provide bounds for the case assuming only (i). We apply these bounds as well as point estimates under other fully-identifying model assumptions to data from the Harvard PEPFAR program in Nigeria, where we evaluate the extent to which the effects of antiretroviral therapy on virological failure are mediated by a patient's adherence, and show that inference on this effect is somewhat sensitive to model assumptions.

email: **calebhmiles@gmail.com**

INTEGRATIVE ANALYSIS OF COMPLEX GENETIC, GENOMIC AND ENVIRONMENTAL DATA USING MEDIATION ANALYSIS

Xihong Lin*, Harvard University

Mediation analysis provides a useful framework for integrative analysis of multiple types of genetic and genomic data and environmental data to understand disease causing mechanisms. Genetic and genomic data include SNP data, such as GWAS or sequencing data, and gene expression data. We discuss in this talk mediation analysis in several complex settings, including the presence of missing data and network analysis. Specifically, GWAS data are often collected on all individuals enrolled in a study. However, genomic data, such as gene expressions and DNA methylations, are often collected in a subset of study subjects. We propose a mediation analysis method using all the data by leveraging the information from the individuals with only the SNP data. We show using all available data, we gained more efficient estimators of the direct effects of SNPs and the indirect effects of SNPs mediated through gene expressions/DNA methylations on a phenotype with varying levels of missingness. We also consider mediation network analysis, where the mediator consists of network data. We applied our method to several existing datasets.

email: **xlin@hsph.harvard.edu**

64. Motivation and Analysis Strategies for Joint Modeling of High Dimensional Data in Genetic Association Studies

REGION-BASED TEST FOR GENE-ENVIRONMENT INTERACTIONS IN LONGITUDINAL STUDIES

Zihuai He, University of Michigan

Min Zhang*, University of Michigan

Seunggeun Lee, University of Michigan

Jennifer Smith, University of Michigan

Xiuqing Guo, Harbor-UCLA Medical Center

Walter Palmas, Columbia University

Sharon L.R. Kardia, University of Michigan

Ana V. Diez Roux, University of Michigan

Bhramar Mukherjee, University of Michigan

There has been tremendous emphasis on searching for interactions between genetic factors and environmental

exposures. Gene-environment interactions (G×E) are typically based on testing the interaction between each single-nucleotide polymorphisms (SNP) and an environmental variable separately, with adjustment for multiple testing. However, the interaction process is probably far more complex than looking for "single locus vs. environment factor" analysis. We propose a novel statistical approach to test for gene-environment interaction between an environmental factor and a set of genetic variants for longitudinal studies, with the consideration of potential time dependency and correlation in the outcomes measured on the same subject. The method integrates the entire genotype-environment-phenotype information contained in a longitudinal study through a region based test. Nonparametric modeling of the environmental exposure is incorporated to alleviate the problem of misspecification of the main or interaction effect, leading to more robust type-I error rate and superior power. As the number of SNPs in a target region can be very large, dimension reduction method is further proposed, which adaptively selects and adjusts for the main effect of genetic variants to achieve numerical feasibility, controlled type I error probability and improvement in power. The performance of the method will be evaluated through simulation studies and illustrated by real data analysis. Survival data collected from prevalent cohorts are subject to left-truncation. The conventional conditional approach using Cox model disregards the information in the marginal likelihood of truncation time thus can be inefficient. On the other hand, the stationary assumption under length-biased sampling (LBS) methods to incorporate the marginal information can lead to biased estimation when it is violated. In this paper, we propose a semiparametric estimation method by augmenting the Cox partial likelihood with a pairwise likelihood, by which we eliminate the unspecified truncation distribution in the marginal likelihood, yet retain the information about regression coefficients and the baseline hazard. Exploring self-consistency of the estimator, we give a fast algorithm to solve for the regression coefficients and the cumulative hazard simultaneously. The proposed estimator is shown to be consistent and asymptotically normal with a sandwich-type consistent variance estimator. Simulation studies show a substantial efficiency gain in both the regression coefficients and the cumulative hazard over Cox model estimators, and that the gain is comparable to LBS methods when the stationary assumption holds. For illustration, we apply the proposed method to the RRI-CKD data.

email: **mzhangst@umich.edu**

STRATEGIES TO IMPROVE THE POWER OF PATHWAY ANALYSIS IN GENETIC ASSOCIATION STUDIES

Kai Yu*, National Cancer Institute, National Institutes of Health

Han Zhang, National Cancer Institute, National Institutes of Health

Jianxin Shi, National Cancer Institute, National Institutes of Health

Nilanjan Chatterjee, National Cancer Institute, National Institutes of Health

It is increasingly recognized that pathway analyses-a joint test of association between the outcome and a group of single nucleotide polymorphisms (SNPs) within a biological pathway-could complement single-SNP analysis and provide additional insights for the genetic architecture of complex diseases. In this talk, we will explore several strategies for enhancing the power of pathway analysis, including the improvement of the signal-to-noise ratio through more informative selection of SNPs based on their potential functional impact, and the increase of sample size by integrating summary statistics on individual SNPs from existing large scale Meta analysis for pathway analysis. We will use numerical simulations and real data applications to evaluate the proposed procedures.

email: **yuka@mail.nih.gov**

A UNIFIED TEST FOR POPULATION-BASED MULTIPLE CORRELATED PHENOTYPE-GENOTYPE ASSOCIATION ANALYSIS

Saonli Basu*, University of Minnesota

Debashree Ray, University of Minnesota

Joint modeling of a disease-related multiple correlated traits may improve power to detect association between a genetic variant and the disease. Moreover this joint analysis can reveal some pleiotropic genes involved in the biological development of the disease. The standard multivariate analysis of variance (MANOVA) is very powerful when a genetic variant is associated with a subset of the traits or the effect of causal variant is in different directions with these correlated traits, but loses significant power when the variant is associated with all the traits and the effect direction is same as the direction of dependence between the traits. We propose a power-

ful computationally efficient unified test that maximizes power by adaptively using the data to optimally combine MANOVA and a test that potentially ignore the correlation between the traits. We illustrate our proposed test through simulation studies and real data applications and compare the performance of different multivariate approaches under various alternative models.

email: **saonli@umn.edu**

MODELLING MULTIPLE CORRELATED GENETIC VARIANTS

Sharon R. Browning*, University of Washington

Many statistical analyses of genetic data rely on being able to model the correlation between genetic variants that are located close together on a chromosome. The processes that create the correlation are complex, and include mutation, recombination, selection, and drift. These factors have variable effects across the genome, so the strength and patterns of correlation are also variable from one genomic region to another. Hence successful modeling efforts need to be data driven, as well as incorporating key elements of genetic processes such as recombination. I will describe the Beagle model, which has proved to be useful for a variety of statistical analyses, including haplotype phase inference and imputation of untyped variants. I will also present recent work applying this model to identity by descent tract detection.

email: **sguy@uw.edu**

65. Recent Developments on Inference for Possibly Time-Dependent Treatment Effects with Survival Data

THRESHOLD REGRESSION FOR LIFETIME DATA

Mei-Ling Ting Lee*, University of Maryland, College Park

George A. Whitmore, McGill University, Canada

Cox regression methods are well-known. It has, however, a strong proportional hazards assumption. In many medical contexts, a disease progresses until a failure event (such as death) is triggered when the latent health level first degrades to a failure threshold. I will present the Threshold Regression (TR) model for the health process that requires few assumptions and, hence, is quite general in its potential application. I will begin with Wiener diffusion process based TR model and the regression methods using inverse-Gaussian distributions. Distribution-free methods for estimations and predictions using the TR models will also be derived. I will demonstrate the methodology and its practical use. Comparisons with the Cox model will also be discussed.

email: **mltlee@umd.edu**

HYPOTHESIS TESTING FOR AN EXTENDED COX MODEL WITH TIME-VARYING COEFFICIENTS

Ying Q. Chen*, Fred Hutchinson Cancer Research Center

The log-rank test has widely been used to test treatment effects under the Cox model for censored time-to-event outcomes, though it may lose power substantially when the model's proportionality assumption does not hold. In this paper, we consider an extended Cox model that uses B-splies or smoothing splines to model a time-varying treatment effect and propose score test statistics for detecting the treatment efffect. The new methods are applied to a randomized clinical trial assessing the efficacy of single-dose Nevirapine against mother-to-child HIV transmission that was conducted by the HIV Prevention Trial Network.

email: **yqchen@fhcrc.org**

TIME-DEPENDENT CUT POINT SELECTION FOR BIOMARKERS IN CENSORED SURVIVAL DATA

Zhezhen Jin*, Columbia University

In biomedical research and practice, continuous biomarkers are often used for diagnosis and prognosis, with cut points being established to monitor treatment effect on survival or time to an event. We will study non-parametric procedure for the selection of time-dependent cut points with censored survival data. Numerical studies will be presented along with real applications.

email: **zj7@columbia.edu**

66. *Journal of Agricultural, Biological and Environmental Statistics* (JABES) Highlights

INFERENCE ON THE SUMMARY MEASURES OF TREATMENT EFFECT WITH SURVIVAL DATA WHEN THERE IS POSSIBLY TREATMENT BY TIME INTERACTION

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

For clinical trials with survival data, the hazard ratio has been the most widely used measure for describing the treatment effect. When there is possibly a treatment by time interaction, summary measures such as average hazard ratio and restricted mean survival difference have been proposed in the literature. We investigate various old and new summary measures, and study their nonparametric and semiparametric estimates, with or without covariate adjustment. Hypothesis testing and confidence intervals of the measures are established. We illustrate these measures and discuss their merits and limitations in applications to clinical trials including the Women's Health Initiative.

email: **yangso@nhlbi.nih.gov**

LIMITED-INFORMATION MODELING OF LOGGERHEAD TURTLE POPULATION SIZE

John M. Grego*, University of South Carolina

David B. Hitchcock, University of South Carolina

In traditional capture-recapture experiments to estimate the size of an animal population, individual animals are tagged and the information about which individuals are captured repeatedly is crucial. We apply these approaches to data in which information about individual identity is not available, specifically nesting data for loggerhead turtles. Rather, we observe only the counts of successful and failed nestings at a location over a series of days. We view the turtles' nesting behavior as an alternating renewal process, model it using parametric distributions, and then derive probability distributions that describe the behavior of the turtles during the days under study. We adopt a Bayesian approach, formulating our model in terms of parameters about which strong prior information is available. We use a Gibbs sampling algorithm to sample from the posterior distribution of our random quantities, the most crucial of which is the number of turtles remaining offshore during the entire sampling period. We illustrate the method using data sets from loggerhead turtle sites along the South Carolina coast.

email: **grego@stat.sc.edu**

NONLINEAR VARYING-COEFFICIENT MODELS WITH APPLICATIONS TO A PHOTOSYNTHESIS STUDY

Damla Senturk*, University of California, Los Angeles

Esra Kurum, Medeniyet University

Runze Li, The Pennsylvania State University

Yang Wang, China Vanke

Motivated by a study on factors affecting the level of photosynthetic activity in a natural ecosystem, we propose nonlinear varying-coefficient models, in which the relationship between the predictors and the response variable is allowed to be nonlinear. One-step local linear estimators are developed for the nonlinear varying-coefficient models and their asymptotic normality is established leading to point-wise asymptotic confidence bands for the coefficient functions. Two-step local linear estimators are also proposed for cases where the varying-coefficient functions admit different degrees of smoothness; bootstrap confidence intervals are utilized for inference based on the two-step estimators. We further propose a generalized F-test to study whether the coefficient functions vary over a covariate. We illustrate the proposed methodology via an application to an ecology data set and study the finite sample performance by Monte Carlo simulation studies.

email: **dsenturk@ucla.edu**

MULTILEVEL LATENT GAUSSIAN PROCESS MODEL FOR MIXED DISCRETE AND CONTINUOUS MULTIVARIATE RESPONSE DATA

Erin M. Schliep*, Duke University

Jennifer A. Hoeting, Colorado State University

We propose a Bayesian model for mixed ordinal and continuous multivariate data to evaluate a latent spatial Gaussian process. Our proposed model can be used in many contexts where mixed continuous and discrete multivariate responses are observed in an effort to quantify an unobservable continuous measurement. In our example, the latent, or unobservable measurement is wetland condition. While predicted values of the latent wetland condition variable produced by the model at each location do not hold any intrinsic value, the relative magnitudes of the wetland condition values are of interest. In addition, by including point-referenced covariates in the model, we are able to make predictions at new locations for both the latent random variable and the multivariate response. Lastly, the model produces ranks of the multivariate responses in relation to the unobserved latent random field. This is an important result as it allows us to determine which response variables are most closely correlated with the latent variable. Our approach offers an alternative to traditional indices based on best professional judgment that are frequently used in ecology. We apply our model to assess wetland condition in the North Platte and Rio Grande River Basins in Colorado. The model facilitates a comparison of wetland condition at multiple locations and ranks the importance of in-field measurements.

email: **erin.schliep@duke.edu**

ANALYSIS OF VARIANCE OF INTEGRO-DIFFERENTIAL EQUATIONS WITH APPLICATION TO POPULATION DYNAMICS OF COTTON APHIDS

Jianhua Huang*, Texas A&M University

The population dynamics of cotton aphids is usually described by mechanistic models, in the form of IDEs with parameters representing some key properties of the dynamics. Investigation of treatments on the population dynamics is a central issue in developing successful chemical and biological controls for cotton aphids. Motivated by this important agricultural problem we propose a framework of ANOVA for IDEs.

email: **jianhua@stat.tamu.edu**

67. Estimation and Inference for High Dimensional and Data Adaptive Problems

A FLEXIBLE FRAMEWORK FOR SPARSE ADDITIVE MODELING

Noah Simon*, University of Washington

In high dimensional modeling problems there is a tradeoff between adding flexibility (to decrease the bias) and removing flexibility (to decrease the variance). Often L1 penalized linear models are used as they give parsimonous fits with few variables and relatively low bias. However, sometimes a linear model is not a good approximation to the true underlying signal. To combat this, authors have begun to consider sparse additive models. In most proposals these models have been constructed using a group lasso penalty and an explicit basis. We take a different route and provide a framework to construct sparse additive models using either an explicit basis expansion, or structure induced by a penalty or constraint. This allows us to build more data-adaptive additive models --- eg. piecewise constant or linear models with knots chosen adaptively, isotonic regression models, spline models etc. We give an efficient algorithm and show that in many cases, fitting these models requires only the same order of computation as the usual linear lasso.

email: **nrsimon@uw.edu**

INFERENCE FOR REGRESSION QUANTILES AFTER MODEL SELECTION

Jelena Bradic*, University of California, San Diego

Mladen Kolar, University of Chicago

Regression quantiles have been a topic of interest for the longest period of time. The last two decade have seen extensive research devoted to this problem in high dimensional regimes when the dimension of parameters overcomes the sample size. WIth the modern data being collected at a fast pace, developing methodology for the inference after model selection becomes ever so important. We address the issue of optimal confidence intervals and bahadur expansions of the regression quantiles. We introduce penalized regression rank scores and propose a novel estimator of the density of the regression noise as a score statistic.

Hence, our method is independent of the error distribution and is nonparametric in nature. Our results are non-asymptotic in nature and reflect the delicate interplay of the signal strength and sample size.

email: **jbradic@ucsd.edu**

FALSE DISCOVERY RATE CONTROL FOR SPATIAL DATA

Alexandra Chouldechova*, Carnegie Mellon University

In many modern applications the aim of the statistical analysis is to identify "interesting" or "differentially behaved" regions from noisy spatial measurements. From a statistical standpoint the task is both to identify a collection of regions which are likely to be non-null, and to associate to this collection a measure of uncertainty. Viewing this task as a large scale multiple testing problem, and borrowing ideas from the Poisson clumping heuristic literature, we present methods for controlling the clusterwise false discovery rate (cFDR), defined as the expected fraction of reported regions that are in truth null. We show that the widely used approach of applying an FDR controlling procedure pointwise to the measurement locations in general fails to control the cFDR. We also describe how the proposed cFDR procedure can be used to incorporate into the analysis quantities such as cluster size and slope at upcrossing.

email: **achould@cmu.edu**

CONDITIONAL OR FIXED? DIFFERENT PHILOSOPHIES IN ADAPTIVE INFERENCE

Max Grazier-G'sell*, Carnegie Mellon University

Ryan Joseph Tibshirani, Carnegie Mellon University

Inferential approaches in adaptive, non-classical estimation settings can be divided, roughly speaking, into two camps. The first camp conditions on the model selected by the adaptive procedure and performs a (conditional) hypothesis test accordingly. The second camp uses the adaptive procedure as a stepping stone to perform marginal (fixed) hypothesis tests and then defines the model of interest according to the results of these tests. There is much recent and exciting work that has been done in both categories. We discuss specific examples of such advances in the literature, and the advantages and disadvantages of the two general approaches.

68. CONTRIBUTED PAPERS: Novel Methods for Bioassay Data

drLUMI: TOOLS FOR THE ANALYSIS OF THE MULTIPLEX IMMUNOASSAYS IN R

Hector Sanz*, Universitat de Barcelona, Spain

John Aponte, Universitat de Barcelona, Spain

Jaroslaw Harezlak, Indiana University Fairbanks School of Public Health, Indianapolis

Magdalena Murawska, Indiana University Fairbanks School of Public Health, Indianapolis

Ruth Aguilar, Universitat de Barcelona, Spain

Gemma Moncunill, Universitat de Barcelona, Spain

Carlota Dobaño, Universitat de Barcelona, Spain

Clarissa Valim, Harvard School of Public Health

Multiplex immunoassays are used to measure concentrations of several analytes simultaneously and are important for biomarker discovery. In addition to the biological samples, assays include control standard curves to calibrate between-plate variability and quantify analyte concentrations. However, their range might result in suboptimal calibration and decrease assay sensitivity, i.e., the number of samples with analyte concentrations within limits of quantification (LOQ). To optimize the assay, we used alternative approaches to fit the standard curves, treat background noise, and estimate LOQ. We developed a comprehensive R package with functions for managing data, calibrating assays and performing quality control (QC). Dose-response five-parameter logistic regression and other parametric functions for standard curves are implemented. Several approaches for treating background noise and estimating LOQ are available to maximize the number of quantifiable concentrations. The package automates QC metrics and includes analysis of residuals and reliability estimates. Using data from a correlates of protection

study of a malaria vaccine candidate, we show the importance and exemplify the functionality of drLumi.

email: **hector.sanz@cresib.cat**

A BAYESIAN ANALYSIS OF BIOASSAY EXPERIMENTS

Luis G. Leon-Novelo*, University of Louisiana at Lafayette

Andrew Womack, Indiana University

Hongxiao Zhu, Virginia Polytechnic Institute and State University

Xiaowei Wu, Virginia Polytechnic Institute and State University

We address model based statistical Bayesian inference to analyze data arising from bioassay experiments. These experiments consist in assigning increasing doses of a chemical substance to different groups of individuals (usually lab rats or mice) while retaining a control group unexposed to the substance. For every individual a 0-1 response is observed according to whether the individual exhibits the adverse event of interest. The objective of the experiment is to conclude if there is an association between the adverse event and the substance. A decision will be made based on the Bayes factor comparing two probit models: the model that assumes increasing dose effects vs. the model that assumes no dose effect. Moreover, the proposed approach incorporates information of (historical) control groups from previews studies and is able tohandle data with very few occurrences of the adverse event. The proposed method is compared to a variation of the Peddada test (Peddada et.al. 2007) via simulation and is shown to have higher power.

email: **leonnovelo@gmail.com**

COMPOUND RANKING BASED ON A NEW MATHEMATICAL MEASURE OF EFFECTIVENESS USING TIME COURSE DATA FROM CELL-BASED ASSAYS

Francisco J. Diaz*, University of Kansas Medical Center

The IC50 concentration has limitations that make it unsuitable for examining a large number of compounds in cytotoxicity studies, particularly when multiple exposure periods are tested. A new approach to measure drug effectiveness is presented, which ranks compounds according to their toxic effects on live cells. This effectiveness measure combines all exposure times tested, compares the growth rates of a cell line in the presence of the compound with its growth rate in the presence of DMSO alone, measures a wider spectrum of toxicity than IC50, and allows automatic analyses of large numbers of compounds. It is easily implemented in linear regression software, provides a comparable measure of effectiveness for each investigated compound, and tests the null hypothesis that a compound is non-toxic versus the alternative that it is toxic. Our approach allows defining an automated decision rule for deciding whether a compound is significantly toxic. We illustrate with a cell based study of the cytotoxicity of 24 analogs of novobiocin; the compounds were ranked in order of cytotoxicity to a panel of 18 cancer cell lines and 1 normal cell line. Our approach may also be a good alternative to computing the EC50.

email: **fdiaz@kumc.edu**

NONPARAMETRIC CLASSIFICATION OF CHEMICALS USING QUANTITATIVE HIGH THROUGHPUT SCREENING (qHTS) ASSAYS

Shuva Gupta*, National Institute of Environmental Health Sciences, National Institutes of Health

Soumendra Lahiri, North Carolina State University

Shyamal Peddada, National Institute of Environmental Health Sciences, National Institutes of Health

Toxicologists (and regulatory agencies) are often interested in identifying toxins and carcinogens humans are exposed to. While the standard 2-year rodent cancer bioassay conducted by the US National Toxicology Program (NTP) is often considered as the “gold standard” to evaluate chemicals, it is typically slow and expensive. Consequently, the NTP, the US Environmental Protection Agency (EPA) and others have begun exploring quantitative high throughput screening (qHTS) assays where thousands of chemicals can be processed in each run of the assay. These are cost effective and take considerably less time than the standard 2-year cancer bioassay. For each chemical, the data obtained from qHTS assay consists of responses

to several doses (e.g. 10 to 14 doses). Typically, chemicals with a sigmoidal shaped dose-response may be regarded as potentially active (i.e. potential toxin). Otherwise they are potentially regarded as in-active (i.e. perhaps not a toxin). Due to various characteristics of the data and the design, these distinctions are not clear cut and hence some chemicals are declared inconclusive. Recently, Lim, Sen and Peddada (2013) developed a robust parametric methodology for classifying chemicals from qHTS assays. As commonly done by toxicologists, Lim et al. (2013) used the Hill function to model the sigmoidal dose-response. While parameters of the Hill function provide important interpretations to a toxicologists, and hence a very preferred model, there are instances where such a parametric function can be too rigid for qHTS data. Lim et al. (2013) describe several challenges and open problems in the analysis of data from qHTS assays. We overcome some of those challenges by taking a nonparametric approach to the problem. In this talk we describe a methodology based on nonparametric monotone and convex functions that can be used for classifying chemicals as active, inactive or inconclusive. The resulting methodology is illustrated using a data obtained from the NTP. Operating characteristics of the proposed methodology are discussed using an extensive simulation study that mimics the real qHTS assay data.

email: **shuvagupta@gmail.com**

ROBUST BAYESIAN METHODS FOR THE INVERSE REGRESSION WITH AN APPLICATION TO IMMUNOASSAY EXPERIMENTS

Magdalena Murawska, Indiana University Fairbanks School of Public Health, Indianapolis

Hector Sanz, Universitat de Barcelona, Spain

Ruth Aguilar, Universitat de Barcelona, Spain

Gemma Moncunill, Universitat de Barcelona, Spain

Carlota Dobaño, Universitat de Barcelona, Spain

John Aponte, Universitat de Barcelona, Spain

Clarissa Valim, Harvard School of Public Health

Jaroslaw Harezlak*, Indiana University Fairbanks School of Public Health, Indianapolis

Immunoassays are a common diagnostic and research tool in medical experiments. The use of such assays requires a calibration method that involves a standard curve estimation that reflects the functional relationship between the concentration of the analytes and the median fluorescence intensity (MFI). Using the inverse standard curve an unknown concentration can be estimated based on the given MFI. The most commonly used calibration methods rely on per plate and per analyte standard curve estimation. Such methods do not use the underlying biological properties and are not robust to the presence of outliers. Therefore we employ and expand alternative Bayesian robust methods proposed by Fong et al. (2012) which allow the specification of correlated non-normal errors. We extend that approach to other than 5 parameter logistic such as exponential and power functions to account for the setting where no upper asymptote is reached. The developed methods are applied in malaria vaccine study using cytokine Luminex platform. The preliminary results of the performed simulations indicate the more robustness of the Bayesian approach comparing to mixed model or plate-by-plate approaches especially when the informative prior information is incorporated.

email: **mmurawsk@iu.edu**

ESTIMATING THE PREVALENCE OF MULTIPLE DISEASES VIA TWO-STAGE HIERARCHICAL POOLING

Md S. Warasi*, University of South Carolina

Joshua M. Tebbs, University of South Carolina

Christopher McMahan, Clemson University

Testing protocols in large-scale disease screening applications often involve pooling biospecimens (e.g., blood, urine, swabs, etc.) to lower costs and/or to increase the number of individuals who are screened. Motivated by the recent development of assays that detect multiple diseases, it is now common to test biospecimen pools for multiple infections simultaneously. In a recent article, Tebbs, McMahan, and Bilder (Biometrics, 2013) developed an expectation-maximization algorithm to estimate the prevalence of

two infections using a two-stage, Dorfman-type testing protocol motivated by current screening practices for chlamydia and gonorrhea in the United States. In this article, we have the same goal, but instead we take a more flexible Bayesian approach. This allows us to incorporate information about assay uncertainty during the screening process (which involves testing both pools and individuals) and also to update information as more individuals are screened. Overall, our approach provides reliable inference for disease probabilities and accurately estimates assay sensitivity and specificity even when little or no information is supplied in the prior distributions. We illustrate our approach using chlamydia and gonorrhea screening data from the Infertility Prevention Project. Extensions to more than two infections are also possible.

email: **sarkerm@email.sc.edu**

A BALLOONED BETA REGRESSION MODEL AND ITS APPLICATION TO BIOASSAY DATA

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. A ballooned beta regression model with expected responses equal to the four parameter logistic model is developed that expands the response boundaries from (0, 1) to (L, U), where L and U are unknown parameters. The ballooned beta regression function differs from the typical four parameter logistic model which has positive probability of responses from negative infinity to positive infinity. Given multiple Elisa plates of bioassay data from different laboratories, the motivating problem was to ascertain whether they all had equivalent boundaries. For this data, we find MLEs using combination of grid searches and the Newton-Raphson method. We first test equivalences of the boundaries among plates. We do this under the ballooned beta model. Then we use a bivariate normal approximation to test the equivalence of the slopes and inflection points, considering L and U to be nuisance parameters. A 95 percent confidence ellipsoid is drawn to detect plates with outlying slopes and interception points.

email: **vincenty43@gmail.com**

69. CONTRIBUTED PAPERS: Infectious Disease

VIRAL GENETIC LINKAGE ANALYSIS IN THE PRESENCE OF MISSING DATA

Shelley Han Liu*, Harvard University

Gabriel Erion, Harvard University

Vladimir Novitsky, Harvard School of Public Health

Victor DeGruttola, Harvard School of Public Health

Phylogenetic linkage, based on viral sequencing data from HIV prevention trials at the community level, can provide insight into HIV transmission dynamics and the impact of prevention interventions. Specifically, phylogenetic linkage has the potential to inform whether recently-infected individuals have acquired viruses circulating within or outside a community. Characteristics and patterns of HIV clustering can help to trace transmission dynamics of viruses circulating across communities. Specifics of HIV clustering can be related to the potential of some individuals to contribute disproportionally to the spread of the virus. However, assessment of the extent to which individual (incident or prevalent) viruses are clustered within a community is biased if only a subset of subjects are observed, especially if that subset is not representative of the entire HIV infected population. To address this concern, we develop a multiple imputation framework in which missing viral sequences are imputed based on a biological model for the diversification of viral genomes. The imputation method decreases the bias in clustering that arises from informative missingness. Data from a household survey conducted in a village in Botswana are used to illustrate these methods. We demonstrate that the multiple imputation approach effectively corrects for bias in the overall proportion of clustering due to informative missingness of individuals from certain demographic groups, and that we can recreate the entire sample of the population by viewing the observed dataset as a biased sample from the population.

email: **shelleyliu@fas.harvard.edu**

A BAYESIAN APPROACH TO ESTIMATING CAUSAL VACCINE EFFECTS ON BINARY POST-INFECTION OUTCOMES

Jincheng Zhou*, University of Minnesota

Haitao Chu, University of Minnesota

Michael G. Hudgens, University of North Carolina, Chapel Hill

M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington

To estimate causal effects of vaccine on post-infection outcomes, Hudgens and Halloran (2006) defined a post-infection causal vaccine efficacy estimand VE I based on the principal stratification framework using the maximum likelihood estimation method. Extending their research, we propose a Bayesian approach to estimate the causal vaccine effects on binary post-infection outcomes. The identifiability of the causal vaccine effect VE I is discussed under different assumptions on selection bias. The performance of the proposed Bayesian method is compared with the maximum likelihood method through simulation studies and two case studies -- a clinical trial of a rotavirus vaccine candidate and a field study of a pertussis vaccine. For both case studies, the Bayesian approach provided similar inference as the frequentist analysis. However, simulation studies with small sample sizes suggest that the Bayesian approach provides smaller bias and shorter confidence interval length.

email: **zhoux801@umn.edu**

EXPLORING BAYESIAN LATENT CLASS MODELS AS A POTENTIAL STATISTICAL TOOL TO ESTIMATE SENSITIVITY AND SPECIFICITY IN PRESENCE OF AN IMPERFECT OR NO GOLD STANDARD

Jay Mandrekar*, Mayo Clinic

Assessment of a new assay or diagnostic test is generally performed using statistical measures such as sensitivity, specificity, negative predictive value, positive predictive value and area under the curve when an established gold standard exists. However, in some cases, the gold standard may be imperfect or may not exist. In such situations, Bayesian latent class models (BLCM) is proposed as one of the possible alternatives. LCM does not assume any gold standard and a true disease state (present/absent) for each individual is also unknown. Bayesian methodology to LCM will be illustrated using a simple example of a real life dataset from Clinical Microbiology research study. This approach is increasingly used to validate diagnostic tests for infectious diseases and clinical microbiology research without assuming a gold standard.

email: **mandrekar.jay@mayo.edu**

MODELING AND INFERENCE FOR ROTAVIRUS DYNAMICS IN NIGER

Joshua Goldstein*, The Pennsylvania State University

Murali Haran, The Pennsylvania State University

Matthew Ferrari, The Pennsylvania State University

Recently developed vaccines provide a new way of controlling rotavirus in sub-Saharan Africa. Models for the transmission dynamics of rotavirus are critical for assessing effects of vaccination and guiding intervention strategies. We examine rotavirus infection in the Maradi area in southern Niger, using hospital surveillance data provided by Médecins Sans Frontières collected over two years. Additionally, a cluster survey of households in the region allows us to estimate the proportion of children with diarrhea who consulted at a health structure. We compare our results across several variants of Susceptible-Infectious-Recovered (SIR) compartmental models to quantify the impact of modeling assumptions on our estimates. Model parameters are estimated by Bayesian inference using Markov chain Monte Carlo. Our approach allows us to quantify the burden of infection in the region, and explore the impact of vaccination on both the short-term dynamics and the long-term reduction of rotavirus incidence under varying levels of coverage. Additionally, we investigate two-strain dynamic models to gain insight into a shift in the observed dominant genotype of rotavirus, consistent with the effects of strain replacement.

email: **jrg326@psu.edu**

COMPARISON OF GROUP TESTING ALGORITHMS FOR CASE IDENTIFICATION IN THE PRESENCE OF DILUTION EFFECT

Dewei Wang*, University of South Carolina

Christopher S. McMahan, Clemson University

Colin M. Gallagher, Clemson University

Group testing, through the use of pooling, has been widely implemented as a more efficient means to screen individuals for infectious diseases. Various testing strategies, such as hierarchical and square array-based testing algorithms, have been proposed. In this talk, I will present the comparison of the operating characteristics, including testing efficiency and classification accuracy, of these algorithms for the purpose of case identification. The differences between our approach and the previous ones are the assumptions regarding testing error rates. We relax previously made assumptions by acknowledging the mechanistic structure of the diagnostic assays. By doing this, we are able to account for the dilution effect; i.e., truly positive specimens could be diluted when they are pooled together with many truly negative ones, and thus cannot be detected. This methodology is illustrated by comparing different testing algorithms via the HIV, HBV and HCV data collected form a study involving Irish prisoners.

email: **deweiwang@stat.sc.edu**

CHOLERA TRANSMISSION IN OUEST REGION OF HAITI: DYNAMIC MODELING AND PREDICTION

Alexander Kirpich*, University of Florida

Alex Weppelmann, University of Florida

Yang Yang, University of Florida

Ira Longini, University of Florida

We present a stochastic compartmental model for cholera transmission that combines a framework of SIRS for human hosts with an environmental reservoir of the bacteria to accounts for both human-to-human and environment-to-human-to-environment transmission routes. In addition, we consider the effect of environmental conditions such as temperature and precipitation on modulating the dynamics. The model distinguishes between symptomatic and asymptomatic infections, each with its own disease course and infectivity level. The asymptomatic subpopulation is not observable, and we perform sensitivity analysis on related parameters. We apply our model to surveillance data in the Ouest region of Haiti during 2010-2014 years. We found that the transmission dynamics in Haiti were shaped jointly by the transmission among human hosts and the environmental reservoir, the waning of immunity in human hosts, the natural life cycle of the bacteria, and the potential effects of other external factors such as phage that infects the bacteria.

email: **akirpich@ufl.edu**

70. CONTRIBUTED PAPERS: Variable Selection

WEAK SIGNAL IDENTIFICATION AND INFERENCE IN PENALIZED MODEL SELECTION

Peibei Shi*, University of Illinois, Urbana-Champaign

Annie Qu, University of Illinois, Urbana-Champaign

Penalized model selection methods are developed to select variables and estimate coefficients simultaneously, which is useful in high-dimensional variable selection. However, identification and inference for weak signals are still quite challenging and are not well-studied. Existing inference procedures for the penalized estimators are mainly focused on strong signals. This motivates us to investigate finite sample behavior for weak signal inference. We propose an identification procedure for weak signals in finite samples, and provide a transition phase in-between noise and strong signal strengths. A new two-step inferential method is introduced to construct better inference for the weak signals being identified. Our simulation studies show that the proposed method leads to better confidence coverages for weak signals, compared with those using asymptotic inference, perturbation and bootstrap resampling approaches. We also illustrate our method for HIV antiretroviral drug susceptibility data to identify genetic mutations associated with HIV drug resistance.

email: **pshi2@illinois.edu**

FEATURE SCREENING FOR TIME-VARYING COEFFICIENT MODELS ULTRAHIGH DIMENSIONAL LONGITUDINAL DATA

Wanghuan Chu*, The Pennsylvania State University

Runze Li, The Pennsylvania State University

Matthew Reimherr, The Pennsylvania State University

This paper is concerned with feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. We propose a new screening method that identifies important predictors after accounting for within-subject correlation and time-varying variance of the longitudinal response. We examine their finite sample performance by comparing with other existing methods via Monte Carlo simulations. In the real data example, Childhood Asthma Management Program (CAMP) datasets are analyzed, where SNPs of genes that affect children's asthma measurements are selected after accounting for baseline predictors. We advocate a two-stage approach by first reducing the ultrahigh dimensionality to a moderate size using the proposed procedure, and then applying model selection techniques to make statistical inference on the coefficient functions and covariance structure. To compare models selected from our screening procedure and other methods, we evaluate the prediction performance through leave-one-out cross validation. Finally, we discuss the joint and individual heritability of SNPs estimated from the best models selected.

email: **wxc228@psu.edu**

A REGULARIZED APPROACH FOR SIMULTANEOUS ESTIMATION AND MODEL SELECTION FOR SINGLE INDEX MODELS

Longjie Cheng*, Purdue University

Peng Zeng, Auburn University

Yu Zhu, Purdue University

The single index model generalizes the linear regression model by incorporating a non-parametric component. It has become increasingly more popular due to its flexibility in modelling. Similar to the linear regression model, the set of predictors for the single index model can contain a large number of irrelevant variables. In this work, we propose a new method for simultaneous estimation and model selection for the single index model. We will develop a coordinate descent algorithm to efficiently implement our method for both low and high dimensional cases. We will show that under certain conditions, the proposed method can consistently estimate the true index and select the true model. Simulations with various settings and a real data analysis are conducted to demonstrate the estimation accuracy, the selection consistency and the computational efficiency of our proposed method.

email: **cheng70@purdue.edu**

MULTI-STEP LASSO

Haileab Hilafu*, University of Tennessee

The traditional linear regression model remains one of the most popular statistical inference tools in a diverse of applications due to its simplicity and intuitive interpretability. Under this model, the LASSO (Tibshirani, 1996) is an attractive penalized least squares approach that provides simultaneous estimation and variable selection. However, the LASSO known to have many limitations, especially when the number of non-zero coefficients exceed the available sample size and the variables corresponding to the non-zero coefficients are highly correlated. In this talk, we will present a novel algorithm, the multi-step LASSO, which shields the LASSO from these limitations. The algorithm exploits the correlation structure among the predictors to improve estimation. Extensive simulation studies and application to a publicly available gene expression data on Diffuse Large-B-Cell Lymphoma show that the proposed method yields superior results under different modeling scenarios.

email: **hhilafu@utk.edu**

BAYESIAN HIERARCHICAL VARIABLE SELECTION INCORPORATING MULTI-LEVEL STRUCTURAL INFORMATION

Changgee Chang*, Emory University

Yize Zhao, Statistical and Applied Mathematical Sciences Institute

Qi Long, Emory University

Recently, considerable effort has been made to incorporate structural or biological information among covariates into variable selection. In this work, we propose a Bayesian approach for hierarchical variable selection in Gaussian

process models while incorporating multi-level structural/biological information. We develop efficient MCMC algorithms for posterior computation. We examine the performance of our proposed method by simulation studies and we apply it to a colorectal cancer study for assessing treatment effects on multiple functional biomarkers.

email: changgee.chang@emory.edu

MODEL SELECTION FOR PROTEIN COPY NUMBERS IN POPULATIONS OF MICROORGANISM

Burcin Simsek*, University of Pittsburgh

Hanna Salman, University of Pittsburgh

Satish Iyengar, University of Pittsburgh

Recent biophysical studies have raised questions about the possible universality of protein copy number fluctuations. We are interested in comparing the fits of several models to those fluctuations. These models include the lognormal, generalized inverse Gaussian, and Frechet using closeness as measured by the Kullback-Leibler divergence. The lognormal results from a large number of multiplicative processes, or exponential growth; the generalized inverse Gaussian arises as a first passage time for diffusions; and the Frechet is an extreme value distribution. In this study, we show that the lognormal gives the best fit, and discuss implications for underlying biophysical processes.

email: **bus5@pitt.edu**

GLOBALLY ADAPTIVE QUANTILE REGRESSION WITH ULTRA-HIGH DIMENSIONAL DATA

Qi Zheng*, Emory University

Limin Peng, Emory University

Xuming He, University of Michigan

Quantile regression has become a valuable tool to analyze heterogeneous covaraite-response associations that are often encountered in practice. The development of quantile regression methodology for high dimensional covariates primarily focuses on examination of model sparsity at a single or multiple quantile levels, which are typically prespecified ad hoc by the users. The resulting models may be sensitive to the specific choices of the quantile levels, leading to conceptual difficulties in identifying relevant variables of interest. We propose a new penalization framework for quantile regression in the high dimensional setting. Our proposed approach achieves consistent shrinkage of regression quantile estimates across a continuous range of quantiles levels, enhancing the flexibility and robustness of the existing penalized quantile regression methods. Our theoretical results include the oracle rate of uniform convergence and weak convergence of the parameter estimators. We also use numerical studies to confirm our theoretical findings and illustrate the practical utility of our proposal.

email: **qi.zheng@emory.edu**

71. CONTRIBUTED PAPERS: Modeling Health Data with Spatial or Temporal Features

MODELING OF CORRELATED OBJECTS WITH APPLICATION TO DETECTION OF METASTATIC CANCER USING FUNCTIONAL CT IMAGING

Yuan Wang*, University of Texas MD Anderson Cancer Center

Brian Hobbs, University of Texas MD Anderson Cancer Center

Jianhua Hu, University of Texas MD Anderson Cancer Center

Kim-Anh Do, University of Texas MD Anderson Cancer Center

Perfusion computed tomography (CTp) is an emerging functional imaging modality that uses physiological models to quantify characteristics pertaining to the passage of fluid through blood vessels. Perfusion characteristics provide physiological correlates for neovascularization induced by tumor angiogenesis. Thus CTp offers promise as a non-invasive quantitative functional imaging tool for cancer detection, prognostication, and treatment monitoring. We first developed a Bayesian probabilistic framework for simultaneous supervised classification of multivariate correlated regions. We demonstrate that simultaneous Bayesian classification yields dramatic improvements in performance in the presence of strong correlation, yet remains competitive with classical methods in the presence of weak or no correlation. A

semi-parametric model is further implemented for estimation and prediction of sparse spatiotemporally correlated CTp characteristics derived from multiple intra-patient metastatic sites. We considered weighted kernel smoothing and joint prediction of curves arising from multiple ROIs within the same patient to improve characterizations of contrast absorption over time. The methodology builds a foundation for probabilistic segmentation of regions of liver that exhibit perfusion characteristic indicative of metastatic sites using CTp maps acquired over the entire liver.

e-mail: **ywang46@mdanderson.org**

A SPATIALLY VARYING COEFFICIENT MODEL WITH PARTIALLY UNKNOWN PROXIMITY MATRIX FOR THE DETECTION OF GLAUCOMA PROGRESSION USING VISUAL FIELD DATA

Joshua L. Warren*, Yale School of Public Health

Jean-Claude Mwanza, University of North Carolina, Chapel Hill

Angelo P. Tanna, Northwestern University

Donald L. Budenz, University of North Carolina, Chapel Hill

Glaucoma is a leading cause of irreversible blindness worldwide. Once a diagnosis is made, careful monitoring of the disease is required to prevent vision loss. However, determining if the disease is progressing remains the most difficult task in the clinical setting. We introduce new spatial methodology in the Bayesian setting in order to properly model the progression status of a patient, as determined by expert clinicians, as a function of changes in sensitivities at each visual field (VF) location jointly. Past modeling attempts include the analysis of global VF measures over time or the separate analyses of sensitivities at individual VF locations over time. The first set of methods ignores valuable spatial information while the second set is inefficient and fails to account for spatial similarities in vision loss across the VF. Our spatial probit model jointly incorporates all VF changes in a single framework while accounting for structural similarities between neighboring VF regions. Results indicate that our method provides improved model fit when compared with previously developed methods. The model is also shown to provide improved predictions of progression status in a validation dataset. This model may be clinically useful for detecting the glaucoma progression status of an individual.

email: **joshua.warren@yale.edu**

MAPPING AND MEASURING THE EFFECT OF PRIVATIZATION ON ALCOHOL AND VIOLENCE: DOES IT REALLY MATTER?

Loni Philip Tabb*, Drexel University

Tony H. Grubesic, Drexel University

The increasing presence 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 the 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 research is to examine the spatio-temporal distribution of alcohol outlets before and after privatization in Seattle, Washington, 2010-2014. This natural experiment allows us to see the effect of privatization on the already well-documented positive relationship between alcohol outlets and violence. 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**

MODELING ADOLESCENT HEALTH DATA USING A BINARY SPATIAL-TEMPORAL GENERALIZED METHOD OF MOMENTS APPROACH

Kimberly Kaufeld*, Statistical and Applied Mathematics Institute and North Carolina State University

Health applications generally contain binary data that are correlated across space and time. A model that accounts for the spatial and temporal dependence is the centered spatial-temporal autologistic regression model. Statistical inference for the autologistic model has been based upon pseudolikelihood, Monte Carlo maximum likelihood, or Monte Carlo expectation-maximization. However, these methods require the full conditional distribution to be defined, which can be computationally expensive given the complexity of spatial and temporal dependence. We propose an alternative approach to likelihood based methods for

binary spatial-temporal data using generalized method of moments. The approach is based on a set of moment conditions constructed with respect to spatial neighborhoods and time, accounting for the spatial and temporal dependence of the data. Comparisons of the estimation methods are demonstrated in a simulation and with the Add Health data to assess the effect of peers at multiple levels (i.e. grade and school) on drug and alcohol use.

email: **kimberly.kaufeld@gmail.com**

A PIECEWISE EXPONENTIAL SURVIVAL MODEL WITH CHANGE POINTS FOR EVALUATING THE TEMPORAL ASSOCIATION OF WORLD TRADE CENTER EXPOSURE WITH INCIDENT OBSTRUCTIVE AIRWAY DISEASE

Charles B. Hall*, Albert Einstein College of Medicine

Xiaoxue Liu, Montefiore Medical Center

Rachel Zeig-Owens, Montefiore Medical Center

Mayris P. Webber, Montefiore Medical Center

Jessica Weakley, Montefiore Medical Center

Theresa M. Schwartz, Montefiore Medical Center

David J. Prezant, Fire Department of the City of New York

The World Trade Center (WTC) disaster presents a unique opportunity to describe the latency period for obstructive airway disease (OAD) diagnoses. This prospective cohort study of New York City firefighters compared the timing and incidence of physician-diagnosed OAD relative to WTC-exposure. Exposure was categorized by WTC arrival time: high (9/11/2001 AM); moderate (9/11/2001 PM or 9/12/2001); or low (9/13/-24/2001). Piecewise exponential survival models with change points were used to model the relative rates (RR) and 95% confidence intervals (CI) of OAD incidence by exposure over the first ten years post-9/11/2001, estimating the time(s) of change in the RR with change point models. We observed change points at 15 and 84 months post-9/11/2001. Before 15 months the RR for the high versus low exposure group was 4.23 (95% CI 2.71-6.60), from 15 to 84 months 1.94 (95% CI 1.51-2.51) and thereafter, 1.01 (95% CI 0.76-1.36). Incidence of physician-diagnosed OAD increased in all exposure groups starting in the sixth year post 9/11/2001 as the program started covering OAD medications for free. This difference in RR by exposure occurred despite full and free access to healthcare for all WTC-exposed firefighters, demonstrating the persistence of WTC-associated OAD risk for up to seven years.

email: **charles.hall@einstein.yu.edu**

DISTRIBUTED LAG MODELS: EXAMINING ASSOCIATIONS BETWEEN THE BUILT ENVIRONMENT AND HEALTH

Jonggyu Baek*, University of Michigan

Brisa N. Sanchez, University of Michigan

Veronica J. Berrocal, University of Michigan

Emma V. Sanchez-Vaznaugh, San Francisco State University

Built environment factors have received heightened attention in recent years as potential contributors to health, given that the built environment can constrain individual-level choices and behaviors. For example, food outlets around schools may affect children dietary choices both through direct access to junk food and exposure to advertisement thereby influencing body weight. Although some research has observed significant associations between the availability of food outlets near schools and childhood obesity, other studies have not. Traditional regression methods have been widely used to examine said associations, but they often rely on measures of the built environment (e.g., number of food outlets) within pre-specified distances from schools. We propose using distributed lag models (DLMs) to describe the association between built environment features and health as a function of distance from the study locations. We demonstrate through simulation studies that traditional regression models can produce severely biased associations when there is spatial correlation among the built environment features. In contrast, inference based on DLMs is robust under various conditions of the built environment. We use this innovative application of DLMs to examine the association between the presence of convenience stores around California public schools and children's body mass index z-score.

email: **jongguri@umich.edu**

CLUSTER DETECTION TEST IN SPATIAL SCAN STATISTICS: ADHD APPLICATION

Ahmad Reza Soltani*, Kuwait University

Suja Aboukhamseen, Kuwait University

We establish hypotheses testing for spatial scan testing hypotheses for cluster detection, then provide a transparent test statistics procedure for cluster detection in a spatial settings. We also specify the limiting distribution of the test statistics. We apply our method to the special needs school students in Kuwait suffering from Attention Deficit Hyper Active Disorder, using real data. We do detects same primary and secondary clusters among districts of the students residential areas.

email: **asoltanir@yahoo.com**

72. CONTRIBUTED PAPERS: Advances in Longitudinal Modeling

CONDITIONAL MODELING OF LONGITUDINAL DATA WITH TERMINAL EVENT

Shengchun Kong*, Purdue University

Bin Nan, University of Michigan

Jack Kalbfleisch, University of Michigan

We consider longitudinal data analysis with a terminal event where the terminal event time is informative. Existing methods include the joint modeling approach using latent frailty and the marginal estimating equation approach using inverse probability weighting approach, and both assume that the relationship between the response variable and a set of covariates is the same no matter whether the terminal event occurs or not. This assumption, however, is not reasonable for many longitudinal studies. Therefore we directly model event time as a covariate, which provides intuitive interpretation. When the terminal event times are right-censored, a semi-parametric likelihood-based approach is proposed for the parameter estimation, where the Cox regression model is used for the censored terminal event time. We consider a two-stage estimation procedure, where the conditional distribution of the right-censored terminal event time given other variables is estimated prior to maximizing the likelihood function for the regression parameters. The proposed method outperforms the complete case analysis in simulation studies, which simply eliminates the subjects with censored terminal event times. Desirable asymptotic properties are provided.

email: **kongsc@umich.edu**

A MARGINALIZED MULTILEVEL MODEL FOR BIVARIATE LONGITUDINAL BINARY DATA

Gul Inan*, Middle East Technical University, Turkey

Ozlem Ilk Dag, Middle East Technical University, Turkey

This study considers analysis of bivariate longitudinal binary data. We propose a model based on marginalized multilevel model framework. The proposed model consists of two levels such that the first level associates the marginal mean of responses with covariates through a logistic regression model and the second level includes subject/time specific random intercepts within a probit regression model. The covariance matrix of multiple correlated time-specific random intercepts for each subject is assumed to represent the within-subject association. The subject-specific random effects covariance matrix is further decomposed into its dependence and variance components through modified Cholesky decomposition method and then the unconstrained version of resulting parameters are modelled in terms of covariates with low-dimensional regression parameters. This provides better explanations related to dependence and variance parameters and a reduction in the number of parameters to be estimated in random effects covariance matrix to avoid possible identifiability problems. Marginal correlations between responses of subjects and within the responses of a subject are derived through a Taylor series-based approximation. Data cloning computational algorithm is used to compute the maximum likelihood estimates and their standard errors of the parameters in the proposed model. The proposed model is illustrated through Mother's Stress and Children's Morbidity study data, where both population-averaged and subject-specific interpretations are drawn through Emprical Bayes estimation of random effects.

email: **ginan@metu.edu.tr**

AUGMENTED BETA RECTANGULAR REGRESSION MODELS: A BAYESIAN PERSPECTIVE

Jue Wang*, University of Texas Health Science Center, Houston

Sheng Luo, University of Texas Health Science Center, Houston

Mixed effects Beta regression models based on Beta distributions have been widely used to analyze longitudinal percentage or proportional data ranging between zero and one. However, Beta distributions are not flexible to extreme outliers or excessive events around tail areas, and they do not account for the presence of the boundary values zeros and ones because these values are not in the support of the Beta distributions. To address these issues, we propose a mixed effects model using Beta rectangular distribution and augment it with the probabilities of zero and one. We conduct extensive simulation studies to assess the performance of mixed effects models based on both the Beta and Beta rectangular distributions under various scenarios. The simulation studies suggest that the regression models based on Beta rectangular distributions improves the accuracy of parameter estimates in the presence of outliers and heavy tails. The proposed models are applied to the motivating Neuroprotection Exploratory Trial in PD Long-term Study-1 (LS-1 study, n=1741), developed by The National Institute of Neurological Disorders and Stroke Exploratory Trials in Parkinson's Disease (NINDS NET-PD) network.

email: **Jue.Wang@uth.tmc.edu**

RANK-BASED REGRESSION MODELS FOR LONGITUDINAL DATA

Rui Chen, University of Rochester

Tian Chen*, University of Rochester

Xin Tu, University of Rochester

Popular mean-based semi-parametric regression models such as the generalized estimating equations (GEE) improve robustness of inference over parametric models. However, such models are not robust against outlying observations. Rank regression (RR), a lesser-known model based on the Wilcoxon score for the Mann-Whitney-Wilcoxon (MWW) test, provides more robust estimates over GEE. Unfortunately, RR does not sufficiently address missing data arising in longitudinal studies. We discuss a new approach to address outliers in longitudinal study data. This robust alternative not only effectively addresses missing data, but has also been applied to extend the MWW to provide causal inference for observational studies. The approach is illustrated with both real and simulated data.

email: **tian_chen@urmc.rochester.edu**

MARKOV CHAINS AND CONTINUOUS TIME MULTI-STATE MARKOV MODELS COMPARISONS IN LONGITUDINAL CLINICAL ANALYSIS

Lijie Wan*, University of Kentucky

Richard J. Kryscio, University of Kentucky

Erin Abner, University of Kentucky

Multi-state Markov models are widely used to analyze longitudinal data describing the progression of a chronic disease or condition, like dementia. Several studies have focused on modeling true disease progression as a discrete time Markov chain, which requires certain assumptions. Recently, continuous-time multi-state Markov models have also become very popular. In this paper, we discuss the relationship as well as differences between these two modeling techniques. Our simulation study shows that when longitudinal data are arise from equally spaced intervals and with no unobserved transition between two contiguous assessment time points, these two types of models work equally well. When the data are not equally spaced, or there are possible unobserved transitions between two contiguous assessment time points (e.g., a patient dies without being observed first passing through severe clinical disease), the continuous-time multi-state Markov model is preferred. We also apply our model to a real dataset, the Nun Study, a cohort of 461 participants who were cognitively normal at study baseline and followed to autopsy.

email: **lijie.wan@uky.edu**

APPLICATIONS OF MULTIPLE OUTPUTATION FOR THE ANALYSIS OF LONGITUDINAL DATA SUBJECT TO IRREGULAR OBSERVATION

Eleanor M. Pullenayegum*, Hospital for Sick Children

Observational cohort studies often feature longitudinal data subject to irregular observation. Moreover, the timings of observations are often associated with the underlying disease process, and must thus be accounted for when analysing the data. Multiple outputation, which consists of repeatedly discarding excess observations, can be a helpful way of approaching the problem. In particular, we show that multiple outputation enables doubly robust inference within standard statistical software, and widens the scope of semi-parametric joint models for the outcome and visit processes to include cases where the visit process includes a time-varying endogenous covariate.

email: **pullena@mcmaster.ca**

A HIDDEN MARKOV MODEL APPROACH TO ANALYZE LONGITUDINAL TERNARY OUTCOME DISEASE STAGE CHANGE SUBJECT TO MISCLASSIFICATION

Julia Benoit*, University of Houston

Wenyaw Chan, University of Texas Health Science Center School of Public Health

Understanding the dynamic disease process is vital in early detection, diagnosis, and measuring progression. Continuous-time Markov chain (CTMC) methods have been used to estimate state change intensities but challenges arise when stages are potentially misclassified. We present an analytical likelihood approach where the hidden state is modeled as a three-state CTMC model using the possibly misclassified observed values. Covariate effects of the hidden process and misclassification probabilities of the hidden state are estimated without information from a 'gold standard' as comparison. Parameter estimates are obtained using a modified EM algorithm and identifiability of CTMC estimation is addressed. Simulation studies and an application studying Alzheimer Disease progression are presented. The method was highly sensitive to detecting true misclassification and did not falsely identify error in the absence of misclassification. In conclusion, we have developed a robust longitudinal methodology for categorical outcome data where the researcher is unsure of the level of uncertainty in the classification of disease severity stage if the purpose is to look at the process' transition behavior without a gold standard.

email: **Julia.Benoit@times.uh.edu**

73. CONTRIBUTED PAPERS: Causal Inference: Average and Mediated Effects

INSTRUMENTAL VARIABLE ESTIMATION OF THE MARGINAL AVERAGE EFFECT OF TREATMENT ON THE TREATED

Lan Liu*, Harvard University

Baoluo Sun, Harvard University

James Robins, Harvard University

Eric Tchetgen Tchetgen, Harvard University

The objective of many studies in health and social sciences is to evaluate the causal effect of a treatment or exposure on a specific outcome using observational data. In such studies the exposure is typically not randomized and therefore confounding bias can rarely be ruled out with certainty. The instrumental variable (IV) design plays the role of a quasi-experimental handle since the IV is associated with the treatment, independent of potential outcomes conditional on observed covariates and it affects the outcome only through treatment. In this paper, we present a novel framework for identification and estimation using an IV, of the marginal average causal effect of treatment amongst the treated (ETT) in the presence of unmeasured confounding. We show that access to an IV allows for partial identification of the association between exposure and the potential outcome under no exposure, which encodes the magnitude of selection bias due to confounding of the treatment. In the special case of a binary IV, we show that ETT is non-parametrically identified under a straightforward assumption that the IV does not interact with an unmeasured confounder in a logistic propensity score model for treatment. For inference, we propose three different semiparametric strategies (i) inverse probability weighting (IPW), (ii) outcome regression and (iii) doubly robust (DR) estimation which combines (i) and (ii) and is more robust than either strategies. Specifically, the DR estimator is shown to be consistent if either strategy (i) or (ii) is consistent. An extensive simulation study is carried out to investigate the finite sample performance of the proposed estimators. The methods are further illustrated in a well known application of the impact of participation in a 401(k) retirement programs on savings.

email: **lanl@email.unc.edu**

WITHIN-SUBJECT DESIGNS FOR CAUSAL MEDIATION ANALYSIS

Yenny Webb-Vargas*, Johns Hopkins Bloomberg School of Public Health

Martin A. Lindquist, Johns Hopkins Bloomberg School of Public Health

Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Michael E. Sobel, Columbia University

Making causal statements about mediation always poses a problem because, even if we can randomize the intervention, it is often difficult - or impossible - to randomize the assignment of the mediator. However, there are cases in which a different design of experiment can be applied, opening new possibilities for identification of the mediation counterfac-

tuals. Within-subject designs, common in experiments using functional Magnetic Resonance Imaging, allow for the observation of an individual's response under both treated and control conditions, often with replicates. We explored this design under the potential outcomes framework, and established identifiability conditions for estimating causal direct and indirect effects. Furthermore, we developed an estimation procedure that is robust to confounding of the mediator-outcome, and we tested it using simulations. We applied this method to a trial analyzing how the brain processes thermal pain. In this case, the mediator is the hemodynamic response function, conceptualized as a continuous function in time, and our method is able to estimate the corresponding functional parameter of a causal indirect effect.

email: **yennywebb@gmail.com**

MEDIATION ANALYSIS OF A SET OF CORRELATED PREDICTORS USING WEIGHTED QUANTILE SUM REGRESSION METHOD

Bhanu Murthy Evani*, Virginia Commonwealth University

Robert A. Perera, Virginia Commonwealth University

Chris Gennings, Icahn School of Medicine at Mount Sinai

Traditional mediation analysis uses the single predictor, multiple regression method; first proposed by Baron & Kenny (1986) and since then advanced by authors Hayes, MacKinnon and Vander-Weele. The application of Mediation analysis to a set of correlated predictors is challenging, due to Multicolinearity and the Reversal paradox. We propose using Weighted Quantile Sum (WQS) regression to analyze mediation effects of a set of correlated predictors on the outcome. The WQS method is a constrained, non-linear optimization algorithm which estimates the regression parameters, and the empirical weights of individual predictors (ranked as quartiles), simultaneously. The result is WQS index, which represents the set of correlated predictors as a single entity, having a composite effect on the outcome. Next, by applying the traditional mediation analysis method using the WQS index, enables finding the significant mediation effect of the predictors acting through a hypothesized mediation pathway, on the outcome of interest. While other constrained optimization methods focus on dimensionality reduction, WQS attempts to pick out the predictors amongst a correlated predictor set, that have a significant effect on the outcome. Preliminary simulation results of WQS applied to a set of two correlated predictors compared to the traditional Multiple Regression Mediation Analysis are presented.

email: **evanibm@vcu.edu**

BAYESIAN SEMIPARAMETRIC LATENT MEDIATION MODEL

Chanmin Kim*, Harvard University

Michael J. Daniels, University of Texas, Austin

Yisheng Li, University of Texas MD Anderson Cancer Center

We propose a Bayesian semiparametric method to estimate natural direct and indirect effects (causal effects) within clusters estimated based on a set of potential effect modifiers. The cluster specific direct and indirect effects can be estimated through a set of regression models whose coefficients differ by cluster. We construct a nonparametric model with a stick breaking prior to identify the clusters based on a large set of potential effect modifiers. We use this approach to estimate the cluster specific causal effects of an expressive writing intervention for patients with renal cell carcinoma (Milbury et al., 2014).

email: **ylit777@gmail.com**

ACCOUNTING FOR UNCERTAINTY IN CONFOUNDER SELECTION WHEN ESTIMATING AVERAGE CAUSAL EFFECTS IN GENERALIZED LINEAR MODELS

Chi Wang*, University of Kentucky

Corwin Matthew Zigler, Harvard School of Public Health

Giovanni Parmigiani, Dana-Farber Cancer Institute and Harvard School of Public Health

Francesca Dominici, Harvard School of Public Health

Confounder selection and adjustment are essential elements of assessing the causal effect of an exposure or treatment in observational studies. Building upon work by Wang et al. (2012) and Lefebvre et al. (2014), we propose and evaluate a Bayesian method to estimate average causal effects in studies with a large number of potential confounders, likely interactions between these confounders and the exposure of interest,

and uncertainty on which confounders should be included. Our method is applicable across all exposures and outcomes that can be handled through generalized linear models. In this setting, models coefficients are not collapsible across different models for confounding adjustment. We implement a Bayesian bootstrap procedure to estimate causal effects while acknowledging uncertainty in the population covariate distribution. Our method permits estimation of both the overall population causal effect and effects in specified subpopulations, providing clear characterization of heterogeneous exposure effects that may vary considerably across different covariate profiles. Simulation studies demonstrate that the proposed method performs well in small sample size situations with 100 to 150 observations and 50 covariates. The method is applied to evaluate the effect of surgery on reducing thirty-day hospital readmissions among 15060 US Medicare beneficiaries diagnosed with a brain tumor between 2000 and 2009.

email: **chi.wang@uky.edu**

VARIABLE SELECTION FOR ESTIMATING AVERAGE CAUSAL EFFECTS

Douglas Galagate*, U.S. Census Bureau

Determining which variables to include at each stage of the modeling process when estimating an average causal effect (ACE) is a topic of debate. In this simulation study, different subsets of the predictor variables are used to estimate the ACE. We estimate the ACE using outcome-focused, treatment-focused, and double-robust models with the goal of gaining some awareness of how variable selection affects the balance of covariates, prediction ability, and accuracy of the estimate. Optimizing all three metrics does not always coincide in the simulations conducted.

email: **galagate@umd.edu**

ESTIMATING MEDIATION EFFECTS UNDER CORRELATED ERRORS WITH AN APPLICATION TO fMRI

Yi Zhao*, Brown University

Xi Luo, Brown University

Mediation analysis assesses the effect passing through a intermediate variable (mediator) in a causal pathway from the treatment variable to the outcome variable. Structure equation models (SEMs) is a popular approach to estimate the mediation effect. However, causal interpretation usually requires strong assumptions which may not hold in many social and scientific studies. In this paper, we use mediation analysis in an fMRI experiment to assess the effect of randomized binary stimuli passing through a brain pathway of two brain regions. We propose a two-layer SEM framework for mediation analysis that provides valid inference even if correlated additive errors are present. In the first layer, we use a linear SEM to model the subject level fMRI data, where the continuous mediator and outcome variables may contain correlated additive errors. We propose a constrained optimization approach to estimate the model coefficients, analyze its asymptotic properties, and characterize the nonidentifiability issue due to the correlation parameter. To address this issue, we introduce a linear mixed effects SEM with an innovation to estimate the unknown correlation parameter in the first layer. Using extensive simulated data and a real fMRI dataset, we demonstrate the improvement of our approach over existing methods.

email: **yi_zhao@brown.edu**

74. CONTRIBUTED PAPERS: Variable Selection with High Dimensional Data

EMPIRICAL LIKELIHOOD TESTS FOR COEFFICIENTS IN HIGH DIMENSIONAL LINEAR MODELS

Honglang Wang*, Michigan State University

Ping-Shou Zhong, Michigan State University

Yuehua Cui, Michigan State University

We consider hypothesis testing problems for low-dimensional regression coefficients in a high dimensional linear model with Gaussian designs. We propose empirical likelihood based test procedures. The empirical likelihood is constructed based on asymptotically unbiased estimating equations. This method is flexible in incorporating auxiliary information to improve the power of testing and it is robust to heterogeneous random errors. Some simulation studies and real data analyses are conducted to demonstrate the proposed methods.

email: **wanghonglang2008@gmail.com**

TPRM: TENSOR PARTITION REGRESSION MODELS WITH APPLICATIONS IN IMAGING BIOMARKER DETECTION

Michelle F. Miranda*, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Many neuroimaging studies have collected ultra-high dimensional imaging data in order to identify imaging biomarkers that are related to normal biological processes, diseases, and the response to treatment, among many others. These imaging data are often represented in the form of a multi-dimensional tensor. Existing statistical methods are insufficient for analysis of these tensor data due to their ultra-high dimensionality as well as complex structure. The aim of this paper is develop a tensor partition regression modeling framework to establish an association between low-dimensional clinical outcomes and ultra-high dimensional tensor covariates. Our TPRM is a hierarchical model with four components: (i) a partition model that divides the high-dimensional tensor covariates into sub-tensor covariates; (ii) a canonical polyadic decomposition model to reduce the sub-tensor covariates to a low-dimensional feature vectors; and (iii) a generalized linear model that uses the feature vectors to predict clinical outcomes; (iv) a sparse inducing normal mixture prior. Under this framework, ultra-high dimensionality is not only reduced to a manageable level, resulting in efficient estimation, but also prediction accuracy is optimized to search for informative sub-tensors. We apply TPRM to a structural magnetic resonance imaging data, to predict diagnostic status of individuals with Attention Deficit Hyperactivity Disorder.

email: **michellemirandaest@gmail.com**

A BOOSTING-BASED VARIABLE SELECTION METHOD FOR SURVIVAL PREDICTION WITH GENOME-WIDE GENE EXPRESSION DATA

Yanming Li*, University of Michigan

Kevin He, University of Michigan

Yi Li, University of Michigan

Ji Zhu, University of Michigan

Motivated by a study using genome-wide gene expression data to predict breast cancer patients' survival, we proposed a Gateaux differential-based boosting (GDBoosting) method for variable selection in the ultra-high dimensional predictor and survival outcome setting. The proposed method can simultaneously select important variables via an early-stopping criterion and provide consistent estimates for the effects of selected variables provided the selection were as good as had been told by an oracle. The GDboosting algorithm is more computationally efficient compared to the lasso, and can be easily adapted to the case when the genome-wide gene expression data assume a grouping structure, such as biological pathways. The proposed method is evaluated through extensive simulations and is applied to the breast cancer NKI data set to predict breast cancer patients' survival.

email: **liyanmin@umich.edu**

STATISTICAL INFERENCE IN HIGH-DIMENSIONAL M-ESTIMATION

Hao Chai*, Yale University

Shuangge Ma, Yale University

This paper studies the asymptotic properties of some low-dimensional parameters under the high-dimensional M-estimation framework. We consider a general M-estimation problem in which penalization is used to select the variables. Based on the low-dimensional penalization projection method, we propose a two-stage estimator for selected low-dimensional parameters. Our framework includes linear and generalized linear models as special cases. Under reasonable conditions, we show that the proposed estimator is consistent and has an asymptotic normal distribution. We find that a stronger requirement on the sample size is needed in order to obtain the asymptotic normality than the consistency. The numerical performance of our estimator is evaluated through simulation studies and a high-dimensional data example is used to illustrate its application.

email: **haochai2@gmail.com**

AUGMENTED WEIGHTED SUPPORT VECTOR MACHINES FOR MISSING COVARIATES

Thomas G. Stewart*, University of North Carolina, Chapel Hill

Michael C. Wu, University of North Carolina, Chapel Hill

Donglin Zeng, University of North Carolina, Chapel Hill

In recent years, support vector machine (SVM) classifiers have demonstrated utility for a wide variety of classification tasks. A key feature of SVMs is that they allow for construction of both linear and non-linear decision rules which can yield better prediction when the data are complex, as is frequently the case in biomedical studies. A practical challenge for SVMs, as with many other classification methods, lies in the accommodation of missing data which commonly occur in real data applications due to imperfect data collection. Currently, many researchers rely on complete-case or imputation solutions which may introduce bias. Additional systematic approaches exist, but these alternative approaches require non-standard algorithms which have slowed their adoption. Therefore, we propose an EM-motivated solution to the incomplete data problem for SVMs which maintains the convex objective function and which allows the researcher to use the same software as the complete case solution. Simulations show that the proposed method often yields classification rules with higher accuracy than existing methods. We apply the approach to analyze data from HCV-TARGET, a longitudinal study of Hepatitis C patients.

email: **tgs@live.unc.edu**

VARIABLE SELECTION ON MODEL SPACES CONSTRAINED BY HEREDITY CONDITIONS

Andrew Womack, Indiana University, Bloomington

Daniel Taylor-Rodriguez*, Statistical and Applied Mathematics Institute and Duke University

Claudio Fuentes, Oregon State University

Often having a linear additive regression model in the predictors is not sufficient to adequately predict the response. Considering higher order polynomial terms and interactions between the predictors might substantially improve the results. In this setting, respecting the polynomial hierarchy of the terms is necessary to ensure that the ensuing model is invariant under location and scale transformations, especially when using shrinkage methods. With this in mind, we propose a Bayesian procedure using adaptive shrinkage estimators that respects the polynomial hierarchy between predictors by enforcing the strong heredity principle. We run simulations on a variety of scenarios to test the performance of our approachand compare with other popular methods found in the literature.

email: **taylor-rodriguez@samsi.info**

75. PRESIDENTIAL INVITED ADDRESS

BIG DATA, BIG OPPORTUNITIES, BIG CHALLENGES

David L. DeMets, Ph.D., Max Halperin Professor of Biostatistics, University of Wisconsin, Madison

Since the 1950's, biostatisticians have been successfully engaged in biomedical research, from laboratory experiments to observational studies to randomized clinical trials. We owe some of that success to the early pioneers, especially those biostatisticians who were present at the National Institutes of Health (NIH). They created a culture of scientific collaboration, working on the methodology as needed to solve the biomedical research problems in design, conduct and analysis. Over the past 5 decades, we have experienced a tremendous increase in computational power, data storage capability and multidimensionality of data, or "big data". Some of this expansion has been driven by genomics. At present, we have the opportunity to contribute to the design and analysis of genomic data, data stored in the electronic health record and continued needs of clinical trials for greater efficiency. However, with these opportunities, we have serious challenges starting with the fact that we need to develop new methodology to design and analyze the "big data" bases. The demand for quantitative scientists exceeds the supply and there is no strategic national plan to meet these demands. Federal funding for biomedical research has been flat and likely to remain so for several years, impacting both the ability

to train additional quantitative scientists and provide them with research funding for new methodologies. We face new or more public scrutiny, demanding that our data and analysis be shared earlier and earlier, even as the data are being gathered such as in clinical trials. Litigation is now part of our research environment. We will examine some of these issues and speculate on ways forward.

e-mail: **demets@biostat.wisc.edu**

76. RECENT ADVANCES IN DYNAMIC TREATMENT REGIMES

THE LIBERTI TRIAL FOR DISCOVERING A DYNAMIC TREATMENT REGIMEN IN BURN SCAR REPAIR

Jonathan Hibbard, University of North Carolina, Chapel Hill

Michael R. Kosorok*, University of North Carolina, Chapel Hill

In this talk we describe the design and analysis plan for the LIBERTI (Laser Induced, Biologically Engineered Remodeling of Thermally Injured) Skin Trial. This is a SMART (Sequential Multiple Assignment Randomized Trial) design to discover the best sequence of treatments over three time intervals to improve outcomes for patients with severe burn scaring as a function of baseline and historical tailoring variables. In addition, a simple randomized comparison of the three treatments under consideration (standard of care plus two different laser treatments) using a surrogate outcome is embedded within the SMART design.

email: **kosorok@unc.edu**

FROM IDEALIZED TO REALIZED: ESTIMATING DYNAMIC TREATMENT REGIMENS FROM ELECTRONIC MEDICAL RECORDS

Erica EM Moodie*, McGill University

David A. Stephens, McGill University

Due to the cost and complexity of conducting a sequential multiple assignment randomized trial, it is often desirable to estimate optimal strategies via other means which may then be trialed in a confirmatory study. Finding good candidate regimes can be done via simulation or using large non-experimental datasets in which treatment allocation were not randomized, such as electronic medical records (EMRs). Unfortunately, EMRs are subject to a variety of limitations. In this presentation, I will present a simulation design for a complex, continuous dosing problem, and discuss ongoing work in which we relax idealized assumptions and move towards more realistic scenarios. I will then present an analysis of EMR data from a London anticoagulation clinic.

email: **erica.moodie@mcgill.ca**

ADAPTIVE TREATMENT AND ROBUST CONTROL

Robin Henderson*, Newcastle University, UK

There has been steadily increasing statistical interest over the last ten years in the data-based development of optimal dynamic treatment rules. Given a sequence of observations the aim is to choose at each decision time the treatment that maximises some target. This is the same fundamental problem that underpins control methodology in applications (primarily engineering) or theory (often mathematical analysis). This talk looks at similarities and differences between the problems considered by the two schools and the methods that are used for their solutions. We examine how established control methods might be adapted for statistical adaptive treatment problems, and how statistical thinking might bring fresh ideas to the control literature, especially as control methods are now increasingly being used in biomedical applications.

email: **Robin.Henderson@ncl.ac.uk**

METHODS TO INCREASE EFFICIENCY OF ESTIMATION WHEN A TEST USED TO DECIDE TREATMENT HAS NO DIRECT EFFECT ON THE OUTCOME

James M. Robins*, Harvard University

A CAT scan of the lung has no effect on mortality from lung cancer except through its role in deciding what chemotherapeutic agent to treat with next. If one is trying to estimate the optimal CAT- scan frequency as a function of a patients evolving laboratory and clinical measures, one should be able to use the fact that the scan has no direct effect on mortality except through treatment to increase the efficiency of estimation. In this talk I show how that can be accomplished.

email: **robins@hsph.harvard.edu**

77. Predictive Models for Precision Medicine

THE POWER OF ELECTRONIC MEDICAL RECORDS AS DATA-GATHERING TOOLS FOR THE CREATION OF (a) LONGITUDINAL PERSONALIZED NEAR-REAL-TIME PREDICTIONS OF ADVERSE OUTCOMES AND (b) DATA-DRIVEN ADVICE SYSTEMS FOR MEDICAL DECISION-MAKING

David Draper*, University of California, Santa Cruz and eBay Research Labs

The recent rapid increase in the use of electronic medical records (EMRs) for near-real-time clinical documentation has created a singular new opportunity for longitudinal statistical predictive modeling, of at least two kinds: (a) vital signs, symptoms and laboratory results now appear in the EMR in an almost real-time manner, making possible much more accurate assessments of the risk that a given hospitalized patient will experience an adverse outcome (such as an unplanned transfer from the general medical wards to the intensive care unit); and (b) each hospitalization, when finished, becomes another row in a growing clinically-rich data base, permitting the development of systems in which physicians can make queries of the following form: among all of the patients in the past with this clinical profile, at this stage of the hospitalization, what clinical courses of action did physicians in the past undertake, and what were the success rates of those actions? In this talk I will describe work, of both types (a) and (b), that I have been doing recently with colleagues in the Division of Research of the Northern California Kaiser Permanente hospital chain.

e-mail: **draper@ams.ucsc.edu**

ASSESSING ILLNESS SEVERITY FROM ELECTRONIC HEALTH DATA

Suchi Saria*, Johns Hopkins University

Nearly one in five patients are harmed by iatrogenic errors during a hospitalization; types of harm include sepsis, central-line associated blood stream infection (CLABSI), urinary tract infection, and ventilator associated pneumonia. These harms result in prolonged length of stay, and increased risk of morbidity including death. Early detection can allow early treatment. However, early signs and symptoms are often subtle, and therefore difficult to detect by the "naked eye". With the HITECH act of 2009, much of an individuals health data — continuous physiologic streaming data, hundreds of laboratory test results, demographic data, personal and family medical history, treatments, imaging results, and so on — are now available for automated analysis. In this talk, I present early work for integrating these diverse measurements collected in the inpatient setting for assessing illness severity. These are useful for triage and early intervention for adverse events. This is joint work with collaborators at the Johns Hopkins School of Engineering and the Johns Hopkins School of Medicine.

email: **suchi.saria@gmail.com**

TOWARD INDIVIDUALIZING HEALTH CARE; STATISTICAL OPPORTUNITIES

Yates Coley, Johns Hopkins University

Zhenke Wu, Johns Hopkins University

Scott L. Zeger*, Johns Hopkins University

A century ago, William Osler, the first Johns Hopkins Chief of Medicine, said: "Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions which we know as disease". Biohealth has dramatically shifted in the last century when that statement was made. Revolutions in biologic and information technologies have unleashed a torrent of new data that make possible the kind of individual-specific medicine Osler envisioned. The Institute of Medicine calls on health systems to continuously learn from these complex data how to optimally care for each individual. This talk will discuss statistical opportunities to promote individualized healthcare. A hierarchical model will be introduced with components that represent (1) the trajectory of an individual's health status over time; (2) the effect of exogenous covariates and possibly endogenous interventions on health status; (3) the measurement of health status; and (4) the embedding of the individual in a relevant population. Bayesian methods will be used to estimate a person's health state or trajectory using a multivariate discrete state space and discrete time. Method for checking model predictions will be described. The approach and methods will be illustrated with recent clinical applications.

email: **sz@jhu.edu**

DANCING WITH BLACK SWANS: A COMPUTATIONAL PERSPECTIVE ON SUICIDE RISK DETECTION

Truyen Tran*, Deakin University and Curtin University, Australia

Santu Rana, Deakin University, Australia

Wei Luo, Deakin University, Australia

Dinh Phung, Deakin University, Australia

Svetha Venkatesh, Deakin University, Australia

Richard Harvey, Barwon Health, Australia

Despite great attention paid to suicide prevention with substantive medico-legal implications, there has been no satisfying method that can reliably predict future attempted or completed suicides. Is this impossible we ask? Are these (baby) black swans? This talk takes an alternate view when faced with such challenging outlier detection in big data problems. Can we focus on moderate and high risk with minimal error? We present an integrated machine learning framework to tackle this challenge. Our proposed framework consists of a novel feature extraction scheme, an embedded feature selection process, a set of risk classifiers and finally, a risk calibration procedure. We perform comprehensive study on data collected for the mental health cohort, and the experiments validate that our proposed framework outperforms risk assessment instruments by medical practitioners. The second part of the talk focuses on the challenges to adoption into clinician practice. We discuss the challenges, our solutions and finally outlines the implementation in clinical practice.

email: **truyen.tran@deakin.edu.au**

78. Electronic Health Records: Challenges and Opportunities

TRIALS AND TRIBULATIONS IN TRIALS USING EHR DATA

Meredith Nahm Zozus*, Duke University

NIH funded researchers and others are using EHR data for pragmatic trials in healthcare settings and other clinical studies. The Health Care Systems Research Collaboratory (www.nihcollaboratory.org) works with pragmatic clinical trials conducted in health systems and using EHR data. The literature is well peppered with empirical reports of data quality problems encountered in using EHR data for purposes other than those for the data were originally collected, however, there is little systematic knowledge about the topic. The collaboratory, other initiatives, and studies using EHR data provide opportunities to observe data quality problems in EHR data and to build our knowledge. This talk casts data quality problems in two parts, representational inadequacy and information loss and degradation. Each part will be explored using case studies of actual data quality problems encountered in using EHR data. Examples of data quality problems will be reviewed along with assessment methods for identifying similar data quality problems.

email: **meredith.nahm@duke.edu**

STATISTICAL METHODS FOR DEALING WITH NON-RANDOM OBSERVATION OF LABORATORY DATA IN EHRs

Jason A. Roy*, University of Pennsylvania

EHRs are potentially a rich source of observational data for comparative effectiveness research. Laboratory data, which are increasingly available in many EHRs, can be used to either improve confounder control or as outcomes. However, which laboratory values are observed when are likely related to the underlying health of the subjects. Further, some subjects will have no values of particular laboratory variables available at all. We develop methods for dealing with this type of informative missing data. Two sets of assumptions are considered -- one which is more appropriate for out-

come models and another that is more appropriate for propensity score-based methods. We compare the methods using simulation studies. The methods are illustrated using data from several EHR-based comparative effectiveness studies.

email: **jaroy@upenn.edu**

EXTENDING BAYESIAN NETWORKS TO ESTIMATE CONDITIONAL SURVIVAL PROBABILITY USING ELECTRONIC HEALTH DATA

David M. Vock*, University of Minnesota

Julian Wolfson, University of Minnesota

Sunayan Bandyopadhyay, University of Minnesota

Gediminas Adomavicius, University of Minnesota

Paul E. Johnson, University of Minnesota

Gabriela Vazquez-Benitez, HealthPartners Institute for Education and Research

Patrick J. O'Connor, HealthPartners Institute for Education and Research

Models for predicting the risk of cardiovascular events based on individual patient characteristics are important tools for managing patient care. Most current and commonly used risk prediction models have been built from carefully selected epidemiological cohorts. However, the homogeneity and limited size of such cohorts restricts the predictive power and generalizability of these risk models to other populations. Electronic health data (EHD) from large health care systems provide access to data on large, heterogeneous, and contemporaneous patient populations. The unique features and challenges of EHD, including missing and high-dimensional risk factor information, non-linear relationships between risk factors and cardiovascular event outcomes, and differing effects in different patient subgroups, demand novel machine learning approaches to risk model development. However, many machine learning methods are not well-suited to handle right-censored outcomes. We describe how to efficiently extend Bayesian networks to estimate the conditional survival distribution. We show that our approach can lead to better predictive performance than the Cox proportional hazards model while more naturally handling the challenges posed by EHD. Our techniques are motivated by and illustrated on data from a large U.S. Midwestern health care system.

email: **vock@umn.edu**

TRACKING AND PREDICTING DISEASE FROM THE ELECTRONIC MEDICAL RECORD

Joseph Edward Lucas*, Duke University

The systematic collection of electronic medical records is creating new opportunities throughout healthcare. Applications based on EHR data have the potential to disrupt they way that physicians and health care systems interact with patients, the way that clinical research is conducted, and the ability of regulators to encourage medical practice that is focused on patient health. However, medical records are complicated and messy. They incorporate many different types of data, those data often contain mistakes or holes, and they are sometimes subject to abrupt changes in structure and content. We present a statistical approach to jointly analyzing text, categorical and continuous data to model a patient's disease state. The approach allows the collection of multiple sources of evidence into a single coherent picture of disease. We put this together with a parametric family of curves used to describe the progression of health and disease through time (as recorded in the medical record). This model allows us to tie together different health outcomes such as the onset of Alzheimer's disease and future admission into skilled nursing facilities into a single picture of health. We demonstrate our approach on collection of 3.55 million patient visits occurring over a seven year period.

email: **joe@stat.duke.edu**

79. Cost-Effective Study Designs for Observational Data

DESIGN AND ANALYSIS OF RETROSPECTIVE STUDIES FOR LONGITUDINAL OUTCOME DATA

Jonathan S. Schildcrout*, Vanderbilt University School of Medicine

Nathaniel D. Mercaldo, Vanderbilt University School of Medicine

We discuss approaches to examine the modifying effects of single nucleotide polymorphisms on lowering LDL-cholesterol among patients on simvastatin. We consider the setting where longitudinal LDL and covariate data are available prior to study conception; however genotyp-

ing stored blood samples is expensive, and it can only be done on a fraction of patients. We consider methods for determining which subjects should be sampled, and methods for analyzing the data once the biased sample is identified. We compare several designs that sample based on features of the longitudinal outcome and covariate data. Further, we compare several approaches to analyses that are of varying degree of complexity including univariate and longitudinal data analyses. Parameter interpretations and estimation precision will be the focus.

email:
jonathan.schildcrout@vanderbilt.edu

ON THE ANALYSIS OF HYBRID DESIGNS THAT COMBINE GROUP- AND INDIVIDUAL-LEVEL DATA

Sebastien Haneuse*, Harvard School of Public Health

Elizabeth Smoot, Harvard School of Public Health

Ecological studies that make use of data on groups of individuals, rather than on the individuals themselves, are subject to numerous biases that cannot be resolved without some individual-level data. In the context of a rare outcome, the hybrid design for ecological inference efficiently combines group-level data with individual-level case-control data. Unfortunately, except in relatively simple settings, use of the design in practice is limited since evaluation of the hybrid likelihood is computationally prohibitive expensive. In this paper we first propose and develop an alternative representation of the hybrid likelihood. Second, based on this new representation, a series of approximations are proposed that drastically reduce computational burden. A comprehensive simulation shows that, in a broad range of scenarios, estimators based on the approximate hybrid likelihood exhibit the same operating characteristics as the exact hybrid likelihood, without any penalty in terms of increased bias or reduced efficiency. Third, in settings where the approximations may not hold, a pragmatic estimation and inference strategy is developed that uses the approximate form for some likelihood contributions and the exact form for others. The strategy gives researchers the ability to balance computational tractability with accuracy in their own settings. Finally, as a by-product of the development, we provide the first explicit characterization of the hybrid aggregate data design which combines data from an aggregate data study (Prentice and Sheppard, 1995) with case-control samples. The methods are illustrated using data from North Carolina on births between 2007 and 2009.

email: **shaneuse@hsph.harvard.edu**

TEST-DEPENDENT SAMPLING DESIGN AND SEMI-PARAMETRIC INFERENCE FOR THE ROC CURVE

Haibo Zhou*, University of North Carolina, Chapel Hill

Beth Horton, University of Virginia

The receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) are used to describe the ability of a screening test to discriminate between diseased and non-diseased subjects. Evaluating the true disease status can be costly. We develop a test dependent sampling (TDS) design where TDS inclusion depends on a continuous screening test measure. We propose semi-parametric empirical likelihood estimators for the AUC, partial AUC, and the covariate-specific ROC curve to avoid making distributional assumptions about the data. These methods are especially beneficial in situations where the disease status is costly to ascertain or when the length of time between the screening test and the outcome of interest is long. This cost-effective sampling design allows for a more powerful study on the same budget.

email: **zhou@bios.unc.edu**

80. Advanced Machine Learning Methods

A NEW APPROACH TO VARIABLE SELECTION VIA ALGORITHMIC REGULARIZATION PATHS

Yue Hu, Rice University

Genevera I. Allen*, Rice University and Baylor College of Medicine

Variable selection for high-dimensional problems yields a trade-off between statistical and computational efficiency. Penalization methods such as the LASSO solve a relaxation of the best subsets problem that run in polynomial time, but only achieve provable statistical guarantees for selection in limited settings. Is possible to achieve better statistical performance for variable selection in a computationally faster manner? We answer in the affirmative, introducing a new approach to variable selection that we term Algorithmic Regularization Paths. Our method quickly finds a sequence of sparse models, similar in spirit to regularization paths, by solving a series of linked

subproblems associated with the Alternating Direction Methods of Multipliers (ADMM) algorithm. Here, we introduce this novel method, provide intuition for where the algorithm originates, study its theoretical properties, and draw connections to existing regularization methods. Empirical results show that our Algorithm Paths yield better performance in terms of variable selection and prediction error and are moreover computationally faster than all existing approaches.

email: **gallen@rice.edu**

LINK PREDICTION FOR PARTIALLY OBSERVED NETWORKS

Yunpeng Zhao, George Mason University

Yun-Jhong Wu, University of Michigan

Elizaveta Levina, University of Michigan

Ji Zhu*, University of Michigan

Link prediction is one of the fundamental problems in network analysis. In many applications, notably in genetics, a partially observed network may not contain any negative examples of absent edges, which creates a difficulty for many existing supervised learning approaches. We develop a new method which treats the observed network as a sample of the true network with different sampling rates for positive and negative examples. We obtain a relative ranking of potential links by their probabilities, utilizing information on node covariates as well as on network topology. Empirically, the method performs well under many settings, including when the observed network is sparse. We apply the method to a protein-protein interaction network and a school friendship network.

email: **jizhu@umich.edu**

GRAPHICAL REGRESSION

Hsin-Cheng Huang, Academia Sinica, Taiwan

Xiaotong Shen*, University of Minnesota

Wei Pan, University of Minnesota

Graphical models have proven useful in describing relations among interacting units. In this paper, we propose graphical models to link the inverse of covariance matrix, called the precision matrix, to covariates, to locate the structural change of a graph as a function of covariates. For instance, in neuroimaging, functional connectivity of regions of interest concerns the relationship between brain activity and specific mental functions. To investigate the structural change of a graph based on regression coefficients, we construct a constrained likelihood, where sparsity constraints are imposed to seek low rank representations. Computational aspects will be discussed in addition to some theoretical aspects.

email: **xshen@umn.edu**

PENALIZED MAXIMUM LIKELIHOOD ESTIMATION ON A TWO-LAYERED NETWORK

George Michailidis*, University of Michigan

Networks are one of the most popular tools for capturing the interactions between nodes, which are used to represent the underlying random variables. In particular, constructing and analyzing a layered structure provides insight into understanding the conditional relationships among nodes within layers while adjusting for and quantifying the effects of nodes of a particular layer on another. We propose a new unified approach for estimating a two-layered network. The proposed method offers an efficient way of estimating edges between and across layers iteratively, by constructing an objective function based on the penalized joint maximum likelihood function (under a Gaussianity assumption), then using block co-ordinate descent to do the optimization. Our method decouples the estimation of undirected and directed edges within each iteration, however the updated estimates are integrated in the next iteration. The performance of the methodology is illustrated via simulations. Applications to Omics problems are also briefly discussed.

email: **gmichail@umich.edu**

81. Statistical Analysis for Deep Sequencing Data in Cancer Research: Methods and Applications

A STATISTICAL METHOD FOR DETECTING DIFFERENTIALLY EXPRESSED MUTATIONS BASED ON NEXT-GENERATION RNAseq DATA

Pei Wang*, Icahn School of Medicine at Mount Sinai

Rong Fu, University of Washington

Ziding Feng, University of Texas MD Anderson Cancer Center

We propose a new statistical method – MutRSeq – for detecting single nucleotide variants (SNVs) differentially expressed in samples with different disease status based on RNA-seq data.

Specifically, we employ a hierarchical likelihood approach to jointly model observed mutation events and read count measurements from RNAseq experiments. We then introduce a likelihood ratio based test statistic, which detects changes not only in overall expression levels, but also in allele specific expression patterns. In addition, this method can jointly test multiple mutations in one gene/pathway. The simulation studies suggest that the proposed method achieves better power than a few competitors under a range of different settings. In the end, we apply this method to a non-smoker lung cancer data set and identify potential disease genes.

email: **pei.wang@mssm.edu**

ACCOUNTING FOR DIFFERENTIAL COVERAGE IN COMPARING MUTATION PREVALENCE

George W. Wright*, National Cancer Institute, National Institutes of Health

RNA-seq is a powerful tool in understanding the biological mechanisms underlying cancer and other diseases. By comparing the prevalence of a given mutation between subsets of samples, it is possible to gain insight into the molecular basis for observed phenotypic differences. However, the ability to detect mutations in samples at a given location will depend on their coverage at that location, so that simply counting the proportion of samples in which the mutation is observed may underestimate the true mutation prevalence in locations of low coverage. In RNA-seq, coverage depends on gene expression, and since compared subtypes may have very different gene expression profiles, it is possible that differences between subsets in observed mutation frequency may be due to differences in expression rather than actual differences in prevalence. We present an empirical model for detection frequency as a function of coverage, which we use to estimate the true mutation prevalence within a given subset of cases. Further, we provide a test for differential prevalence which takes into account the possibility of differential coverage between the subsets.

email: **wrightge@mail.nih.gov**

SCALABLE BAYESIAN NONPARAMETRIC LEARNING FOR HIGH-DIMENSIONAL LUNG CANCER GENOMICS DATA

Chiyu Gu, University of Missouri

Subharup Guha*, University of Missouri

Veerabhadran Baladandayuthapani, University of Texas MD Anderson Cancer Center

Through array-based and next-generation sequencing, 'omics datasets involve intrinsically different sizes and scales of high-throughput data, providing genome-wide, high-resolution information about the biology of lung cancer. A common goal is the identification of differential genomic signatures between samples that correspond to different treatments or biological conditions, e.g., treatment arms, tumor (sub)types, or cancer stages. We construct an encompassing class of nonparametric models called generalized Poisson-Dirichlet processes (g-PDPs) that are applicable to mixed, heterogeneously scaled datasets. Each platform can choose from diverse parametric and nonparametric models, which include finite mixtures, finite and infinite hidden Markov models, Dirichlet processes, and zero and first order PDPs that cover a broad range of data correlation structures. An outstanding feature of g-PDP models is their ability to infer in an unsupervised manner and with high accuracy, latent clusters of genes that may represent joint regulatory mechanisms. Simulation studies demonstrate that g-PDPs outperform many existing techniques in terms of accuracy of signature identification. The detected disease genomic signatures in RNA-seq lung cancer TCGA data highlight the ability of g-PDP models to handle multiple conditions without Bonferroni-type adjustments. The pathway analysis identified upstream regulators of many genes that are common genetic markers in multiple tumor cells.

email: **guhasu@missouri.edu**

UNDERSTANDING MicroRNA SEQUENCING DATA DISTRIBUTION

Li-Xuan Qin*, Memorial Sloan Kettering Cancer Center

Tom Tuschl, Rockefeller University

Sam Singer, Memorial Sloan Kettering Cancer Center

MicroRNAs (miRNAs) are a prevalent class of small single-stranded non-coding RNAs that negatively regulate gene expression. MiRNAs are involved in a wide variety of cellular functions such as proliferation, differentiation, and apoptosis. Their roles in carcinogenesis are being increasingly studied with the advent of the deep sequencing technology. Comparing with mRNA sequencing, miRNA sequencing has the advantages

of relatively homogeneous gene length and more straightforward gene mapping and assembly; at the same time, it faces the unique challenges that miRNA abundance has a broad range among genes and its distribution can be highly skewed among samples. In order to better understand miRNA sequencing data distribution, we carried out a miRNA sequencing study at Memorial Sloan Kettering Cancer Center using tumor samples including both biological and technical replicates. We examined the distributional patterns of technical and biological variations and we proposed a novel statistical model for miRNA sequencing data.

email: **qinl@mskcc.org**

82. Spatial and Spatio-Temporal Modeling

SPATIAL LOCAL GRADIENT MODELS OF BIOLOGICAL INVASIONS

Joshua Goldstein, The Pennsylvania State University

Murali Haran*, The Pennsylvania State University

Ottar N. Bjornstad, The Pennsylvania State University

Andrew M. Liebhold, U.S. Forest Services

Understanding the processes that influence the spread of biological invasions is key to developing effective strategies for intervention. Of particular interest is the ability to quantify spread rates and to understand factors that affect invasion speed. We also need to identify geographic patterns of spread, for instance distinguishing highly directional "wave-like" spread versus spread that is radially outward, indicating the source of the invasion or the site of a sudden long-range dispersal. We describe methodology that addresses these questions, building largely on Gaussian process models for observations of the time of first infestation. Our methods allow us to estimate local speed and direction of spread while also identifying key invasion features in rigorous fashion. We illustrate the application of our methods to two biological invasions, the gypsy moth and the emerald ash borer. This is joint work with Joshua Goldstein, Ottar Bjornstad and Andrew Liebhold.

email: **mharan@stat.psu.edu**

A GENERALIZED CONDITIONALLY AUTOREGRESSIVE (CAR) MODEL

Veronica J. Berrocal*, University of Michigan

Alan E. Gelfand, Duke University

Spatial areal data is encountered in a broad range of applications. Within the Bayesian framework, this type of data is most commonly analyzed by introducing, in the second stage of a hierarchical model, spatial random effects modeled as a Gaussian Markov random field, specified locally through a conditionally autoregressive (CAR) model (Besag 1974). The key specification in a CAR model is the proximity matrix, W, with entries wij , proximities that encode the strength of association among the various areal units. The most frequently adopted choice for the proximities is to assume that they are binary and based upon some notion of adjacency. In this paper, we propose an extension of the binary adjacency proximities CAR model where the proximities, defined through a suitable transformation of a latent Gaussian process, are random and directional, thus allowing for varying strength of association among an areal unit and its neighbors. Our specification of the proximities allows us to derive distributional properties of the proximities and of the spatial random effects, and leads to tractable Bayesian inference with closed form full conditionals. In the case of large dimensional datasets, it is possible to introduce dimension reduction for the proximities to alleviate computational burden.

email: **berrocal@umich.edu**

MULTIVARIATE SPATIAL MODELING OF CONDITIONAL DEPENDENCE IN MICROSCALE SOIL ELEMENTAL COMPOSITION DATA

Joseph Guinness*, North Carolina State University

Montserrat Fuentes, North Carolina State University

Dean Hesterberg, North Carolina State University

Matthew Polizzotto, North Carolina State University

Elevated concentrations of toxic trace elements pose threats to human health through contamination of food and drinking water. We describe an experiment that maps the composition of elements

on a sand grain using X-ray fluorescence microprobe analyses, before and after the grain is treated with arsenic solutions, resulting in multivariate spatial lattice maps of elemental abundance. To understand the behavior of arsenic in soils, it is important to disentangle the complex multivariate relationships among the elements in the sample. The abundance of most elements, including arsenic, correlates strongly with that of iron, but conditional on the amount of iron, some elements may mitigate or potentiate the accumulation of arsenic. This problem motivates our work to define conditional correlation in spatial lattice models and give general conditions under which two components are conditionally uncorrelated given the rest. We describe how to enforce that two components are conditionally uncorrelated given a third in parametric models, which provides a basis for likelihood ratio tests of conditional correlation between arsenic and chromium given iron. We show how to apply our results to big datasets using the Whittle likelihood, and we demonstrate through simulation that tapering improves Whittle likelihood parameter estimates governing cross covariance.

email: **jsguinne@ncsu.edu**

GAUSSIAN PROCESS MODELS FOR EMULATING SPATIAL COMPUTER MODEL OUTPUT

Dave M. Higdon*, Los Alamos National Laboratory and Virginia Tech

Mengyang Gu, Duke University

Gaussian process models have proven to be very useful in modeling computer simulation output. This is because many computer codes are essentially noiseless and respond very smoothly to changes in input settings. In a typical setting, the simulation output is a function of a p-dimensional input vector x. In some applications, p may be as large as 60, however for the application of interest, the output is typically affected by a much smaller subset of these p inputs. After a fixed number of simulations are carried out, a GP model can be used to predict the simulation output at untried settings. Recently we have encountered situations where the computer model output is a spatial field. The high-dimensionality of this output, along with its spatial dependence, leads to a number of interesting issues in computer model emulation and calibration. We compare a number of different approaches for GP-based emulation for spatial fields, using two applications – one in nuclear physics, and one in vulcanology.

email: **dhigdon@lanl.gov**

83. CONTRIBUTED PAPERS: Study Design and Power

COMPARISON OF RISK ESTIMATES DERIVED FROM FULL COHORT, SUB-SAMPLE, AND NESTED CASE-COHORT METHODOLOGIES

Kathleen A. Jablonski*, The George Washington University

Madeline M. Rice, The George Washington University

Drawing subsamples from a large study is often done to measure biomarkers as a cost savings measure. We compared risk estimates from three sampling designs that were utilized in a completed trial; the full cohort, a subsample, and a nested case-cohort design. The subsample represents participants that provided samples for ancillary studies. Sub-cohort controls were weighted by the inverse of the sampling fraction. We used proportional hazard models, stratified by treatment group and adjusted for known confounders, to examine the association between baseline eGFR and morality in each sampling design. Demographic and baseline characteristics were similar between the full cohort and subsample. We show that estimates from all three sampling designs agree reasonably well. These results lend a measure of confidence in these designs.

email: **kjablons@gwu.edu**

POWER ESTIMATION FOR ORDINAL CATEGORICAL DATA IN THE PRESENCE OF NON PROPORTIONAL ODDS

Roy N. Tamura*, University of South Florida

Xiang Liu, University of South Florida

Ordinal categorical data are extremely common in clinical trials where subjective outcomes are being measured. The most common analyses for ordinal categorical data are based on the proportional odds model. However, the assumption of proportional odds is an assumption of the model and not necessarily a property of the data. In many situations, clinical researchers may suspect that an effect on an ordinal scale will not satisfy the proportional odds assumption. In these situations, statisticians need to be able

to investigate alternative models and estimate the power for these alternative models. We examine the trend odds model and the saturated model and compare these models to the proportional odds model under a wide class of alternative hypotheses. We also examine the exemplary dataset (ED) approach for power estimation for these models and determine how well this approach approximates the actual power. A recently developed ordinal scale for assessing nausea in the pediatric population illustrates the issues and recommendations.

email: **roy.tamura@epi.usf.edu**

SINGLE ARM PHASE II CANCER SURVIVAL TRIAL DESIGNS

Jianrong John Wu*, St. Jude Children's Research Hospitial

In this talk, a modified one-sample log-rank test statistic is proposed. In general, the proposed test can be used to design single-arm phase II survival trials under any parametric survival distribution. Simulation results showed that it preserves type I error well and provides adequate power for phase II cancer survival trial designs.

email: **jianrong.wu@stjude.org**

EMPIRICAL DETERMINATION OF STATISTICAL POWER AND SAMPLE SIZE FOR RNA-Seq STUDIES

Milan Bimali*, University of Kansas Medical Center

Jonathan D. Mahnken, University of Kansas Medical Center

Brooke L. Fridley, University of Kansas Medical Center

RNA-Seq studies produce count-based data that do not follow normal distribution, but rather Poisson or Negative-Binomial distribution. In designing RNA-Seq studies, the estimate of sample size to achieve a desired statistical power is important. Several methods have been proposed in the literatures that are designed specifically for RNA-Seq studies, however none have attracted consensus yet. The mean-variance relationship in Poisson and Negative-Binomial distribution has been one of the challenges in deriving exact analytic form. Proposed methods have been based on large sample approximation thereby making the use of such estimates for small scale RNA-Seq studies questionable. We propose a simulation based approach for estimating power for given sample size under the Poisson and Negative-Binomial assumption. Our method does not assume large sample approximation. The sample sizes based on our approach are compared with existing methods. The simulation based sample size estimates based on the Poisson distribution were smaller than the estimates computed from large sample Poisson based methods and larger than an approach based on the data following Gaussian distribution. As expected, empirical based sample size estimates based on Negative-Binomial distribution were larger than the estimates computed from Poisson distribution thus accounting for the over-dispersion observed in RNA-Seq data.

email: **mbimali@kumc.edu**

FUNCTIONAL SIGNAL-TO-NOISE RATIO ANALYSIS WITH APPLICATIONS IN QUANTITATIVE ULTRASOUND

Yeonjoo Park*, University of Illinois, Urbana-Champaign

Douglas G. Simpson, University of Illinois, Urbana-Champaign

Motivated by research on diagnostic ultrasound to evaluate tissue regions of interest such as tumors and cysts via their ultrasound backscatter properties, this paper develops an approach to functional effect size estimation, testing and visualization. Extending methods from functional analysis of variance we introduce the functional signal-to-noise ratio (fSNR), discuss its use for visualizing the magnitude of effects over the domain of interest, and develop bootstrap inferences based on global summary functions of the fSNR. The approach allows for irregular functional

data in which the ranges of the curves may vary, as long as the full ensemble of curves covers the domain of interest. We also develop simulation based power analysis for the global fSNR based tests. The methods are illustrated in the analysis of irregular functional data from inter-laboratory quantitative ultrasound measurements.

email: **ypark61@illinois.edu**

ANALYSIS OF A NON-MORTALITY OUTCOME IN CLINICAL TRIAL OF A POTENTIALLY LETHAL DISEASE

Roland A. Matsouaka*, Duke University

Rebecca Betensky, Harvard University

We consider a randomized clinical trial of a potentially lethal disease where patients are treated, followed for a fixed period of time, and assessed on a non-mortality outcome. For patients who die before the end of follow-up, the outcome of interest is not assessed. Any statistical analysis based solely on patients who survived can be misleading since survivors may substantially differ from those who died. An alternative approach to assess the treatment effect is to create a composite outcome including death and the non-mortality outcome. For this talk, we examine the use of Wilcoxon—Mann—Whitney test on such worst-rank composite outcomes, where the patients who survived are ranked based on the magnitude of their responses while those who died are assigned "worst-rank" scores. These scores are (chosen) worse than any observed responses: they are either (1) set to a single value or (2) rank based the time of death, an earlier death being worse than a later death. We evaluate the power of the test through Monte Carlo simulation studies and apply our method to a clinical trial of the effects of normobaric oxygen therapy on patients who had an acute ischemic stroke.

email: **roland.matsouaka@duke.edu**

SAMPLE SIZE DETERMINATION BASED ON QUANTILE RESIDUAL LIFE

Jong Hyeon Jeong*, University of Pittsburgh

In the analysis of time-to-event data, the concept of residual life provides straightforward interpretation, and has recently drawn much attention in the literature. Clinical studies are usually designed based on the hazard rates under the proportional hazards model. In this presentation, we consider sample size determination to detect a difference in quantile residual lifetimes between two groups, given operating characteristics. The results are compared to ones calculated from the hazard rate approach. An extension to a competing risks case is also considered.

email: **jeong@nsabp.pitt.edu**

84. CONTRIBUTED PAPERS: Missing Data

A MIXED EFFECTS MODEL FOR INCOMPLETE DATA WITH EXPERIMENT-LEVEL ABUNDANCE-DEPENDENT MISSING-DATA MECHANISM

Lin S. Chen, University of Chicago

Jiebiao Wang*, University of Chicago

Xianlong Wang, Fred Hutchinson Cancer Research Center

Pei Wang, Icahn Medical School at Mount Sinai

Nonignorable missing data exist in the iTRAQ (isobaric tag for relative and absolute quantitation) proteomic experiment. The missing mechanism in the data is defined as experiment-level abundance-dependent missing-data mechanism (EADMM). We propose a new method, mixEMM, to explicitly model the missing mechanism, using the expectation conditional maximization (ECM) algorithm under the setting of mixed effects model. The performance of the proposed method is evaluated in simulation studies and in a proteomic data illustration. We also discuss some extensions of this approach in the end.

email: **jwang88@uchicago.edu**

MULTIPLE IMPUTATION FOR GENERAL MISSING PATTERNS IN THE PRESENCE OF HIGH-DIMENSIONAL DATA

Yi Deng*, Emory University

Qi Long, Emory University

Zhao and Long (2013) investigated several approaches of using regularized regression and Bayesian lasso regression to conduct multiple imputation (MI) in the presence of high-dimensional data, though their methods are not directly applicable to the case of general missing patterns. Based on the technique of chained equations, we extend their methods to handle general missing patterns in the presence of high-dimensional data and implement our methods in an R

package. The proposed MI methods are evaluated in extensive simulation studies and are further illustrated using a data set from the Georgia Coverdell Acute Stroke Registry.

email: **ydeng26@emory.edu**

A MIXED-EFFECTS MODEL FOR NONIGNORABLE MISSING LONGITUDINAL DATA

Xuan Bi*, University of Illinois, Urbana-Champaign

Annie Qu, University of Illinois, Urbana-Champaign

Nonignorable missing data occurs frequently in longitudinal studies. Estimation bias may arise if a missing mechanism is misspecified. To address this issue, we introduce a mixed-effects estimating equation approach, which enables one to recover missing information from the measurement process and the missing process simultaneously. The proposed method proves consistency and asymptotic normality of the fixed-effect estimation under shared-parameter models and an extended shared-parameter model. In simulation studies, we show the effectiveness of the proposed method under different missing patterns in conjunction with robustness against model assumption violation. In addition, it is applied to the election poll survey data from 2007-2008 Associated Press-Yahoo! News which involves multiple refreshment samples.

email: **xuanbi2@illinois.edu**

EM ALGORITHM IN GAUSSIAN COPULA WITH MISSING DATA

Wei Ding*, University of Michigan

Peter X. K. Song, University of Michigan

Rank-based correlation is widely used to measure dependence between variables when their marginal distributions are skewed. Estimation of such correlation is challenged by both the pres ence of missing data and the need for adjusting for confounding factors. In this paper, we consider a unified framework of Gaussian copula regression that enables us to estimate either Pearson cor- relation or rank-based correlation (e.g. Kendall's tau or Spearman's rho), depending on the types of marginal distributions. To adjust for confounding covariates, we utilise marginal regression models with univariate location-scale family distributions. We establish the EM algorithm for estimation of both correlation and regression parameters with missing values. For implementation, we propose an effective peeling procedure to carry out iterations required by the EM algorithm. We compare the performance of the EM algorithm method to the traditional multiple imputation approach through simulation studies. For structured types of correlations, such as exchangeable or first-order auto-regressive (AR-1) correlation, the EM algorithm outperforms the multiple imputation approach in terms of both estimation bias and efficiency.

email: **dingwei@umich.edu**

ON IDENTIFICATION ISSUES WITH BINARY OUTCOMES MISSING NOT AT RANDOM

Jiwei Zhao*, University at Buffalo, SUNY

In epidemiology, regression models with binary outcomes are often used to investigate the relation between disease status and other exposures or covariates of interest, especially for case control studies. Clinically, these studies usually encounter the problem of missing disease status and the missing data mechanism is highly suspected to be nonignorable (Little and Rubin, 2002). Therefore, we have to be careful resolving the identifiability conditions of the unknown parameters for each method we use (Robins, 1997). In this paper, we systemically study the identifiability conditions with missing response data when the mechanism is nonignorable. Although we focus on logistic regression and probit regression, the theory can be extended to other generalized linear models. Comprehensive simulation studies and a real data analysis are conducted to illustrate our theory and method.

email: **zhaoj@buffalo.edu**

KENWARD-ROGER APPROXIMATION FOR LINEAR MIXED MODELS WITH MISSING COVARIATES

Akshita Chawla*, Michigan State University

Tapabrata Maiti, Michigan State University

Samiran Sinha, Texas A&M University

Partially observed variables are common in scientific research. Ignoring the subjects with partial information may lead

to biased and or inefficient estimators, and consequently any test based only on the completely observed subjects may inflate the error probabilities. Missing data issue has been extensively considered in the regression model, especially in the independently identically (IID) data setup. Relatively less attention has been paid for handling missing covariate data in the linear mixed effect model-- a dependent data scenario. In case of complete data, Kenward-Roger's F test is a well-established method for testing of fixed effects in a linear mixed model. In this paper, we present a modified Kenward-Roger type test for testing fixed effects in a linear mixed model when the covariates are missing at random. In the proposed method, we attempt to reduce bias from three sources, the small sample bias, the bias due to missing values, and the bias due to estimation of variance components. The operating characteristics of the method is judged and compared with two existing approaches, listwise deletion and mean imputation, via simulation studies.

email: **chawlaak@stt.msu.edu**

NONPARAMETRIC SEQUENTIAL MULTIPLE IMPUTATION FOR SURVIVAL ANALYSIS WITH MISSING COVARIATES

Paul Hsu, University of Arizona

Mandi Yu*, National Cancer Institute, National Institutes of Health

Cancer registry data has been the cornerstone for monitoring cancer survival at the population level. However, the presence of censoring and missing covariates could cause efficiency losses in estimating survival statistics and pose potential risks for bias if censoring depends upon survival time or the missing covariates are not ignorable. This presentation proposes a nonparametric approach to impute censored survival or missing covariates by substituting with information from a matched observation with complete data (called donor). This procedure is applied sequentially for one variable at a time for multiple rounds. The completed data can be analyzed as if there were no missing or censoring data. This approach is appealing because it utilizes two working models, one predicts the variable subjecting to missing or censoring and the other predicts the missingness of that variable, to define the donor pool, such that the imputed value is double robust against possible incorrect assumptions about the missing mechanisms. Multiple imputation is used to propagate imputation uncertainty. The approach is applied to cancer registry data for assessing the survival and incidence of breast cancer by HER2 status. The performance of the proposed method is assessed by a simulation study.

email: **yum3@mail.nih.gov**

85. CONTRIBUTED PAPERS: Innovative Methods for Clustered Data

CORRELATION STRUCTURE SELECTION PENALTIES FOR IMPROVED INFERENCE WITH GENERALIZED ESTIMATING EQUATIONS

Philip M. Westgate*, University of Kentucky

Woodrow W. Burchett, University of Kentucky

Generalized estimating equations (GEE) are often used for the marginal analysis of correlated data. With GEE, a working correlation structure must be selected. Accurate modeling of this structure can improve efficiency. However, estimation of correlation parameters can inflate the covariance matrix of the regression parameter estimates, which should be accounted for, or penalized by, criteria that are used to select a working structure. We therefore discuss different penalties, and give practical considerations for data analysts on how these penalties may influence regression parameter estimation.

email: **philip.westgate@uky.edu**

HANDLING NEGATIVE CORRELATION AND/OR OVERDISPERSION IN GAUSSIAN AND NON-GAUSSIAN HIERARCHICAL DATA

Geert Molenberghs*, Hasselt University and Leuven University

Non-negative correlation and overdispersion are phenomena that received a lot of study, separately and in conjunction.

As a result, a large suite of modeling approaches has been proposed, for about three decades now. That said, also negative correlation is perfectly possible in real-life biometric and other experiments. The same is true for underdispersion. Allowing for these in a flexible and elegant modeling approach, preferably with hierarchical intepretation, is less than straightforward. We sketch the problem, offer modeling approach, and discuss implications for such tasks as: model formulation, parameter and precision estimation, hypothesis testing (in a marginal and hierarchical fashion). Results synthetized encompass both historic findings as well as very recent results.

email: **geert.molenberghs@uhasselt.be**

REFLECTING THE ORIENTATION OF TEETH IN RANDOM EFFECTS MODELS FOR PERIODONTAL OUTCOMES

Rong Xia*, University of Michigan

Thomas M. Braun, University of Michigan

William V. Giannobile, University of Michigan

Clinical attachment level (CAL) is a tooth-level measure that quantifies the severity of periodontal disease. The within-mouth correlation of tooth-level measures of CAL is difficult to model because it must reflect the three-dimensional spatial geography of teeth and their functional similarity. Thus, traditional approaches have included (1) applying a t-test/regression model to mouth-level averages or (2) using generalized estimating equations (GEE) with simple correlation structures. As an alternative, we propose two linear mixed models with random effects that quantify the within-mouth correlation of teeth and their shared functionality. Via simulation, we compare the bias and efficiency of fixed effect 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**

DETECTING HETEROGENEITY BASED ON EFFECT SIZE OF RESPONSE MEASURES

Xin Tong*, University of South Carolina, Columbia

This study was motivated by the potential heterogeneity in clusters identified by pre-existing methods which are designed to separate heterogeneous data into groups of similar objects such that objects within a group are similar. 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. To improve the quality of clusters, we propose a clustering method which can cover skewed and fat tailed error distribution. Specifically, the variables are clustered based on the agreement of relationships (unknown) 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 clustering. Simulation studies are used to examine the performance and efficiency of the proposed method. The method is then applied to a population of patients evaluated for laser refractive surgery to further understand the effect of aging on measurements of higher-order aberrations.

email: **cuzntone@gmail.com**

STATISTICAL METHODS FOR MANIFOLD-VALUED DATA FROM LONGITUDINAL STUDIES

Emil A. Cornea*, University of North Carolina, Chapel Hill

Hongtu T. Zhu, University of North Carolina, Chapel Hill

Joseph G. Ibrahim, University of North Carolina, Chapel Hill

The aim of the paper is to present a general regression framework for the analysis of manifold-valued data from longitudinal studies. We develop semi-parametric intrinsic random effect regression models for analyzing manifold longitudinal data. We focus on directional data, symmetric positive definite (SPD) matrices, and landmark-based planar shapes rising from manifold-valued imaging data to illustrate our methodological development. We

apply our semi-parametric models to the shape analysis of the corpus callosum from longitudinal studies and investigate whether the corpus callosum shape information is a potential biomarker for the diagnosis of Alzheimer's disease and attention deficit hyperactivity disorder.

email: **ecornea@bios.unc.edu**

ANALYZING DEPENDENT DATA USING EMPIRICAL LIKELIHOOD AND QUADRATIC INFERENCE FUNCTION

Chih-Da Wu*, University of North Carolina, Chapel Hill

Naisyin Wang, University of Michigan

Dependent data can be modeled under generalized estimating equation framework. However there are always same number of estimating functions as the number of parameters and the correlation matrix to describing the dependent struction is often unknown. Even though the sandwich estimator method shows that the parameter estimation using a working correlation structure is still consistent. In practice, people still want to cooperate the partial information about dependent structure by assuming the covariance matrix to have specific simpler format that can be represented by not so many argumanet. For example, compound symmetry stucture needs only one argument and auto-regressive needs two. Quadratic inference function is the method to cooperate such information by creating multiple numbers of estimating function, more than the number of parameters. For estimating equation models with more estimating functions than parameters, empirical likelihood can provide a robust parameter estimation. In literature, such combination of empirical likelihood and quadratic inference function already exist. However, only a few type of correlation structure are mentioned. Such limitation is built-in in the quadratic inference function method because it requires the inverse of correlation matrix of specific format can be decomposed into linear combination of basis matrix. Such decomposition is not always viable. In this paper, we will investigate a general stretagy of applying quadratic inference function method when the assumed structure are not limited to compound symmetry or AR(1). Simulation study will be provided to comparing the performance of several methods. And a functional data consisted of diffusino tensor tract statistics, preocessed from neuroimaging dataset will be analyzed to show the application of our proposed strategy.

email: **chihdawu@email.unc.edu**

FAST ESTIMATION OF REGRESSION PARAMETERS IN A BROKEN STICK MODEL FOR LONGITUDINAL DATA

Ritabrata Das*, University of Michigan

Moulinath Banerjee, University of Michigan

Bin Nan, University of Michigan

Estimation of change-point(s) in the broken-stick model has significant applications in modeling important biological phenomena. In this article we present a computationally economical likelihood-based approach for estimating change-point(s) efficiently in both cross-sectional and longitudinal settings. Our method, based on local smoothing in a shrinking neighborhood of each change-point, is shown via simulations to be computationally more viable than existing methods that rely on search procedures, with dramatic gains in the multiple change-point case. The proposed estimates
are shown to have n-consistency and asymptotic normality; in particular,they are asymptotically efficient in the cross-sectional setting allowing us to provide meaningful statistical inference. As our primary and motivating (longitudinal) application, we study the Michigan Bone Health and Metabolism Study cohort data to describe patterns of change in log estradiol levels, before and after the final menstrual period, for which a two change-point broken stick model appears to be a good fit.

email: **ritob@umich.edu**

86. CONTRIBUTED PAPERS: Biopharmaceutical Applications and Survival Analysis

PSEUDO-VALUE APPROACH FOR TESTING CONDITIONAL RESIDUAL LIFETIME FOR DEPENDENT SURVIVAL AND COMPETING RISKS DATA

Kwang Woo Ahn*, Medical College of Wisconsin

Brent R. Logan, Medical College of Wisconsin

Quantile residual lifetime analysis is often performed to evaluate the distributions of remaining lifetimes for survival and competing risks data. Residual lifetimes may depend on some patients' characteristics. In addition, the event times and the censoring times of survival and competing risks data are often clus-

tered or correlated. Thus, it is crucial to develop statistical methods for assessing conditional residual lifetimes for dependent survival and competing risks data. The current literature on quantile residual lifetime analysis is restricted to independent survival data. We propose a pseudo-value approach to compare conditional residual lifetimes for independent/dependent survival and competing risks data. The pseudo-values are obtained based on jackknife using the Kaplan-Meier estimates for survival data and the cumulative incidence estimates for competing risks data. Statistical inference for comparing conditional residual lifetimes is made by relying on generalized estimating equations to account for correlation among patients. The statistical properties of the proposed method are studied. Simulation studies show that the proposed method controls Type I errors very well. The proposed method is illustrated by a bone marrow transplant data set.

email: **kwooahn@mcw.edu**

FALLBACK TYPE FDR CONTROLLING PROCEDURES FOR TESTING A PRIORI ORDERED HYPOTHESES

Anjana Grandhi*, New Jersey Institute of Technology

Gavin Lynch, New Jersey Institute of Technology

Wenge Guo, New Jersey Institute of Technology

In large scale multiple testing problems in applications such as stream data, statistical process control, etc., the tested hypotheses are ordered a priori by time and it is desired to control False Discovery Rate (FDR) while making real time decisions about rejecting or accepting a hypothesis. The existing FDR controlling procedures, such as Benjamini Hochberg procedure, are not applicable here as they are unable to make real time decisions based on incomplete information pertaining to the p-values available. In this talk, I present a powerful fallback type procedure for controlling FDR that awards the critical constants on rejection of a hypothesis and penalizes on acceptance. This procedure overcomes the drawback of the conventional FDR controlling procedures by making real time decisions based on partial information available when a hypothesis is tested and allowing testing of each a priori ordered hypothesis. The procedure is shown to strongly control FDR under positive regression dependence (PRDS) of p-values. FDR control under arbitrary dependence can also be achieved by applying a correction factor to the critical constants. A Simulation study demonstrates effectiveness of the procedure in terms of FDR control and average power. A real data analysis is presented that supports our findings.

email: **ag454@njit.edu**

PARAMETRIC INFERENCE ON QUANTILE RESIDUAL LIFE

Kidane B. Ghebrehawariat*, University of Pittsburgh

Ying Ding, University of Pittsburgh

Jong-Hyeon Jeong, University of Pittsburgh

Statistical inference via quantile residual life enjoys some practical advantages over other existing approaches (e.g., hazard based or mean based) to time-to-event analyses. Most of the proposed inference procedures for quantile residual life in the literature are non-parametric or semi-parametric. However, parametric approaches are expected to be asymptotically efficient under a correct specification of the model. Furthermore, the parametric approach does not require nonparametric estimation of the probability density function of the underlying distribution under informative or noninformative censoring to evaluate the variance of the quantile estimator. In this presentation we develop parametric inferences on the quantile residual lifetimes for one-sample and two sample cases both under competing and non-competing risks settings. Simulation results indicate that the methods perform well. The proposed methods will be illustrated with a real dataset.

email: **kig11@pitt.edu**

STUDY DESIGN ISSUES IN PRECISION STUDY FOR OPTICAL COHERENCE TOMOGRAPHY DEVICE

Haiwen Shi*, U.S. Food and Drug Administration

Optical Coherence Tomography (OCT) is a new medical imaging technology and more eye specialists are relying on it to diagnose the Ocular diseases, such as Diabetic Retinopathy. It is critical for OCT to accurately monitor the thickness of eye anatomy such as retina, which is the indication of some ocular disease. Hence, the OCT device needs to have good precision in the measurement of the thickness. A typical precision study for OCT device uses either a nested or crossed design. For the economic rea-

son, sponsors prefer the nested design, especially when the study involves multiple sites. In this talk, I will discuss my investigation of the consequence if the random effect model that is used to obtain the variance component is mis-specified. Specifically, I investigate the consequences of fitting the nested design data by a crossed model and the consequences of treating some random effect factors as fixed effect using simulations and a data example. In addition, I will discuss some other issues related to the precision study for the OCT device. This includes whether an interaction term should be included in a crossed ANOVA random effect model and the problem of negative variance component estimates and imbalanced data.

email: **haiwen.shi@fda.hhs.gov**

MODELING GAP TIMES BETWEEN RECURRENT INFECTIONS AFTER HEMATOPOIETIC CELL TRANSPLANT

Chi Hyun Lee*, University of Minnesota

Xianghua Luo, University of Minnesota

Chiung-Yu Huang, Johns Hopkins University

Recurrent infections after transplantation can cause significant morbidity in hematopoietic cell transplantation (HCT) recipients. Patients who received HCT at the University of Minnesota Fairview Hospital had been monitored for various types of infectious complications, including bacterial, viral, and fungal infections. The effects of patient- and transplant-related characteristics on the interoccurrence times or gap times between infections of each type are often of scientific interest. However, there are certain difficulties in modeling the recurrent gap time data of infections after transplant. First, there are multiple types of infections, hence multivariate recurrent event processes, which could be correlated due to the shared, compromised host immunity. Second, the effects of covariates on different episodes of gap times of the same event type could be different. Hence, a model which can handle episode-specific effect is necessary. Third, the observation of recurrent infection processes may be terminated by events such as death and a second transplant. Ignoring the possible dependence of recurrent event processes with the terminal events could lead to incorrect inferential results. In this paper, we will present the analysis of post-HCT infection data prospectively collected from patients who received HCT between 2000 and 2010 at the University of Minnesota.

email: **leex5865@umn.edu**

ASSESSING TREATMENT EFFECTS WITH SURROGATE SURVIVAL OUTCOMES USING AN INTERNAL VALIDATION SUBSAMPLE

Jarcy Zee*, Arbor Research Collaborative for Health

Sharon X. Xie, University of Pennsylvania

In studies with surrogate outcomes available for all subjects and true outcomes available for only a subsample, survival analysis methods are needed that incorporate both endpoints in order to assess treatment effects. We develop a semiparametric estimated likelihood method for the proportional hazards model with discrete time data and a binary covariate of interest. Our proposed estimator is consistent and asymptotically normal. Through numerical studies, we showed that our proposed method for estimating a covariate effect is unbiased compared to the naive estimator that uses only surrogate endpoints and is more efficient with moderate missingness compared to the complete-case estimator that uses only true endpoints. We also illustrated the use of our proposed method by estimating the effect of gender on time to detection of Alzheimer's disease using data from the Alzheimer's Disease Neuroimaging Initiative. The proposed method is able to account for the uncertainty of surrogate outcomes by using a validation subsample of true outcomes in estimating a binary covariate effect. The proposed estimator can outperform standard semiparametric survival analysis methods, and can therefore save on costs of a trial or improve power in detecting treatment effects.

email: **Jarcy.Zee@arborresearch.org**

INFERENCE CONCERNING THE DIFFERENCE BETWEEN TWO TREATMENTS IN CLINICAL TRIALS

Krishna K. Saha*, Central Connecticut State University

This article focuses on confidence interval construction for the difference between two treatment means in clinical trials and other similar fields. In this study, the interval methods based on the generalized estimating equation (GEE) approach and the ratio estimator approach are developed. The three other interval methods following the procedures studied for proportions are also developed. Monte Carlo simulations indicate that all the

procedures have reasonably well coverage properties. However, the GEE-based interval procedure outperforms other interval procedures in terms of all three confidence interval criteria. An example in clinical trials is also presented to illustrate the proposed confidence interval procedures.

email: **sahakrk@ccsu.edu**

87. CONTRIBUTED PAPERS: Computational Methods

DNase2TF: AN EFFICIENT ALGORITHM FOR FOOTPRINT DETECTION

Songjoon Baek*, National Cancer Institute, National Institutes of Health

Myong-Hee Sung, National Cancer Institute, National Institutes of Health

Gordon L. Hager, National Cancer Institute, National Institutes of Health

By deep sequencing of DNase-seq data and analyzing the nucleotide-resolution DNase cleavage profiles, it is possible to achieve digital footprinting of transcription factors. The DNA regions that are bound by proteins and relatively protected from enzymatic cutting are termed footprints. Decreasing costs and higher yields of improved sequencing methods make digital footprinting more feasible, making de novo discovery of relevant transcription factors possible. However, reliable and fast computational methods must be widely available to enable footprint detection from DNase-seq data. Here we present DNase2TF, a new detection algorithm that scans DNase I hypersensitive sites for putative footprints. When compared to previous methods, DNase2TF is faster and more accurate in predicting actual transcription factor binding sites. We also assess a limitation of using footprints for binding prediction that may be caused by insufficient sequencing and/or certain binding events that do not produce footprints. DNase2TF allows rapid identification of footprint candidates, but care should be taken when inferring transcription factor binding through footprints. The MATLAB source code and C code are provided as Supplementary Material and updated in http://sourceforge.net.

email: **baeks@mail.nih.gov**

SPECTRAL PROPERTIES OF MCMC ALGORITHMS FOR BAYESIAN LINEAR REGRESSION WITH GENERALIZED HYPERBOLIC ERRORS

Yeun Ji Jung*, University of Florida

James P. Hobert, University of Florida

We study MCMC algorithms for Bayesian analysis of a linear regression model with generalized hyperbolic errors. The Markov operators associated with the standard data augmentation algorithm and a sandwich variant of that algorithm are shown to be trace-class. This means the Markov chains underlying the DA and sandwich algorithms are geometrically ergodic; that is, the Markov chains converge to the target posterior distribution at a geometric rate. This result is highly important because geometric ergodicity of the Markov chain guarantees the existence of CLT (Central Limit Theorem) s which allow for the computation of valid asymptotic standard errors for MCMC-based estimates.

email: **snow0907@gmail.com**

GROUP FUSED MULTINOMIAL REGRESSION

Brad Price*, University of Miami

Charles J. Geyer, University of Minnesota

Adam J. Rothman, University of Minnesota

We propose a penalized likelihood method to reduce the number of response categories in multinomial logistic regression. An l2 rusion peanlty is used to introduce shrinkage and exploit vectorwise simialrity of the regression coefficieints. An ADMM algortihm is used for optimization, and tuning parameter selection is also adressed.

email: **bprice@bus.miami.edu**

ANALYSIS OF MCMC ALGORITHMS FOR BAYESIAN LINEAR REGRESSION WITH LAPLACE ERRORS

Hee Min Choi*, University of California, Davis

Let pi denote the intractable posterior density that results when the standard default prior is placed on the parameters in a linear regression model with iid Laplace errors. We analyze the Markov chains underlying two different Markov chain Monte Carlo algorithms for exploring pi. In particular, it is shown that the Markov operators associated with the data augmentation (DA) algorithm and a sandwich variant are both trace-class. Consequently, both Markov chains are geometrically ergodic. It is also established that for each i=1,2,3,..., the ith largest eigenvalue of the sandwich operator is less than or equal to the

corresponding eigenvalue of the DA operator. It follows that the sandwich algorithm converges at least as fast as the DA algorithm. (Joint work with Dr. Jim Hobert.)

email: **hmchoi@ucdavis.edu**

ON THE USE OF CAUCHY PRIOR DISTRIBUTIONS FOR BAYESIAN BINARY REGRESSION

Joyee Ghosh*, University of Iowa

Yingbo Li, Clemson University

Robin Mitra, University of Southampton

Cauchy prior distributions for regression parameters have been popular in the Bayesian linear regression literature for a long time. One of the main advantages of these prior distributions is that they are heavy tailed and hence much more robust compared to normal prior distributions. More recently a Cauchy prior distribution has been recommended as the default choice for logistic regression by Gelman et al. (2008). It is known that the mean does not exist for the Cauchy distribution and a natural question is if there are some scenarios under which the posterior mean of the regression parameters may not exist either. In this talk we focus on complete separation in logistic regression, a scenario that is not that uncommon with many binary covariates. We provide some theoretical justification to show that the posterior mean will not exist for a Cauchy prior distribution when there is complete separation, under certain conditions. We also use simulation studies and real data analysis to illustrate the effect of this property on the behavior of the Markov chain that is used to sample from the posterior distribution of the parameters. Based on our theoretical and simulation results, our recommendation is that a Cauchy prior needs to be used with care because the existence of posterior moments is not always guaranteed. As a result, for a full Bayesian analysis of a binary regression model using Markov chain Monte Carlo, Student t priors with somewhat larger degrees of freedom could serve as a safer choice.

email: **joyee123in@gmail.com**

FAST, EXACT BOOTSTRAP PRINCIPAL COMPONENT ANALYSIS FOR p > 1 MILLION

Aaron Fisher*, Johns Hopkins University

Brian Caffo, Johns Hopkins University

Brian Schwartz, Johns Hopkins University

Vadim Zipunnikov, Johns Hopkins University

Many have suggested a bootstrap procedure for estimating the sampling variability of principal component analysis (PCA) results. However, when the number of measurements per subject (p) is much larger than the number of subjects (n), calculating and storing the leading principal components from each bootstrap sample can be computationally infeasible. To address this, we outline methods for fast, exact calculation of bootstrap principal components, eigenvalues, and scores. Our methods leverage the fact that all bootstrap samples occupy the same n-dimensional subspace as the original sample. As a result, all bootstrap principal components are limited to the same n-dimensional subspace and can be efficiently represented by their low dimensional coordinates in that subspace. Several uncertainty metrics can be computed solely based on the bootstrap distribution of these low dimensional coordinates, without calculating or storing the p-dimensional bootstrap components. We apply fast bootstrap PCA to a dataset of brain magnetic resonance images (MRIs) (p= approx 3 million, n=352). Our method allows for standard errors for the first 3 principal components based on 1000 bootstrap samples to be calculated on a standard laptop in 47 minutes, as opposed to approximately 4 days with standard methods.

email: **fisher@jhu.edu**

88. Biostatistical Methods for Heterogeneous Genomic Data

INVESTIGATING TUMOR HETEROGENEITY TO IDENTIFY ETIOLOGICALLY DISTINCT SUB-TYPES

Colin B. Begg*, Memorial Sloan Kettering Cancer Center

Many investigators have conducted studies that examine the molecular profiles of tumors to identify sub-types with distinctive patterns based on gene expression, copy numbers changes, mutations, or other somatic or epigenetic events. Ideally the sub-types so identified display distinct clinical phenotypes. However, molecular profiles can also be examined with the goal of identifying etiologically distinct sub-types. In this talk a general strategy for establishing and optimizing such etiologic heterogeneity is presented.

A scalar measure of etiologic heterogeneity that can be used to characterize a set of sub-types is defined. It is shown how this can then be used to direct a clustering strategy to identify the sub-types that are most clearly etiologically distinct. The ideas are illustrated using data from ongoing studies of cancer epidemiology.

email: **beggc@mskcc.org**

STATISTICAL CHALLENGES IN CANCER RESEARCH: HETEROGENEITY IN FUNCTIONAL IMAGING AND MULTI-DIEMNSIONAL OMICS DATA

Kim-Anh Do*, University of Texas MD Anderson Cancer Center

Thierry Chekouo, University of Texas MD Anderson Cancer Center

Francesco Stingo, University of Texas MD Anderson Cancer Center

Brian Hobbs, University of Texas MD Anderson Cancer Center

Yuan Wang, University of Texas MD Anderson Cancer Center

Jianhua Hu, University of Texas MD Anderson Cancer Center

James Doecke, CSIRO, Australian e-Health Research Centre, Brisbane, Australia

Understanding different types of heterogeneity is one of the challenges that needs to be addressed to drive cancer research forward. I will describe, at a high level, the statistical questions posed by cancer research at MD Anderson that involve the different facets of heterogeneity. I will focus on two main research projects: (i) Simultaneous supervised classification ofmultivariate correlated objects collected from perfusion computed tomography. The aim is to distinguish between biologically distinct tissue types, metastatic versus normal liver, through the evaluation of vasculature heterogeneity; (ii) Differential networks between treatment groups, cancer subtypes, or prognostic features, using different modalities of genomic data (mRNA expression, copy number, microRNA), in association with heterogeneous survival times in glioblastoma patients from the Cancer Genome Atlas (TCGA) study. I will discuss the development of computer-intensive statistical models,simulation studies conducted, and inferential results. This is joint work with Francesco Stingo, Thierry Chekouo, James Doecke, Yuan Wang, Brian Hobbs, Jianhua Hu.

email: **kimdo@mdanderson.org**

ACCOUNTING FOR CELLULAR HETEROGENEITY IS CRITICAL IN EPIGENOME-WIDE ASSOCIATION STUDIES

Rafael Irizzary*, Harvard University

Epigenome-wide association studies (EWAS) of human disease and other quantitative traits are becoming increasingly common. A series of papers reporting age-related changes in DNA methylation (DNAm) profiles in peripheral blood have already been published. However, blood is a heterogeneous collection of different cell types, each with a very different DNA methylation profile. Using a statistical method that permits estimating the relative proportion of cell types from DNAm profiles, we examine data from five previously published studies, and find strong evidence of cell composition change across age. We also demonstrate that, in these studies, cellular composition explains much of the observed variability in DNAm. Furthermore, we find high levels of confounding between age-related variability and cellular composition at the CpG level. We also present data from brain samples were the cell composition issue also arises.

email: **rafa@jimmy.harvard.edu**

MODELLING SOURCES OF VARIABILITY IN SINGLE-CELL TRANSCRIPTOMICS DATA

Sylvia Richardson*, MRC Biostatistics Unit Cambridge, UK

Catalina Vallejos, MRC Biostatistics Unit Cambridge and European Bioinformatics Institute, Hinxton, UK

John Marioni, European Bioinformatics Institute, Hinxton, UK

Current technology has made possible the analysis of gene expression levels with high resolution. Instead of measuring overall expression across groups of cells, scientists are now able to report measures at a single-cell level, with typical data represented by a matrix which entries correspond to the observed expression counts for each gene across cells. It is known that high levels of technical noise are usually observed when dealing with small amount of genetic material. This creates new challenges for identifying genes, which show genuine within-tissue heterogeneity beyond that induced by technical noise. An additional challenge in this context is the normalization of the expression counts. This is due to cells having different amounts genetic material and technical aspects such as sequencing depth. In this talk, statisti-

cal approaches to model the different sources of variability of such data will be presented. As opposed to previous literature, we implement a self-normalization procedure, where normalizing constants are treated as unknown model parameters.

e-mail:
sylvia.richardson@mrc-bsu.cam.ac.uk

89. Innovative Approaches in Competing Risk Analysis

FLEXIBLE MODELING OF COMPETING RISKS AND CURE RATE

Qi Jiang, Northern Illinois University

S**anjib Basu***, Northern Illinois University

The cumulative incidence functions based approach to competing risks modeling has the advantage of providing direct inference on the survival probabilities from each risk. A unified competing risks cure rate model is proposed in this work where the cumulative incidence functions of the competing risks are directly modeled. The proposed model further accounts for the possibility of cure from one or more of the competing risks. Bayesian analyses of these models are explored, and conceptual, methodological and computational issues related to Bayesian model fitting and model selection are discussed. The performance of the proposed model is investigated in simulation studies. The unified model is used to analyze cancer survival data from SEER and a clinical study.

e-mail: **basu@niu.edu**

COMPETING RISKS PREDICTION IN TWO TIME SCALES

Jason Fine*, University of North Carolina, Chapel Hill

In the standard analysis of competing risks data, proportional hazards models are fit to the cause-specific hazard functions for all causes on the same time scale. These regression analyses are the foundation for predictions of cause-specific cumulative incidence functions based on combining the estimated cause-specific hazard functions. However, in predictions arising from disease registries, where only subjects with disease enter the database, disease related mortality may be more naturally modelled on the time since diagnosis time scale while death from other causes may be more naturally modelled on the age time scale. The single time scale methodology may be biased if an incorrect time scale is employed for one of the causes and alternative methodology is not available. We propose inferences for the cumulative incidence function in which regression models for the cause-specific hazard functions may be specified on different time scales. Using the disease registry data, the analysis of other cause mortality on the age scale requires left truncating the event time at the age of disease diagnosis, complicating the analysis. In addition, standard martingale theory is not applicable when combining regression models on different time scales. We establish that the covariate conditional predictions are consistent and asymptotically normal using empirical process techniques and propose consistent variance estimators which may be used to construct confidence intervals. Simulation studies show that the proposed two time scale methods perform well, outperforming the single time scale predictions when the time scale is misspecified. The methods are illustrated with stage III colon cancer data obtained from the Surveillance, Epidemiology, and End Results (SEER) program of National Cancer Institute.

e-mail: **jfine@email.unc.edu**

CHECKING FINE AND GRAY'S SUBDISTRIBUTION HAZARDS MODEL WITH CUMULATIVE SUMS OF RESIDUALS

Jianing Li, Medical College of Wisconsin

Thomas H. Scheike, University of Copenhagen

Mei-Jie Zhang*, Medical College of Wisconsin

Recently, Fine and Gray (1999) proposed a semi-parametric proportional regression model for the subdistribution hazard function which has been used extensively for analyzing competing risks data. However, failure of model adequacy could lead to severe bias in parameter estimation, and only a limited contribution has been made to check the model assumptions. In this talk, we present a class of graphical and analytical methods for checking the assumptions of Fine and Gray's model. The proposed goodness-of-fit test procedures are based on the cumulative sums of residuals. We validate the model in three aspects: (1) proportionality of hazard ratio, (2) the linear functional form and (3) the link function. For each proposed test, we provided a visualized plot and a testing p-value

against the null hypothesis using a simulation-based approach. We also consider an omnibus test for overall evaluation against any model misspecification. The proposed test methods performed well in simulation studies and are illustrated with real data example.

e-mail: **meijie@mcw.edu**

90. Biomarker Evaluation in Diagnostics Studies with Longitudinal Data

COMBINATION OF LONGITUDINAL BIOMARKERS WITH MISSING DATA

Danping Liu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

A common practice in disease prediction is to measure a biomarker repeatedly over time, where the trajectory information of the biomarker greatly improves the prediction accuracy. Missingness occurs in most longitudinal studies, resulting in incomplete observations of both biomarkers and the disease condition. Liu and Albert (2014) proposed a pattern mixture model (PMM) to combine the longitudinal biomarkers, but their approach cannot handle missing data. We extend their results by allowing both the longitudinal biomarkers and the disease outcome to be missing at random. We develop a doubly robust procedure with PMM for prediction. Under the PMM framework, the missingness in the biomarkers can be handled by using a likelihood-based inference. To account for the missingness in the disease status, we assume a disease model given covariates and a missing mechanism model. The likelihood function is then reweighted by the missing probability and the disease probability. The double robust estimation allows either the disease model or the missing mechanism model to be incorrect. We examine the performance of double robust PMM approach in extensive simulation studies. The proposed methods are motivated from and applied to a fetal growth study, in which neonatal outcomes are predicted by the longitudinal ultrasound measurements.

e-mail: **danping.liu@nih.gov**

MEASURES TO EVALUATE BIOMARKERS AS PREDICTORS OF INCIDENT CASES

Chao-Kang Jason Liang*, University of Washington

Patrick J. Heagerty, University of Washington

In many biomedical applications a primary goal is to predict incident or future cases, and appropriate measures that characterize a biomarker's predictive potential or incremental value are needed. We first review existing non-parametric methods proposed for incident time-dependent accuracy (Zheng and Heagerty, 2005; Saha and Heagerty, 2013) and then overview extensions of integrated discrimination index (IDI) that are appropriate for hazard models. The proposed new methods are also connected to information theory based criteria for model choice. We outline estimation methods and applications to benchmark data sets to illustrate the methodology.

e-mail: **liangcj@uw.edu**

PREDICTION ACCURACY OF LONGITUDINAL MARKER MEASUREMENT

Paramita Saha Chaudhuri*, McGill University

Patrick Heagerty, University of Washington

Longitudinal marker measurements serves several purposes, from disease screening to treatment management. With the progress of medicine in the last couple of decades, there is now considerable focus on early detection of disease via population screening and longitudinal measurements, giving the patient more treatment options if the disease is found early and consequently a better prognostic outlook. However, dispute still remains as to whether multiple marker measurement is predictive over and above baseline or less frequent measurements. Sometimes, it is not clear if a relaxed marker measurement frequency can be used without compromising the benefits (e.g., annual versus biennial). In this talk, I will introduce a summary from time-dependent ROC that can be used to compare the accuracy of longitudinal marker measurements. I will demonstrate the approach with simulated and real data.

e-mail: **paramita.sahachaudhuri.work@gmail.com**

ESTIMATING TIME-DEPENDENT ACCURACY MEASURES FOR SURVIVAL OUTCOME UNDER TWO-PHASE SAMPLING DESIGNS

Dandan Liu*, Vanderbilt University

Tianxi Cai, Harvard University

Anna Lok, University of Michigan

Yingye Zheng, Fred Hutchinson Cancer Research Center

Large prospective cohort studies of rare chronic diseases such as cancer often require thoughtful planning in study designs, especially for biomarker study when measurement are based on stored tissue or blood specimens. Two phase designs, including nested case control (Thomas, 1977) and case-cohort (Prentice, 1986) sampling designs, provide cost-effective tool in the context of biomarker evaluation, especially when the clinical condition of interest is rare. Existing literature for biomarker assessment under two phase designs has been based on simple inverse probability weighting (IPW) estimators (Cai and Zheng, 2011; Liu et al., 2012). Drawing on recent theoretical development on the maximum likelihood estimators for two-phase studies (Scheike and Martinussen, 2004; Zeng et al., 2006), we propose statistical methods to evaluate accuracy and predictiveness of a risk prediction biomarker, with censored time-to-event outcome under both types of two-phase designs. Hybrid estimators that combine IPW estimators and MLE procedures are proposed to improve efficiency and alleviate computational burden. We derive large sample properties of proposed estimators and evaluate their finite sample performance using numerical studies. We illustrate new procedures using a two-phase biomarker study aiming to evaluate the accuracy of a novel biomarker, des-γ-carboxy prothrombin, for the early detection of hepatocellular carcinoma (Lok et al., 2010).

e-mail: **Dandan.Liu@Vanderbilt.edu**

COMPRESSION OF LONGITUDINAL GENOMIC BIOMARKERS FOR DIAGNOSIS STUDY

Le Bao*, The Pennsylvania State University

Xiaoyue Niu, The Pennsylvania State University

Kayee Yeung, University of Washington

In the genome-wise association study (GWAS), one of the objectives is to establish the associations between genes and diseases. Besides traditional gene expression data that are collected in the form of a two dimensional gene by individual matrix, with new technology, individual gene expression levels are also followed longitudinally for a series of time points. The resulting data form a three dimensional gene by individual by time array. In order to infer associations from these newly available data, we propose a novel two-step approach that uses model-based clustering and contingency tables to analyze the array data. The proposed method is computational efficient and suites the analysis goal.

e-mail: **lebao@psu.edu**

91. Solving Clinical Trial Problems by Using Novel Designs

SOME DESIGN APPROACHES TO ADDRESS MISSING DATA DUE TO EARLY DISCONTINUATION IN CLINICAL TRIALS

Sonia M. Davis*, University of North Carolina, Chapel Hill

Clinical trials of some indications such as neuroscience often have markedly high rates of early patient discontinuation. Statistical methods to compare treatment groups in the face of missing data have become more developed in the past decade. However, high rates of missing data due to patient discontinuation pose substantial complications for addressing bias and interpretation of efficacy and safety results. In these settings, studies designed to reduce the likelihood of participant drop-out are warranted. This talk identifies and discusses clinical trial design elements aimed at (1) minimizing the rate of early discontinuation, (2) minimizing the impact of discontinuation on the assessment of the outcome measure, and (3) maximizing the ability to gather patient status information after treatment discontinuation. Topics include recommendations by the National Academy of Science panel on missing data, a discussion of 3-arm 2-period cross-over designs, and use of time to treatment discontinuation as an outcome measure.

e-mail: **sonia.davis@unc.edu**

INTRODUCTION TO THE SEQUENTIAL ENRICHED DESIGN

Yeh-Fong Chen*, U.S. Food and Drug Administration

Roy Tamura, University of South Florida

For many disease areas, the placebo response can be high in clinical trials. Enrichment designs are thought of as being able to provide a way to address this issue. Some of the interesting enrichment designs such as sequential parallel design (Fava et al., 2003), two way enriched design (Ivanova and Tamura, 2011) and sequential enriched design (SED) (Chen et al., 2014) were proposed recently in the literature. The SED was devised not only to reduce placebo response but also to enhance the capability of detecting a targeted treatment effect. As it is a new design, the SED's implementation in real clinical trial settings and its advantage over other enrichment designs need careful evaluation. In this presentation, I will describe the SED and evaluate the selection of required design parameters in terms of power optimization and sample size planning. I will also discuss the issues of missing data and considerations of interim analysis implementation related to its applicability.

e-mail: **yehfong.chen@fda.hhs.gov**

INTEGRITY AND EFFICIENCY OF ENRICHMENT AND ADAPTIVE TRIAL DESIGN AND ANALYSIS OPTIONS TO ENABLE ACCURATE AND PRECISE SIGNAL DETECTION

Marc L. de Somer*, PPD

Traditional clinical trial design and analysis options require large sample sizes to detect the true signal when noise is high due to excessive placebo response and variability. Increasing sample size is self-defeating because it can increase in variability, due to the multiplication of investigational sites, countries and subjects. Several partial enrichment trial design solutions have been recently proposed: the sequential parallel comparison design (SPCD: Fava M, 2003), the two-way enriched design (TED: Ivanova A, 2012), and the sequential enriched design (SED: Chen YF, 2014). Their aim is to filter excessive placebo response and variability, and identify a target responder subject set in which signal detection is enhanced. The present research evaluates their performance in terms of type I error control, accuracy, precision and power, using multiple analysis models and missing data handling methods (ANCOVA.LOCF and MMRM). The reliability of inference based on a linear combination of stage-wise statistics is evaluated. Finally, the three enrichment design options are compared in terms of trial integrity, efficiency, feasibility and economics. Simulation across a wide range of assumptions reveals the optimal choice of the trial design, analysis and missing data handling method in each specific context.

e-mail: **marcldesomer@gmail.com**

92. Ensuring Biostatistical Competence Using Novel Methods

WHAT DO NON-BIOSTATISTICS CONCENTRATORS NEED FROM THE INTRODUCTORY BIOSTATISTICS COURSE?

Jacqueline N. Milton*, Boston University

Successful careers in public health require practitioners to have an understanding of biostatistical concepts, applications and techniques. Developing a curriculum that ensures that all students understand and can apply biostatistical techniques requires an understanding of which concepts in biostatistics are most critical and how they will be applied in various disciplines. Here we sought to understand the learning goals of graduate students in public health in the introductory biostatistics course. We surveyed professors and alumni to determine which biostatistics concepts and applications were most critical. We used this information to inform and further develop our curriculum so that it could be tailored more towards students' interests and career needs.

e-mail: **jnmilton@bu.edu**

CREATING THE INTEGRATED BIOSTATISTICS-EPIDEMIOLOGY CORE COURSE: CHALLENGES AND OPPORTUNITIES

Melissa D. Begg*, Columbia University

Roger D. Vaughan, Columbia University

Dana March, Columbia University

Biostatistics and Epidemiology are considered the "basic sciences" of public health, yet many students fear taking these courses due to their technical nature. One of the challenges for educators is to persuade students of the power and utility of quantitative methods, and deliver the skills required to interpret the literature and identify effective public health interventions. One approach to this problem is to integrate the teaching of these 2 disciplines into one core course, with the goal of helping students to see more immediately how these skill sets can be applied together to better understand the factors that hinder or promote health. However, one must take care when delivering an integrated course to ensure that students have adequate exposure to both disciplines, and that both are capably taught. We share some insights and recommendations after 3 offerings of a combined core course as part of the MPH curriculum.

e-mail: **mdb3@columbia.edu**

MEETING PUBLIC HEALTH CAREER GOALS: COURSE OPTIONS IN BIOSTATISTICS AND EPIDEMIOLOGY

Marie Diener-West*, Johns Hopkins Bloomberg School of Public Health

The heterogeneity of previous backgrounds and experiences, as well as future professional goals, of public health graduate students motivated our development of several different introductory biostatistics course sequence options. The basic understanding of statistical principles and applications is paramount in each of these sequences but they vary by level of achieved skills in performing data analysis as well as understanding of statistical theory. In addition, our epidemiology course options beyond fundamental concepts have evolved to focus on either epidemiologic research methods or professional methods for public health practice. We have recognized the ability to coordinate between the varying biostatistics and epidemiology course sequences to offer students cohesive curricular options that match their diverse needs.

e-mail: **mdiener@jhsph.edu**

93. Methodological Frontiers in the Analysis of Panel Observed Data

SECOND-ORDER MODELS OF WITHIN-FAMILY ASSOCIATION IN CENSORED DISEASE ONSET TIMES

Yujie Zhong*, University of Waterloo

Richard John Cook, University of Waterloo

In preliminary studies of the genetic basis for chronic conditions, interest routinely lies in examining the within-family dependence in disease status. When probands are selected from disease registries and their respective families are recruited, a variety of methods are available which correct for the selection bias, which are typically based on models for correlated binary data. This approach ignores the age of family members at the time of assessment. We consider likelihood and composite likelihood based methods for modeling within-family dependence in the disease onset times using copula models for settings in which data from non-probands are subject to right censoring and current status observation schemes. The advantages of pairwise and partial pairwise composite likelihoods are discussed in terms of computation speed and statistical efficiency. These models and methods are also used to examine the factors influencing the commonly used measures of within-family dependence based on binary responses.

e-mail: **zyujie@uwaterloo.ca**

MODELING COGNITIVE STATES IN THE ELDERLY: THE ANALYSIS OF PANEL DATA USING MULTI-STATE MARKOV AND SEMI-MARKOV PROCESSES

Richard J. Kryscio*, University of Kentucky

Continuous-time multi-state models are commonly used to describe the movement of elderly subjects among various cognitive states in dementia studies. The cognition of each subjects is periodically assessed leading to interval-censoring for the cognitive states: intact cognition, Mild Cognitive Impairment (MCI), and Dementia. In these studies death without dementia is a competing risk which is not interval censored. We discuss two approaches to modeling this type of panel data: Markov chains and semi-Markov processes in terms of computational issues that arise when some cognitive

states such as MCI involve back transitions and when parametric models are used to model time spent in states. Data from the Statistical Modeling of Risk Transitions project, a consortium of six longitudinal studies of cognition in the elderly, will be used to illustrate the results.

e-mail: **kryscio@email.uky.edu**

MULTI-STATE MODELS: A VARIETY OF USES

Vern Farewell*, MRC Biostatistics Unit, Cambridge, UK

Recent uses of multi-state models in the analysis of longitudinal data reflect their usefulness in the specification of data structures and their flexibility. The use of multi-state models for a variety of problems will be illustrated to demonstrate these characteristics. These problems will involve the challenges of panel data, time to event analyses for events defined only by prolonged observation and correlated processes. The application of causal reasoning in the context of multi-state models will also be briefly discussed.

e-mail:
vern.farewell@mrc-bsu.cam.ac.uk

COMPUTATIONALLY SIMPLE STATE OCCUPANCY PROBABILITY ESTIMATES FOR MULTI-STATE MODELS UNDER PANEL OBSERVATION

Andrew Titman*, Lancaster University

A desirable way of assessing the appropriateness of a parametric multi-state model is to compare the model-based estimates of state occupancy probabilities with some non-parametric estimate. Existing non-parametric estimates of state occupancy either make crude assumptions, leading to bias, or else are more computationally intensive to implement than the original parametric model. A computationally simple method for obtaining non-parametric estimates of the state occupation probabilities is proposed for progressive multi-state models where transition times between intermediate states are subject to interval censoring. The method separates estimation of overall survival, using standard methods for survival data, and estimation of the conditional cumulative incidence of progression to a series of subsets of the state space performed using methods for current status competing risks data. The resulting estimates of state occupancy are unbiased, without requiring a Markov assumption, when the disease process and examination times are independent. An inverse visit-intensity weighted estimator is proposed for cases where the time to next examination depends on the last observed state. The method can also be extended to provide approximate estimates of the marginal transition probabilities.

e-mail: **a.titman@lancaster.ac.uk**

94. CONTRIBUTED PAPERS: Ordinal and Categorical Data

EXPLICIT ESTIMATES FOR CELL COUNTS AND MODELING THE MISSING DATA INDICATORS IN THREE-WAY CONTINGENCY TABLE BY LOG-LINEAR MODELS

Haresh D. Rochani*, Georgia Southern University

Robert L. Vogel, Georgia Southern University

Hani M. Samawi, Georgia Southern University

Daniel F. Linder, Georgia Southern University

Missing observations in cross-classified data are an extremely common problem in the process of research in public health, clinical sciences and social sciences. Ignorance of missing values in the analysis can produce biased results and low statistical power. The purpose of this research was to expand Baker, Rosenberger and Dersimonian (BRD) model approach to compute the explicit maximum likelihood estimates for cell counts for three-way cross-classified data. Derivation of explicit cell counts for three-way table with supplementary margins can be obtained by controlling the missingness in third variable and by modeling the missing-data indicators using homogeneous log-linear models. Previous methods for contingency tables with supplementary margins required an iterative algorithms, however, expected cell counts can be obtained by simple algebraic formula. Simulation study with source of knowledge of cancer data

illustrate that how well the explicit maximum likelihood estimates can produce consistent results in idyllic circumstances. Application of the BRD model approach to Slovenian public opinion survey data reveals the effect of smaller sample size to the validity of the method.

e-mail: **hrochani@georgiasouthern.edu**

ADDITIVE INTERACTIONS AND THE METABOLIC SYNDROME

Matthew J. Gurka*, West Virginia University

Baqiyyah N. Conway, West Virginia University

Michael E. Andrew, National Institute for Occupational Safety and Health (NIOSH)

Cecil M. Burchfiel, National Institute for Occupational Safety and Health (NIOSH)

Mark D. DeBoer, University of Virginia

The metabolic syndrome (MetS) is generally defined as a cluster of cardiovascular risk factors, including obesity, high blood pressure, elevated triglycerides, low HDL, and elevated fasting glucose, that has been observed to be associated with future disease (diabetes, cardiovascular disease). MetS has been argued to be a stronger risk factor for future disease than the individual components that comprise it, but this assertion is hotly debated amongst clinicians and researchers alike. In addition, questions remain regarding whether such additive interactions are similar across racial/ethnic groups. Utilizing data from large cardiovascular cohort studies (Jackson Heart Study, Atherosclerosis Risk in Communities Study), we systematically apply epidemiologic principles in studying additive interactions in determining whether or not synergy amongst the traditional MetS components exists in predicting future disease. Specifically, we calculated the relative excess risk due to interaction (RERI) in estimating the additive interactions across the five MetS components. This study of RERI measures across five components, and the resulting comparisons among racial/ethnic groups, provided interesting methodologic challenges that helped us ultimately address the question of whether "the sum is truly greater than its parts."

e-mail: **mgurka@hsc.wvu.edu**

FLEXIBLE LINK FUNCTIONS IN NON-PARAMETRIC BINARY REGRESSION WITH GAUSSIAN PROCESS PRIORS

Dan Li*, University of Cincinnati

Xia Wang, University of Cincinnati

Lizhen Lin, University of Texas, Austin

Dipak K. Dey, University of Connecticut

In many scientific fields, a sequence of 0-1 measurements is frequently collected from a subject across time, space, or a collection of covariates. Researchers are interested in finding out how the expected binary outcome is related to covariates, and aim at better prediction in the future 0-1 outcomes. Gaussian processes have been used to model the latent structure in a binary regression model, but little is known about the adequacy on the choice of link functions and its resulting effects on model fitting, predictive power and flexibility. Commonly used link functions such as probit and logit links have fixed skewness and lack the flexibility to allow the data to determine the degree of skewness. To address this limitation, we propose a Bayesian nonparametric model which combines a generalized extreme value link function with a Gaussian process prior on the latent structure for flexible modeling. The efficiency and gains of our proposed model are illustrated through the analysis of two real data examples with one collected in an experimental paradigm that studies the monkey attention and the other studying the course of development of pneumoconiosis among coal miners when exposed to certain mining conditions.

e-mail: **lid7@mail.uc.edu**

PENALIZED NON-LINEAR PRINCIPAL COMPONENTS ANALYSIS FOR ORDINAL VARIABLES

Jan Gertheiss*, Georg August University, Germany

Nonlinear principal components analysis (PCA) for categorical data constructs new variables by assigning numerical values to categories such that the proportion of variance in those new variables that is explained by a predefined number of principal components is maximized. We propose a penalized version of nonlinear PCA for ordinal variables that is an intermediate between standard PCA on category labels and nonlinear PCA as used so far. Our approach offers both better interpretability of the nonlinear transformation of the category labels as well as better performance on validation data than unpenalized nonlinear PCA. The new method is applied to the International Classification of Functioning, Disability and Health (ICF).

e-mail: **jgerthe@gwdg.de**

COVARIANCE ESTIMATION OF PROPORTION FOR MISSING DICHOTOMOUS AND ORDINAL DATA IN RANDOMIZED LONGITUDINAL CLINICAL TRIAL

Siying Li*, University of North Carolina, Chapel Hill

Gary Koch, University of North Carolina, Chapel Hill

This paper presents a closed form method for sensitivity analysis of a randomized multi-visit multi-center clinical trial that possibly has missing not at random (MNAR) dichotomous data and an extension to ordinal data. Counts of missing data are redistributed to the each category of the outcome probabilistically to adjust for possibly informative missing; adjusted proportion estimates as well as their closed form covariance estimates are provided. Treatment comparisons over time are addressed with adjustment for a stratification factor and/or baseline covariates. The parameter estimates are computed via weighted least square asymptotic regression through randomization based methods. Application of such sensitivity analyses are illustrated with an example.

e-mail: **siying@live.unc.edu**

BAYESIAN NONPARAMETRIC MULTIVARIATE ORDINAL REGRESSION

Junshu Bao*, University of South Carolina

Timothy E. Hanson, University of South Carolina

Multivariate ordinal data are modeled as a stick-breaking mixture of multivariate probit models. Parametric multivariate probit models are first developed for ordinal data, then generalized to finite mixtures of multivariate probit models. Specific recommendations for prior settings are carefully reasoned and found to work well in simulations and data analyses. Interpretation of the model is carried out by examining aspects of the mixture components as well as through averaged effects focusing on the mean responses. A simulation verifies that the nonparametric model is capable to model bivariate ordinal data with latent variables generated from different distributions. In all simulations, nonparametric models perform better than the parametric models in terms of LPML (log pseduo marginal likelihood) and MED (maximal expected discrepancy in probability). An analysis of alcohol drinking behavior data illustrates the usefulness of the proposed model.

e-mail: **bao3@email.sc.edu**

95. CONTRIBUTED PAPERS: Statistical Genetics

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 (DKFZ)

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 the 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 single nucleotide polymorphisms (SNPs) discovered through recent genome-wide association studies. The analysis clearly indicates a 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.

e-mail: **songm4@mail.nih.gov**

PLEMT: A NOVEL PSEUDOLIKELIHOOD BASED EM TEST FOR HOMOGENEITY IN GENERALIZED EXPONENTIAL TILT MIXTURE MODELS

Chuan Hong*, University of Texas School of Public Health, Houston

Yong Chen, University of Texas School of Public Health, Houston

Yang Ning, Princeton University

Shuang Wang, Columbia University

Hao Wu, Emory University

Raymond J. Carroll, Texas A&M University

Motivated by analyses of DNA methylation data, we propose a semiparametric mixture model, namely the generalized exponential tilt mixture model, to account for heterogeneity between differentially methylated and non-differentially methylated subjects in the cancer group, and capture the differences in higher order moments (e.g. mean and variance) between subjects in the cancer and normal groups. A pairwise pseudolikelihood is constructed to eliminate the unknown nuisance function. To circumvent boundary and non-identifiability problems as in parametric mixture models, we modify the pseudolikelihood by adding a penalty function. In addition, as epigenetic and genetic data are usually high dimensional, computationally efficient tests have great advantages over permutation based tests. To this end, we propose a pseudolikelihood based expectation--maximization test, and show the proposed test follows a simple chi-squared limiting distribution. Simulation studies show that the proposed test performs well in controlling Type I errors and has superior power compared to the current tests. In particular, the proposed test outperforms the commonly used tests under all simulation settings considered, especially when there is variance difference between two groups. The proposed test is illustrated by an example of identifying differentially methylated sites between ovarian cancer subjects and normal subjects.

e-mail: **chuan.hong@uth.tmc.edu**

REGRESSION-BASED METHODS TO MAP QUANTITATIVE TRAIT LOCI UNDERLYING FUNCTION-VALUED PHENOTYPES

Il Youp Kwak*, University of Minnesota

Karl W. Broman, University of Wisconsin, Madison

Genetic loci that contribute to variation in a quantitative trait are called quantitative trait loci (QTL). We are developing simple regression-based methods to map QTL that influence a function-valued outcome (such as body weight measured over time) in an experimental cross. In order to handle noisy trait measurements and to account for the correlation structure among time points, we apply an initial smoothing followed by functional principal component analysis. Functional PCA reduces the functional data a small number of principal components without much loss of information. We then consider multiple methods for QTL analysis with these dimension-reduced traits, including a multi-trait mapping method proposed by Knott and Haley (2000), and simple combinations of the single-trait analysis results. All of these methods have been implemented in an R package, funqtl.

e-mail: **ikwak@umn.edu**

A FRAMEWORK FOR CLASSIFYING RELATIONSHIPS USING DENSE SNP DATA AND PUTATIVE PEDIGREE INFORMATION

Zhen Zeng*, University of Pittsburgh

Daniel E. Weeks, University of Pittsburgh

Wei Chen, Children's Hospital of Pittsburgh of UPMC

Nandita Mukhopadhyay, University of Pittsburgh

Eleanor Feingold, University of Pittsburgh

When genome-wide association studies (GWAS) or sequencing studies are performed on family-based datasets, the genetic marker data can be used to check the structure of putative pedigrees. Even in datasets of putatively unrelated people, close relationships can often be detected using dense single-nucleotide polymorphism (SNP) data. A number of methods for finding relationships using genetic data exist, but they all have certain limitations, such as being intended for uncorrelated genetic markers or for correctly-phased genotype data. Also, a common limitation of existing methods is that they use average genetic sharing, which is only a subset of the available information. In this paper we present a set of approaches for classifying relationships in GWAS datasets or large-scale sequencing datasets. We first propose an empirical method for detecting identity-by-descent segments in close relative pairs using dense SNP data, and then demonstrate how that information can be used to build a relationship classifier. We then develop a strategy to take advantage of putative pedigree information to

enhance classification accuracy. Finally, we propose classification pipelines for checking and identifying relationships in datasets containing a large number of small pedigrees.

e-mail: **zhenhouse@msn.com**

A NEGATIVE BINOMIAL MODEL-BASED METHOD FOR DIFFERENTIAL EXPRESSION ANALYSIS BASED ON NANOSTRING nCOUNTER DATA

Hong Wang*, University of Kentucky

Arnold Stromberg, University of Kentucky

Chi Wang, University of Kentucky

The NanoString nCounter system is a new and promising technology that enables the digital quantification of multiplexed target RNA molecules. In this talk, we present a novel bioinformatics method to identify differential expression between two different groups based on NanoString nCounter data. Our negative binomial model-based method is specifically designed for this type of count data, which fully utilizes positive control, negative control and housekeeping probes for data normalization. We propose an empirical Bayes shrinkage approach to estimate the dispersion parameter and a likelihood ratio test to identify differential expression. Our simulation results show competitive performance of our method versus existing methods.

e-mail: **hong.wang@uky.edu**

TWO-STAGE BAYESIAN REGIONAL FINE MAPPING OF A QUANTITATIVE TRAIT

Shelley B. Bull*, University of Toronto and Lunenfeld-Tanenbaum Research Institute

Zhijian Chen, Lunenfeld-Tanenbaum Research Institute

Radu V. Craiu, University of Toronto

In focused studies designed to follow up associations detected in a genome-wide association study (GWAS), investigators can proceed to fine-map a genomic region by targeted sequencing or dense genotyping of all variants in the region, aiming to identify a functional sequence variant. For analysis of a quantitative trait, we consider a Bayesian approach to fine-mapping study design that incorporates stratification according to a promising GWAS tag SNP in the same region. Improved cost-efficiency can be achieved when the fine-mapping phase incorporates a two-stage design, with identification of a smaller set of more promising variants in a subsample taken in stage 1, followed by their evaluation in an independent stage 2 subsample. To avoid the potential negative impact of genetic model misspecification on inference we incorporate genetic model selection based on posterior probabilities for each competing model. Our simulation study shows that, compared to simple random sampling which ignores genetic information from GWAS, tag-SNP-based stratified sample allocation methods reduce the number of variants continuing to stage 2, and are more likely to promote the functional sequence variant into confirmation studies. This approach also permits additional stratification according to the quantitative trait value.

e-mail: **bull@lunenfeld.ca**

OPTIMAL RANKING PROCEDURES IN LARGE-SCALE INFERENCE: THRESHOLDING FAMILIES AND THE R-VALUE

Nicholas C. 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 (ML) approaches tend to favor units with high measurement error. Available Bayesian and empirical Bayesian approaches rely on specialized loss functions that result in similar deficiencies. We describe and evaluate a novel 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. The procedure relates collections of unit-specific posterior upper tail probabilities with their empirical distribution to yield a ranking variable. It discounts high-variance units less than popular non-ML methods and thus achieves improved operating characteristics in the models considered.

e-mail: **nhenders@stat.wisc.edu**

96. CONTRIBUTED PAPERS: Ecology and Forestry Applications

A STATISTICAL FRAMEWORK FOR THE GENETIC DISSECTION OF EVOLUTION INDUCED BY ECOLOGICAL INTERACTIONS

Cong Xu*, The Pennsylvania State University

Libo Jiang, Beijing Forestry University

Meixia Ye, Beijing Forestry University

Rongling Wu, The Pennsylvania State University

An increasing body of research has suggested that genes play a pivotal role in determining the spatial and temporal changes of ecological interactions in response to environmental perturbations. Here, we develop novel theory that can map genes, known as quantitative trait loci (QTLs), and their epistasis that affect intra- or interspecifc interactions involved in an ecological process. We derive a statistical framework to synthesize genetic mapping, an approach widely used in the field of quantitative genetics, and ecological experiments of species competition through mathematical equations. We quantify ecological competition between species using a system of ordinary differential equations (ODE) and further determine the genetic architecture of species interactions via estimating and testing QTL genotype-dependent differences in ODE parameters that specify a web of ecological interactions singly or jointly. The new framework is particularly equipped to characterize how genomes from different species interact with each other to govern species interactions and their ecological dynamics. The kinetic integration of QTL mapping and ecological experiments, which could not be made by previous theory, provides an innovative incentive to understand the intrinsic complexity of ecosystems and their evolutionary mechanisms.

e-mail: **congxu@hmc.psu.edu**

ANALYSIS OF VARIANCE OF INTEGRO-DIFFERENTIAL EQUATIONS WITH APPLICATION TO POPULATION DYNAMICS OF COTTON APHIDS

Xueying Wang, Washington State University

Jiguo Cao*, Simon Fraser University

Jianhua Huang, Texas A&M University

The population dynamics of cotton aphids are usually described by mechanistic models, in the form of integro-differential equations (IDEs), with the IDE parameters representing some key properties of the dynamics. Investigation of treatment effects on the population dynamics of cotton aphids is a central issue in developing successful chemical and biological controls for cotton aphids. Motivated by this important agricultural problem, we propose a framework of analysis of variance (ANOVA) of IDEs. The main challenge in estimating the IDE-based ANOVA model is that IDEs usually have no analytic solution, and repeatedly solving IDEs numerically leads to a high computational cost. We propose a penalized spline method in which spline functions are used to estimate the IDE solutions and the penalty function is defined by the IDEs. The estimated IDE solutions, as implicit functions of the parameters, are inputs in a nonlinear least squares criterion, which in turn is minimized by a Gauss-Newton algorithm. The proposed method is illustrated using simulation and an observed cotton aphids data set.

e-mail: **jiguo_cao@sfu.ca**

NEW INSIGHTS INTO THE USEFULNESS OF ROBUST SINGULAR VALUE DECOMPOSITION IN STATISTICAL GENETICS: ROBUST AMMI AND GGE MODELS

Paulo Canas Rodrigues*, Federal University of Bahia, Brazil

Andreia Monteiro, Nova University of Lisbon, Portugal

Vanda M. Lourenço, Nova University of Lisbon, Portugal

Two of the most widely used models to analyse genotype-by-environment data are the genotype main effects and genotype-by-environment interaction (GGE) model and the additive main effects and multiplicative interaction (AMMI) model. The GGE and AMMI models apply singular value decomposition (SVD) to the residuals of a specific linear model, to decompose the genotype-by-environment interaction (GEI) into a sum of multiplicative terms. However, SVD is highly sensitive to contamination and the presence of outliers may result in misinterpretations and, in turn, lead to bad practical decisions. Since, as in many other real life studies, the distribution of these data is usually not normal due to the presence of outlying observations, robust SVD methods have been suggested to help overcome this handicap. Therefore, a new approach, where robust

statistical methods replace the classic ones to model and analyse GEI in the context of multi-location plant breeding trials, is presented. The performance of the proposed robust extensions of the AMMI and GGE models is assessed through a Monte Carlo study where several contamination schemes are considered. An application to a real data set is also presented to illustrate the benefits of the methodology.

e-mail: **paulocanas@gmail.com**

A ROBUST MIXED LINEAR MODEL FOR HERITABILITY ESTIMATION IN PLANT STUDIES

Vanda M. Lourenço*, Nova University of Lisbon, Portugal

Paulo C. Rodrigues, Federal University of Bahia, Brazil

Miguel S. Fonseca, University of Lisbon, Portugal

Ana M. Pires, University of Lisbon, Portugal

Heritability (H2) refers to the extent of how much a certain phenotype is genetically determined. Knowledge of H2 is crucial in plant studies to help perform effective selection. Once a trait is known to be high heritable, association studies are performed so that the SNPs underlying those traits' variation may be found. Here, regression models are used to test for associations between phenotype and candidate SNPs. SNP imputation ensures that marker information is complete, so both the coefficient of determination (R2) and H2 are equivalent. One popular model used in these studies is the animal model, which is a linear mixed model (LMM) with a specific layout. However, when the normality assumption is violated, as other likelihood-based models, this model may provide biased results in the association analysis and greatly affect the classical R2. Therefore, a robust version of the REML estimates for the LMM to be used in this context is proposed, as well as a robust version of a recently proposed R2. The performance of both classical and robust approaches for the estimation of H2 is thus evaluated via simulation and an example of application with a maize data set is presented.

e-mail: **vmml@fct.unl.pt**

CANCER INCIDENCE AND SUPERFUND SITES IN FLORIDA

Emily Leary*, University of Missouri

Alexander Kirpich, University of Florida

Uncontrolled hazardous waste sites have the potential to adversely impact human health and damage or disrupt ecological systems and the greater environment. Decades have passed since the Superfund law was enacted, allowing increased exposure time to these potential health hazards but also allowing advancement of analysis techniques. In this study, statewide cancer incidence in Florida is analyzed to determine if differences in incidence exist in counties containing Superfund sites compared to counties that do not. Spatial and non-spatial analyses models are utilized and results compared. Preliminary results indicate evidence that level of hazard, proportion of water located around Superfund site, and aggregated exposure score have potential positive association with cancer incidence from 1986 to 2010 in Florida. Additionally, results indicate evidence of heterogeneity among cancer incidence rates.

e-mail: **learye@health.missouri.edu**

97. CONTRIBUTED PAPERS: Pooled Biospecimens and Diagnostic Biomarkers

HIERARCHICAL GROUP TESTING FOR MULTIPLE INFECTIONS

Peijie Hou*, University of South Carolina

Joshua M. Tebbs, University of South Carolina

Christopher R. Bilder, University of Nebraska, Lincoln

Group testing, where individuals are tested initially in pools, is often used to screen a large number of individuals for rare diseases. Triggered by the recent development of assays that detect multiple infections, large-scale screening programs now involve testing individuals in pools for multiple infections simultaneously. Tebbs, McMahan, and Bilder (2013, Biometrics) recently evaluated the performance of a two-stage hierarchical algorithm used to screen for chlamydia and gonorrhea as part of the Infertility Prevention Project in the United States. In this article, we generalize this work to accommodate a larger number of stages. To derive the operating characteristics of higher-stage hierarchical algorithms with more than one infection, we view the

pool decoding process as a finite-state Markov chain. Taking this conceptualization enables us to derive closed-form expressions for the expected number of tests and classification accuracy rates in terms of transition probability matrices. When disease probabilities are small, we offer compelling evidence that higher-stage algorithms can provide significant savings when screening a population for multiple infections. We also demonstrate that if prevalence estimation is an additional goal, two-stage algorithms provide most of the benefits in terms of estimation efficiency.

e-mail: **houp@email.sc.edu**

KEEPING RISK CALCULATORS CURRENT

Donna Pauler Ankerst*, Technical University Munich and University of Health Science Center at San Antonio

Andreas Strobl, Technical University Munich

As clinical practice increasingly focuses on personalized medicine and more contemporary large scale data from research consortiums become publicly available, so too must commonly-used clinical risk prediction tools evolve. As a particular example, the prostate cancer clinical landscape has undergone several changes over the past decade, and these likely necessitate that existing prostate cancer risk calculators be re-calibrated in order to remain accurate. In this talk we outline revision combined with shrinkage methods proposed by Steyerberg (2010) for periodically updating an existing risk calculator to account for serial changes over time. We apply the methods to the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) using yearly data from the Prostate Biopsy Collaborative Group (PBCG) comprising 25,772 prostate biopsies from five international cohorts. We evaluate the annual discrimination and calibration performance characteristics of the multiple revision alternatives relative to static use of the PCPTRC.

e-mail: **ankerst@ma.tum.de**

EVALUATION OF MULTIPLE BIOMARKERS IN A TWO-STAGE GROUP SEQUENTIAL DESIGN WITH EARLY TERMINATION FOR FUTILITY

Nabihah Tayob*, University of Texas MD Anderson Cancer Center

Kim-Anh Do, University of Texas MD Anderson Cancer Center

Ziding Feng, University of Texas MD Anderson Cancer Center

Motivated by an ongoing study to develop a screening test able to identify patients with undiagnosed Sjogren's Syndrome in a symptomatic population, we propose methodology to combine multiple biomarkers and evaluate their performance in a two-stage group sequential design that proceeds as follows: biomarker data is collected from first stage samples; the biomarker panel is built and evaluated; if the panel meets pre-specified performance criteria the study continues to the second stage and the remaining samples are assayed. The design allows us to conserve valuable specimens in the case of inadequate biomarker performance. We propose a nonparametric conditional bootstrap algorithm that uses all the study data to provide unbiased estimates of the biomarker combination rule and the sensitivity of the panel corresponding to specificity of 1-t on the receiver operating characteristic curve (ROC). The Copas & Corbett (2002) correction, for bias resulting from using the same data to derive the combination rule and estimate the ROC, was also evaluated and a modified version was incorporated. An extensive simulation study was conducted to evaluate finite sample performance and propose guidelines for designing studies of this type.

e-mail: **tayob@umich.edu**

FLEXIBLE AND ACCESSIBLE SEMI-PARAMETRIC METHODS FOR ANALYZING POOLED BIOSPECIMENS

Emily M. Mitchell*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Robert H. Lyles, Emory University

Amita K. Manatunga, Emory University

Enrique F. Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Pooling involves the physical combination of biospecimens into a single composite sample prior to performing lab assays. While pooling has various benefits, researchers may be hesitant to adopt a pooling strategy since appropriate statistical methods are still being developed, and existing methods for

individually-measured specimens may not directly apply to pools. This is particularly true when a biomarker is treated as the outcome in a regression model, since measurements are often positive and right-skewed. Current methods for analyzing this type of data are either computationally expensive or limited to specific pool types. In this study, we propose a novel, flexible and accessible estimation technique for a right-skewed outcome subject to pooling, regardless of pool type. We use simulations to demonstrate the efficacy of our proposed method compared with existing methods. Our simulations, along with analysis of data from the Collaborative Perinatal Project (CPP), demonstrate that when appropriate estimation techniques are applied to strategically-formed pools, valid and efficient estimation can be achieved. This novel method contributes to the base of available statistical tools to analyze pooled specimens and will help empower researchers to more confidently consider pooling as a potential study design.

e-mail: **emily.mitchell@nih.gov**

ESTIMATING INDIVIDUALIZED DIAGNOSTIC RULES IN THE ERA OF PERSONALIZED MEDICINE

Ying Liu*, Columbia University

Yuanjia Wang, Columbia University

Chaorui Huang, Cornell University

Donglin Zeng, University of North Carolina, Chapel Hill

Recent trend in disease screening calls for shifting from population-based screening strategies to more personalized risk-based strategies. For example, in breast cancer screening, advanced imaging technologies have made it possible to move away from "one-size-fits-all" screening guidelines to targeted risk-based screening for those who are in need. Similarly, for neurological disorders, multiple imaging modalities may be measured and their diagnostic performances vary across subjects so that applying the most accurate modality to the patients who would benefit the most requires personalized strategy. To address these needs, we propose novel machine learning methods to estimate personalized decision rules for medical screening or diagnosis to maximize a weighted combination of sensitivity and specificity for subgroups of subjects. Specifically, we frame the optimization as a weighted classification problem where we use a weighted supportive vector machine to obtain the solutions. First, we develop methods that can be applied to estimate personalized diagnostic rules where competing modalities or screening strategies are observed on each subject (paired design). Second, we also develop a kernel-based method for studies where not all subjects receive both modalities (unpaired design). We study theoretical properties including consistency and risk bound of the personalized diagnostic rule under the causal inference framework. We conduct extensive simulation studies for both paired and unpaired design to demonstrate that our proposed method can significantly improve the empirical area under the receiver operating curve (AUC). Lastly, we analyze data collected from a brain imaging study of Parkinson's disease using FDG-PET and DTI-MRI imaging modalities with paired and unpaired designs, where a personalized modality assignment is estimated to improve empirical AUC significantly compared to a "one-size-fits-all" assignment.

e-mail: **summeryingl@gmail.com**

ANALYSIS OF UNMATCHED POOLED CASE-CONTROL DATA

Neil J. Perkins*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Emily M. Mitchell, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Enrique F. Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

When studying new biomarkers, pooled study designs, where individual biospecimens are combined and assayed, can minimize cost while maintaining statistical efficiency. Logistic regression and maximum likelihood methods have been developed to estimate the association between a dichotomous outcome and a pooled exposure, where pools are matched on disease status. Exploiting characteristics of the gamma distribution, we have developed a more flexible model for analyzing data containing pooled measurements, where pools can be of mixed disease status. In studies employing pooling strategies, such a flexible approach will be essential to analyzing secondary and conditional outcomes when pools are formed based on the primary outcome. We use simula-

tion studies to assess consistency and efficiency of risk effect estimates by comparing maximum likelihood estimates of odds ratios using all pools to the more standard approaches that only allow for matched pools. In a conditional analysis of pregnancy outcomes and pooled cytokine measurements, we demonstrate the efficacy of our method when application of the standard approaches is precluded by an insufficient number of matched pools.

e-mail: **perkinsn@mail.nih.gov**

ESTIMATING TP53 MUTATION CARRIER PROBABILITY IN FAMILIES WITH LI-FRAUMENI SYNDROME USING LFSpro

Gang Peng*, University of Texas MD Anderson Cancer Center

Jasmina Bojadzieva, University of Texas MD Anderson Cancer Center

Mandy L. Ballinger, Peter MacCallum Cancer Centre, Melbourne, Australia

David M. Thomas, The Kinghorn Cancer Centre and Garvan Institute, Sydney, Australia

Louise C. Strong, University of Texas MD Anderson Cancer Center

Wenyi Wang, University of Texas MD Anderson Cancer Center

Given the cancer spectrum and onset in Li-Fraumeni syndrome (LFS) and limitations of the clinical criteria, accurate identification of candidates for prospective TP53 mutation testing has been difficult. A more efficient prediction tool is needed for LFS identification, management and screening, which should ultimately decrease mortality. We proposed LFSpro that is built on a Mendelian model and estimates TP53 mutation probability through the Elston-Stewart algorithm, incorporating de novo mutation rates. With independent validation data from 765 families (19,530 individuals in the United States [pediatric-onset sarcoma] and Australia [adult-onset sarcoma]), we compared estimations using LFSpro versus classic LFS and Chompret clinical criteria. LFSpro outperformed Chompret and classic criteria in the pediatric sarcoma cohort and was comparable to Chompret criteria in the adult sarcoma cohort. Sensitivity analysis on de novo mutation rates showed that in both cohorts, a rate of 5e-4 gave LFSpro the best prediction performance. We developed and validated a clinically accessible tool that incorporates de novo mutation rates to accurately estimate TP53 mutation carriers. Family history of cancer evolves, and LFSpro is sensitive to mutation carriers in families newly presenting in high-risk clinics and in families followed for years. It is more broadly applicable than the clinical criteria.

e-mail: **gpeng1@mdanderson.org**

98. CONTRIBUTED PAPERS: Multiple Testing and Variable Selection

BAYES FACTOR APPROACHES FOR HYPOTHESIS TESTING IN ANOVA MODELS

Min Wang*, Michigan Technological University

We examine the issue of hypothesis testing in analysis-of-variance (ANOVA) designs. We first reparameterize the ANOVA model with constraints for uniqueness into a classical linear regression model without constraints. Such reparameterization allows us to consider mixtures of g-priors on the regression coefficients with a hyperprior to g. With a special choice of prior specifications, we propose an explicit closed-form expression of Bayes factor, which is easy to apply in practice, and also easy to teach in undergraduate statistics with emphasize on Bayesian thinking. In particular, we present asymptotic properties of Bayes factors with various choices of g for ANOVA models with divergence dimensionality under different asymptotic scenarios. We show that Bayes factor under the mixture g-priors have very similar asymptotic properties: they are always consistent under the null and are consistent under the alternative except for a small region around the null model. Applications to two real-data sets are analyzed for illustrative purposes.

e-mail: **minwang@mtu.edu**

A MULTIFUNCTIONAL BAYESIAN PROCEDURE FOR DETECTING COPY NUMBER VARIATIONS FROM SEQUENCING READ DEPTHS

Yu-Chung Wei*, U.S. Food and Drug Administration and National Chiao Tung University, Taiwan

Guan-Hua Huang, National Chiao Tung University, Taiwan

Copy number variations (CNVs) are genomic structural mutations with abnormal gene fragment copies. Read depths signal mirrors the variants directly from the next generation sequencing data.

Some tools have been published to predict CNVs by depths, but most of them just apply to a specific data type. Providing a multifunctional detection algorithm that can easily make use of a variety of data types is difficult but valuable. We develop a multifunctional COpy Number variation detection tool by a BaYesian procedure, CONY, which adopts an efficient reversible jump Markov chain Monte Carlo inference algorithm for analyzing sequencing read depths. CONY is suitable for reads from both whole genome and targeted exome sequencing. Additionally, CONY can be applied not only to an individual for estimating the absolute number of copies but also to case-control samples for detecting patient specific variations. We demonstrate this pragmatic approach with targeted region exome sequencing data from National Taiwan University Hospital. We also evaluate the performance of CONY and compare it with competing approaches using both simulations and real data from the 1000 Genomes Project.

e-mail: **weiyuchung@gmail.com**

INFERRING THE GLOBAL GENETIC ARCHITECTURE OF GENE TRANSCRIPTS FROM ULTRAHIGH-DIMENSIONAL MOLECULAR DATA

Kirk Gosik*, The Pennsylvania State University

Rongling Wu, The Pennsylvania State University

Knowledge about how changes in gene expression are encoded by expression quantitative trait loci (eQTLs) is a key to construct the genotype-phenotype map for complex traits or diseases. Traditional eQTL mapping is to associate one transcript with a single marker at a time, thereby limiting our inference about a complete picture of the genetic architecture of gene expression. In this talk, I present an innovative application of variable selection approaches to systematically detect main effects and interaction effects among all possible loci on differentiation and function of gene expression. Forward-selection-based procedures were particularly implemented to tackle complex covariance structures of gene-gene interactions. We reanalyzed a published genetic and genomic data collected in a mapping population of Caenorhabditis elegans, gaining new discoveries on the genetic origin of gene expression differentiation, which could not be detected by a traditional one-locus/one-transcript analysis approach.

e-mail: **kgosik@hmc.psu.edu**

STATISTICAL INFERENCE FOR HIGH DIMENSIONAL LINEAR REGRESSION WITH LINEAR CONSTRAINTS AND APPLICATION TO MICROBIOME STUDY

Pixu Shi*, University of Pennsylvania

Anru Zhang, University of Pennsylvania

Hongzhe Li, University of Pennsylvania

We consider the statistical inference problem for high-dimensional linear regression models with linear constraints on the regression coefficients. Such models include the log-contrast model for compositional covariates as a special case. We develop a penalized estimation procedure for estimating the regression coefficients and for selecting variables to account for the linear constraints. We also propose a method to obtain de-biased estimates that are asymptotically unbiased and derive its joint asymptotic distribution. The results provide valid confidence intervals of the regression coefficients and can be used to obtain the p-values. Simulation results show that variable selection based on the confidence intervals or the p-values can improve those from Lasso regression. Application to a gut microbiome data set has identified three bacterial genera that are associated with the body mass index, which can explain about 24% of the variance.

e-mail: **pixushi@mail.med.upenn.edu**

TAKING INTO ACCOUNT OVERREPRESENTED PATTERNS IN GENE EXPRESSION ANALYSIS

Megan Orr*, North Dakota State University

Ekua Bentil, North Dakota State University

Gene expression technologies allow expression levels to be compared across treatments for thousands of genes simultaneously. Many methods exist for identifying differentially expressed genes while controlling multiple testing error. However, most methods do not take into account the overrepresentation of observed patterns (compared to the expected patterns under the null hypothesis) across groups if such patterns exist. An example of an overrepresented pattern includes a large majority of genes being up-regulated compared to down-regulated in a two-sample study.

Another example includes a high proportion of genes exhibiting monotonicity in the sample mean expression levels as treatment dose levels increase in a dose-response experiment with more than two treatments. We propose new methods that take into account these overrepresented patterns and identify differentially expressed genes while controlling false discovery rate. The proposed methods are compared to traditional gene expression analysis procedures through simulation studies. Real gene expression data sets are analyzed to illustrate the usefulness of the proposed methods.

e-mail: **megan.orr@ndsu.edu**

BAYESIAN SCREENING FOR GROUP DIFFERENCES IN METHYLATION ARRAY DATA

Eric F. Lock*, University of Minnesota

In modern biomedical research, it is common to screen for differences between groups in many variables that are measured using the same technology. Motivated by DNA methylation data, this talk focuses on screening for equality of group distributions for many variables with shared distributional features such as common support, common modes and common patterns of skewness. We propose a Bayesian nonparametric testing methodology, which improves performance by borrowing information across the different variables and groups through shared kernels and a common probability of group differences. The inclusion of shared kernels in a finite mixture, with Dirichlet priors on the different weight vectors, leads to a simple framework for testing and we describe an implementation that scales well for high-dimensional data. We provide some theoretical results, compare with existing frequentist and Bayesian nonparametric testing methods, and describe an application to breast cancer methylation data from the Cancer Genome Atlas.

e-mail: **elock@umn.edu**

INCORPORATING ENCODE INFORMATION INTO SNP-BASED PHENOTYPE PREDICTION

Yue-Ming Chen*, University of Texas School of Public Health, Houston

Peng Wei, University of Texas School of Public Health, Houston

Recent studies show that most genome-wide association studies (GWAS) identified single nucleotide polymorphisms (SNPs) fall outside of the protein-coding regions (Hindorff et al 2009) and these trait-associated SNPs may play a role in regulatory networks and tend to be highly correlated, i.e., in high linkage disequilibrium (LD), with the functional SNPs (Maurano et al 2012, Schaub et al 2012). Tools for annotating functional variation in human genome, such as RegulomeDB (Boyle et al 2012), integrate enriched regulatory information from multiple resources including the Encyclopedia of DNA Elements (The ENCODE Project Consortium 2012). Using ENCODE information, we annotated all SNPs by assigning scores which reflect their regulatory function in the human genome. Based on these functional annotations, we constructed a weighted genetic prediction framework for complex traits. Two frameworks were considered as prototypes for incorporating the ENCODE information: the adaptive lasso (Zou 2006) and the weighted false discovery rate (Genovese et al 2006). Using simulations and real data analysis, we compared and contrasted our methods to a benchmark best linear unbiased prediction (BLUP) method that did not consider prior biological information (Speed and Balding 2014).

e-mail: **ychen18@mdanderson.org**

99. CONTRIBUTED PAPERS: Parameter Estimation in Hierarchical and Non Linear Models

A HIERARCHICAL BAYESIAN METHOD FOR WELL-MIXED AND TWO-ZONE MODELS IN INDUSTRIAL HYGIENE

Xiaoyue Zhao*, University of Minnesota

Susan Arnold, University of Minnesota

Dipankar Bandyopadhyay, University of Minnesota

Gurumurthy Ramachandran, University of Minnesota

Sudipto Banerjee, University of California, Los Angeles

In industrial hygiene and occupational exposure assessment, a worker's exposure to chemical, physical and biological agents is customarily modeled using deterministic physical models that study exposures based on the distance to a contaminant source. When field or experimental observations are available, non-linear stochastic regression models have been employed for filtering the noise and estimating the physical parameters. This, however, has been shown to be inefficient. Here, we develop

a likelihood-based model using a discrete version of the underlying differential equations. We cast this within a Bayesian dynamic linear model framework and use posterior predictive measures to predict future exposure concentrations and estimate the underlying physical parameters. We show that this method excels over simple non-linear regression methods by providing more accurate predictions, while it offers more reliable physical parameter estimates than Bayesian melding approaches using Gaussian processes. We implement our method entirely within the RJAGS software environment using two well-established physical models: (i) a well-mixed model, and (ii) a two-zone model.

e-mail: **zhao0378@umn.edu**

PARAMETER ESTIMATION: A BAYESIAN INFERENCE APPROACH

Romarie Morales*, Arizona State University

We focus on improving the current methodology for estimating transmission parameters by applying the Bayesian statistical framework to a probabilistic model of disease transmission. We then generalize this formulation to any disease. The Bayesian method takes into account the stochasticity of disease transmission and provides more robust parameter estimates. Increasing estimation accuracy through adoption of the Bayesian framework will equip policymakers with better tools for mitigating the effects of an epidemic.

e-mail: **rmorale7@asu.edu**

BIAS AND CONFIDENCE INTERVAL CORRECTION IN FOUR PARAMETER LOGISTIC MODELS

Bronlyn Wassink*, Michigan State University

Tapabrata Maiti, Michigan State University

Using a four parameter logistic model is a commonly used method to model dose-response data. The four parameters are the minimum expected response, the maximum expected response, the EC50, and the Hill parameter. The EC50, a method of quantifying a drug's potency, refers to the dose that produces an expected response halfway between baseline and the maximal expected response, and the Hill parameter is a measure of the steepness of the expected response when the dose is near the EC50. In the small sample setting, such as in a pilot study, the maximum likelihood estimates of the four parameters are biased and not near-normally distributed. Therefore, asymptotically-based methods for estimating the parameter biases and confidence intervals do not always produce reliable results. In this talk, we investigate different methods for computing bias-corrected parameters, and propose the use of Beta and Gamma distributions to improve confidence intervals. Simulation results on the coverage probability and interval widths show the improved performance of the proposed method over the existing classical procedures.

e-mail: **bronlyn@gmail.com**

ROBUST MIXED-EFFECTS MODEL FOR CLUSTERED FAILURE TIME DATA: APPLICATION TO HUNTINGTON'S DISEASE EVENT MEASURES

Tanya P. Garcia*, Texas A&M University

Yanyuan Ma, University of South Carolina

Yuanjia Wang, Columbia University

Karen Marder, Columbia University

An important goal in clinical and statistical research is describing the distribution for clustered failure times, which have a natural intra-class dependency and are subject to censoring. We propose to handle these inherent challenges with a novel approach that does not impose a proportional hazards assumption nor treat the dependency with a random effect having a prespecified distribution. Rather, using a logit transformation, we relate the distribution for clustered failure times to covariates and a random intercept. To avoid any misspeciffication issues, the covariates are modeled using unknown functional forms and the random intercept is kept distribution-free and allowed to depend on covariates. Over a range of time points, the model is shown to be reminiscent of an additive logistic mixed effect model, from which we can handle censoring via pseudo-value regression and apply semiparametric techniques to factor out the unknown random effect. The resulting estimators are shown to be simple, consistent, and robust to any nuances of the random effect distribution. We illustrate the method's robustness and competitiveness to existing methods in a simulation study that involves diferent random-intercept distributions and different dependency structures between

the random intercept and model covariates. Lastly, we apply our method to the Cooperative Huntington's Observational Research Trial data, to provide new insights into differences between motor and cognitive impairment event times in genetically predisposed patients.

e-mail: **tpgarcia@stat.tamu.edu**

STACKED SURVIVAL MODELS FOR CENSORED QUANTILE REGRESSION

Kyle Rudser*, University of Minnesota

Andrew Wey, University of Hawaii

John Connett, University of Minnesota

Inference on quantiles of survival is an attractive alternative to using the hazard ratio for comparing groups that has a meaningful interpretation based on units of time in the censored data setting. Censored quantile regression allows contrasts across groups while adjusting for other factors. Current censored quantile regression methods often rely upon the fairly strong assumptions of unconditionally independent censoring or linearity in all quantiles. We examine the use of stacked survival models in a distribution-free framework for adjusted contrasts of quantiles of survival. By minimizing prediction error, stacking estimates optimally weighted combinations of survival models that can span parametric, semi-parametric, and non-parametric models. As such, the low variance of approximately correct parametric models can be exploited while maintaining the robustness of nonparametric models. We analyze the performance on estimation and inference via simulations and found that using stacked survival models can have potential to provide robust estimation at little loss of precision. We also illustrate the approaches using lung transplantation data.

e-mail: **rudser@umn.edu**

THE CoGAUSSIAN DISTRIBUTION: A MODEL FOR RIGHT SKEWED DATA

Govind S. Mudholkar, University of Rochester

Ziji Yu*, University of Rochester

Saria S. Awadalla, University of 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 paper, the essentials, namely the concept of R-symmetry, the roles of the mode and harmonic variance as, respectively, the centrality and dispersion parameters of the CoGaussian distribution, are introduced. The pivotal role of the CoGaussian family in the class of R-symmetric distributions and the estimation, testing and characterization properties are discussed. The similarities between the Gaussian and CoGaussian distribution, namely the G-CoG analogies, are summarized.

e-mail: **ziji_yu@urmc.rochester.edu**

100. New Statistical Methods in the Environmental Health Sciences

NEW STATISTICAL MODELS TO DETECT VULNERABLE PRENATAL WINDOW TO CARCINOGENIC POLYCYCLIC AROMATIC HYDROCARBONS ON FETAL GROWTH

Lu Wang*, University of Michigan

Prenatal exposure to carcinogenic polycyclic aromatic hydrocarbons (c-PAHs) through maternal inhalation induces higher risk for a wide range of fetotoxic effects. However, the most health-relevant dose function from chronic gestational exposure remains unclear. Whether there is a gestational window during which the human embryo/fetus is particularly vulnerable to PAHs has not been examined thoroughly. We consider a longitudinal semiparametric-mixed effect model to characterize the individual prenatal PAH exposure trajectory, where a nonparametric cyclic smooth function plus a linear function are used to model the time effect and random effects are used to account for the within-subject correlation. We propose a penalized least squares approach to estimate regression coefficients and the nonparametric function of time. The smoothing parameter and variance components are selected using the generalized cross-validation criteria. The estimated subject-specific trajectory of prenatal exposure is linked to the birth outcomes through a set of functional linear models, where the coefficient of log PAH exposure is a fully nonparametric

function of gestational age. This allows the effect of PAH exposure on each birth outcome to vary at different gestational ages, and the window associated with significant adverse effect is identified as a vulnerable prenatal window to PAHs on fetal growth.

e-mail: **luwang@umich.edu**

DIMENSION REDUCTION FOR SPATIALLY MISALIGNED MULTIVARIATE AIR POLLUTION DATA

Adam Szpiro*, University of Washington

Emerging monitoring technologies provide high-dimensional characterizations of air pollution, promising a more nuanced understanding of which pollutants/mixtures are responsible for health effects observed in single pollutant epidemiology studies. There are two interrelated challenges (i) interpreting the association parameters requires dimension reduction and (ii) spatial misalignment requires prediction modeling. We propose a paradigm for spatially predictive dimension reduction, exemplified by predictive sparse principal component analysis. We seek sparse principal component loadings that explain a large proportion of the variance in the monitoring data, while ensuring the corresponding low-dimensional representations are predictable at subject locations. We apply the proposed method to long-term multi-pollutant data from regulators monitors across the United States and utilize the predicted low-dimensional exposures to refine our understanding of a previously observed association between hypertension and exposure to fine particulate matter.

e-mail: **aszpiro@u.washington.edu**

EVALUATING ALTERATIONS IN REGRESSION COEFFICIENTS DIRECTED BY TOXICANT MIXTURES

Peter X. K. Song*, University of Michigan

Shujie Ma, University of California, Riverside

We propose a new linear mixed effects model for longitudinal data that enables us to study dynamics of interest in the process of somatic growth. This new methodology is useful to assess if and how the rate of growth may be intervened by exposure variables such as mixtures of toxicants (e.g. PBA and phthalates). Interestingly, most of such interveners are of small size in their effects, and the traditional statistical method fails to detect their statistical significance. Our new modeling strategy incorporates a certain type of principal component in the formation of regression coefficients, termed as index coefficients in that low-effect toxicants are combined into possibly strong toxicant groups. Statistical estimation and inference in such model is challenging because it contains nonlinear interactions between the toxicant groups and covariates of interest (e.g. age or time). The proposed models and methods are motivated and illustrated by an analysis of child growth data to evaluate alterations in growth rates incurred by mother's exposures to endocrine disrupting compounds during pregnancy.

e-mail: **pxsong@umich.edu**

101. Novel Phase II and III Clinical Trial Designs for Cancer Research that Incorporate Biomarkers and Non-standard Endpoints

NOVEL PHASE II AND III DESIGNS FOR ONCOLOGY CLINICAL TRIALS, WITH A FOCUS ON BIOMARKER VALIDATION

Daniel J. Sargent*, Mayo Clinic

Increasing scientific knowledge is creating both substantial opportunities and challenges in oncology drug development. As diseases are sub-stratified into often biomarker-based groups, usual paradigms for phase II and III disease may no longer apply. Enrichment designs are appropriate when preliminary evidence suggest that patients with/without that marker profile do not benefit from treatments in question; however this may leave questions unanswered (e.g. Herceptin and breast cancer). An unselected design is optimal where preliminary evidence regarding treatment benefit and assay reproducibility is uncertain. Adaptive analysis designs allow for pre-specified marker defined subgroup analyses of data from a RCT. We discuss features of these various novel design strategies in the context of real trials.

e-mail: **sargent.daniel@mayo.edu**

STRATIFIED SINGLE ARM PHASE 2 DESIGN FOR FINDING A BIOMARKER GROUP THAT BENEFITS FROM TREATMENT

Irina Ostrovnaya*, Memorial Sloan Kettering Cancer Center

Emily Zabor, Memorial Sloan Kettering Cancer Center

In phase II studies of cancer treatments, there is growing interest in investigating whether a treatment is effective among the general population of patients, or only among a subgroup of patients defined by a biomarker value or genetic mutation. While there are many such statistical "enrichment" designs developed for the randomized clinical trials, there are very few available for single arm studies when current treatment is compared to historical controls, and these designs are not easily applied. Here we propose a simple two-stage single arm stratified design for binary endpoint similar to Simon two stage design that allows for discontinuation of the marker negative subgroup in the first stage if there is not enough evidence of treatment efficacy in that subgroup. The software for calculating the sample size for the proposed design is publically available. This design often requires fewer patients than two parallel Simon two stage designs in marker positive and negative patients independently.

e-mail: **ostrovni@mskcc.org**

LUNG-MAP: A PHASE II/III BIOMARKER-DRIVEN MASTER PROTOCOL FOR SECOND LINE THERAPY OF SQUAMOUS CELL LUNG CANCER

Mary W. Redman*, Fred Hutchinson Cancer Research Center

Lung-MAP is a large scale, screening/clinical registration protocol that genomically screens patients with advanced stage lung squamous cell cancer moving to second-line therapy, and uses the screening results to direct each patient to a therapeutic phase II/III sub-study. Based on the results of the genomic analysis, patients will either be assigned to one of the biomarker-driven sub-studies or to the "non-match" sub-study for patients with none of the eligibility biomarkers, and subsequently randomized between an investigational therapy or standard of care. The biomarker-driven sub-studies are designed around a genotypically-defined alteration in the tumor and a drug that targets it. The non-match study is designed around an investigational agent with the potential for efficacy in a broader/less selected population. Each Lung-MAP sub-study functions autonomously and will open and close independently of the other sub-studies. When an endpoint for a sub-study is met, that drug-biomarker specific combination may proceed to FDA for approval review of the new drug with its matching companion diagnostic. When an endpoint is not met, that sub-study will be closed and another modular sub-study of a different agent will be initiated. While organized by SWOG, Lung-MAP is the result of unprecedented public-private collaboration between government, non-profit, and for-profit organizations and involves the entire National Clinical Trials Network (NCTN) in study leadership and trial participation. The Lung-MAP study activates[activated] in June 2014 with the intention to enroll 1,000 patients per year onto the therapeutic sub-studies. In this talk I will describe the statistical design of the Lung-MAP study and discuss the current status of the study.

e-mail: **mredman@fhcrc.org**

RANDOMIZED PHASE II DESIGN TO STUDY 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 extracrainially 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. Through the use of simulations, we describe how these issues affect power and sample size in Phase II studies and propose a design that reduces their impact.

e-mail: **patils@mskcc.org**

102. Novel Statistical Methods to Decipher Gene Regulation Using Sequence Data

ON THE DETECTION OF NONLINEAR AND INTERACTIVE RELATIONSHIPS IN GENOMIC DATA

Bo Jiang, Harvard University

Jun Liu*, Harvard University

I will discuss a few recent results from my group aiming at the detection of non-lineardependence and interactive effects of several random variables. These approaches were developed by taking a Bayesian view on the inverse-slicing idea first proposed by Ker-Chau Li for dimension reduction. To detect whether a new covariate X can influence the continuous response Y conditional on a set of selected covariates Z, we assume that there is a latent slicing variable Y which indicates how Y can be sliced into a few levels. We then model the conditional distribution of [X|Z] at each level of Y and compare with the overall conditional distribution [X|Z] unconditional of Y. We can also provide a prior on the latent slicing variable and average over all posible slicing schemes weighted by their prior probabilities. We will show how these methods are applied to bioinformatics problems such as gene-set enrichment analysis, transcription regulation analysis, eQTL studies, and others.

e-mail: **jliu@stat.harvard.edu**

STATISTICAL ANALYSIS OF DIFFERENTIAL ALTERNATIVE SPLICING USING RNA-Seq DATA

Mingyao Li*, University of Pennsylvania

Yu Hu, University of Pennsylvania

Cheng Jia, University of Pennsylvania

RNA sequencing (RNA-seq) allows an unbiased survey of the entire transcriptome in a high-throughput manner. It has rapidly replaced microarrays as the major platform for transcriptomics studies. A major application of RNA-seq is to detect differential alternative splicing (DAS), or differential transcript usage, across experimental conditions. Differential analysis at the transcript level is of great biological interest due to its direct relevance to protein function and disease pathogenesis. However, DAS analysis using RNA-seq data is challenging because of the difficulty of quantifying alternative splicing and various biases present in RNA-seq data. In this talk, I will present several statistical issues related to the analysis of DAS. I will discuss methods for detecting DAS for both paired and unpaired data, and compare the performance of exon-based and gene-based tests of DAS. I will show simulation results as well as some examples from real transcriptomics studies.

e-mail: **mingyao@mail.med.upenn.edu**

A CASE STUDY OF RNA-Seq DATA IN BREAST CANCER PATIENTS

Wei Sun*, University of North Carolina, Chapel Hill

We carry out a systematic study of RNA-seq data and its genetic architecture in 550 breast cancer patients from The Cancer Genome Atlas project. eQTL mapping of gene expression (measured by RNA-seq in tumor tissue) vs. germline genotype and tumor copy number aberrations show that both types of genetic variants have substantial influence on gene expression. We further assess such associations after deconvoluting gene expression from tumor cells and normal cells (e.g. stromal cells within the tumor tissue), and discuss possible scenarios to use such eQTL results to obtain further biological insights.

e-mail: **weisun@email.unc.edu**

UNIT-FREE AND ROBUST DETECTION OF DIFFERENTIAL EXPRESSION FROM RNA-Seq DATA

Hui Jiang*, University of Michigan

Ultra high-throughput sequencing of transcriptomes (RNA-Seq) has recently become one of the most widely used methods for quantifying gene expression levels due to its decreasing cost, high accuracy and wide dynamic range for detection. However, the nature of RNA-Seq makes it nearly impossible to provide absolute measurements of transcript concentrations. Several units or data summarization methods for transcript quantification have been proposed to account for differences in transcript lengths and sequencing depths across genes and samples. However, none

of these methods can reliably detect differential expression directly without further proper normalization. We propose a statistical model for joint detection of differential expression and data normalization. Our method is independent of the unit in which gene expression levels are summarized. We also introduce an efficient algorithm for model fitting. Due to the L0-penalized likelihood used by our model, it is able to reliably normalize the data and detect differential expression in some cases when more than half of the genes are differentially expressed in an asymmetric manner. The robustness of our proposed approach is demonstrated with simulations.

e-mail: **jianghui@umich.edu**

103. Flow Cytometry: Data Collection and Statistical Analysis

FLOW, MASS AND IMAGING CYTOMETRY FOR SINGLE CELL ANALYSIS: A FERTILE FIELD FOR BIOSTATISTICS RESEARCH

Richard H. Scheuermann*, J. Craig Venter Institute and University of California, San Diego

Yu Qian, J. Craig Venter Institute

Chiaowen Hsiao, University of Maryland, College Park

Monnie McGee, Southern Methodist University

Flow, mass and imaging cytometry are used to quantitatively assess the phenotypic characteristics of large numbers of single cells in complex biological specimens, with current technologies supporting the quantification of 10's to 100's of individual features in every cell. These technologies are being used to study normal and abnormal cell activation, differentiation, and function, to diagnose leukemia, lymphoma, and myeloproliferative disorder, and to identify novel biomarkers of therapeutic response and treatment outcome. In this presentation we will review each of these technologies, summarize some of the computational challenges associated with the high dimensionality, biological variability and technical artifacts inherent in the resulting data, and describe some of the computational and statistical methods that have been developed to process, analyze and interpret cytometry data. We will also present the results from the FlowCAP challenges (http://flowcap.flowsite.org) that were developed to compare the performance of these methods.

e-mail: **rscheuermann@jcvi.org**

COMPUTATIONAL IDENTIFICATION OF CELL POPULATIONS FROM CYTOMETRY DATA: METHODS, APPLICATIONS, AND INFRASTRUCTURE

Yu Qian*, J. Craig Venter Institute

Hyunsoo Kim, J. Craig Venter Institute

Shweta Purawat, University of California, San Diego

Rick Stanton, J. Craig Venter Institute

Ilkay Altintas, University of California, San Diego

Richard H. Scheuermann, J. Craig Venter Institute

This talk will present the design and implementation of a data-clustering algorithm, FLOCK, for computational identification of cell populations from multi-dimensional flow cytometry data. Then we will discuss improvements to FLOCK that allows for the analysis of higher dimensionality data such as those generated by mass and imaging cytometry. The methods will be presented in the context of their applications in biomedical basic and translational research. Both the impact and the limitations of the methods will be discussed. Besides FLOCK, several other computational methods have been developed and shown to provide excellent performance when compared with manual analysis. We will summarize the basic principles behind these approaches and review the existing infrastructure support for computational single cell data analysis. Through transforming the methods and their accessories into workflow steps, we are able to integrate and compare multiple methods in the same running environment for data-driven selection and optimization of the computational methods. We will report progress in the development of FlowGate; a Scientific Gateway that combines graphical user interfaces, data analytical platforms and workflow engines including GenePattern and bioKepler, and parallel computing support for processing and analyzing cytometry single cell data in an extensible, scalable, and reproducible way.

email: **mqian@jcvi.org**

MAPPING CELL POPULATIONS IN FLOW CYTOMETRY DATA FOR CROSS-SAMPLE COMPARISON USING THE FRIEDMAN-RAFSKY TEST

Chiaowen Joyce Hsiao*, University of Maryland, College Park

Mengya Liu, Southern Methodist University

Rick Stanton, J. Craig Venter Institute

Monnie McGee, Southern Methodist University

Yu Qian, J. Craig Venter Institute

Richard H. Scheuermann, J. Craig Venter Institute and University of California, San Diego

This talk presents FlowMap-FR, a novel method for comparative analysis of cell populations across flow cytometry (FCM) experiment samples. FlowMap-FR is based on the Friedman-Rafsky (FR) non-parametric test statistic, which is used to measure the equivalence of multivariate distributions. As applied to FCM data by FlowMap-FR, the FR test objectively quantifies the similarity between cell populations based on their shapes, sizes, and positions in the high-dimensional feature space. We will present the method and discuss the performance of FlowMap-FR in mapping cell populations under the different kinds of biological and technical sample variations that are commonly observed in FCM data. Our evaluation results show that FlowMap-FR is able to effectively identify equivalent cell populations across samples under scenarios of proportion changes and modest distribution shifts. As a statistical test, FlowMap-FR thus can be used to objectively determine when the expression of a cellular marker has become significantly different from a comparison population, thereby defining a new cellular phenotype by providing an objective measure for when a cell population has become functionally distinct. Through comparing cell populations, FlowMap-FR can also detect situations in which inappropriate splitting or merging of cell populations has occurred during gating procedures. We have implemented FlowMap-FR as a stand-alone R/Bioconductor package that is publicly available to the community.

e-mail: **chsiao@umiacs.umd.edu**

A NOVEL APPROACH TO MODELING IMMUNOLOGY DATA DERIVED FROM FLOW CYTOMETRY

Jacob A. Turner*, Baylor Institute for Immunology Research

This presentation will illustrate some of the distributional properties the variables from flow cytometry (FCM) studies exhibit. A novel modeling strategy denoted Layered Dirichlet Modeling (LDM) will be introduced to model proportions derived from FCM data. The LDM strategy takes into account that the variables are compositional and have a hierarchical structure that imposes correlation between the variables. The properties of the LDM testing procedures are explored. A data-driven tree finding algorithm is provided to find a hierarchy of relationships among FCM subpopulation when the hierarchy is missing or unknown. The motivation of LDM comes from a generalization of the Dirichlet distribution known as the Nested Dirichlet distribution.

e-mail: **jacob.turner1@baylorhealth.edu**

104. Statistical Methods In Chronic Kidney Disease

JOINT MODELING OF KIDNEY FUNCTION DECLINE, END STAGE KIDNEY DISEASE (ESRD), AND DEATH WITH SPECIAL CONSIDERATION OF COMPETING RISKS

Dawei Xie*, University of Pennsylvania

Wensheng Guo, University of Pennsylvania

Wei Yang, Merrill Lynch

Qiang Pan, University of Pennsylvania

Methods have been developed previously to jointly model a repeatedly measured continuous variable such as estimated glomerular filtration rate (eGFR) and a time-to-event outcome (such as ESRD), or jointly model two time-to-event outcomes that represent competing risks (such as ESRD and death). We propose to jointly model repeated measures of eGFR, ESRD and death to address the correlation between repeated measures of eGFR and ESRD/death and the competing risks between ESRD and death. Specifically, repeated measures of eGFR are modelled via a linear mixed effects model and the times to ESRD and death accelerated failure time frailty models. We further assume the linear and the frailty models share the same mixed effects. An EM algorithm is used to calculate the maximum likelihood estimates of the parameters. The method will be illustrated using data from the Chronic Renal Insufficiency Cohort (CRIC) study.

e-mail: **dxie@mail.med.upenn.edu**

JOINT MULTIPLE IMPUTATION FOR LONGITUDINAL OUTCOMES AND CLINICAL EVENTS WHICH TRUNCATE LONGITUDINAL FOLLOW-UP

Bo Hu*, Cleveland Clinic

Liang Li, University of Texas MD Anderson Cancer Center

Tom Greene, University of Utah

Longitudinal cohort studies often collect both repeated measurements of longitudinal outcomes and times to clinical events whose occurrence precludes further longitudinal measurements. Both types of data are usually subject to non-ignorable missingness due to informative dropout as well as intermittent missed visits. Although joint modeling of the clinical events and the longitudinal data can be used to provide valid statistical inference for target estimands in certain contexts, the application of joint models in medical literature is currently rather restricted due to the complexity of the joint models and the intensive computation involved. We propose a multiple imputation (MI) approach to jointly impute missing data of both the longitudinal and clinical event outcomes. With complete imputed datasets, analysts are then able to use simple and transparent statistical methods and standard statistical software to perform various analyses without dealing with the complications of missing data and joint modeling. We show that the proposed MI approach is flexible and easy to implement in practice. Numerical results are also provided to demonstrate its performance.

e-mail: **hub@ccf.org**

MODELING THE EFFECT OF BLOOD PRESSURE ON DISEASE PROGRESSION IN CHRONIC KIDNEY DISEASE USING MULTISTATE MARGINAL STRUCTURAL MODELS

Alisa J. Stephens*, University of Pennsylvania

Wei Peter Yang, University of Pennsylvania

Marshall M. Joffe, University of Pennsylvania

Tom H. Greene, University of Utah

In patients with chronic kidney disease, clinical interest often centers on determining treatments and exposures that are causally related to progression. Analyses of longitudinal clinical data in this population are often complicated by clinical events, such as end stage renal disease (ESRD) or death, and time-dependent confounding, where patient factors that are affected by past exposures are predictive of later exposures and outcomes. We developed multistate marginal structural models to assess the effect of time-varying systolic blood pressure on disease progression in subjects with CKD. The multistate nature of our models allows us to consider jointly as outcomes disease progression characterized by changes in the estimated Glomerular Filtration Rate (eGFR), the onset of (ESRD), and death, and thereby avoid unnatural assumptions of death or ESRD as an informative censoring event after which disease progression can occur. Under a Markov assumption, we model the causal effect of systolic blood pressure on the probability of transitioning into one of five disease states given the current state. We use inverse probability weights to account for potential time-varying confounders, including urine protein, creatinine, and hemoglobin, in estimating the effect of blood pressure on the probability of transitioning among states. We apply our model to data from the Chronic Renal Insufficiency Cohort, a multisite observational study of patients with CKD.

e-mail: **alisaste@mail.med.upenn.edu**

DYNAMIC PREDICTION OF CLINICAL EVENTS USING LONGITUDINAL BIOMARKERS IN A COHORT STUDY OF CHRONIC RENAL DISEASE

Liang Li*, University of Texas MD Anderson Cancer Center

In longitudinal studies, prognostic biomarkers are often measured longitudinally. It is of both scientific and clinical interest to predict the risk of clinical events, such as disease progression or death, using these longitudinal biomarkers and possibly other time-dependent and time-independent information. This problem can be done in two ways. One is to build a joint model of longitudinal data and clinical events data, and draw predictions from the fitted model using the posterior distributions. The other approach is the landmark dynamic prediction model, which is a system of prediction models that evolve with the landmark times. We review the pros and cons of the two approaches in the context of the chronic renal disease studies, and present our research using the landmark approach. One drawback of the current landmark methodology is that the predictors are difficult to define when the

longitudinal data are measured at irregularly spaced time points. We present a solution to this problem by an augmentation of the landmark model. We apply our proposed methodology to the African American Study of Kidney Disease and Hypertension (AASK) to derive and test a dynamic prediction model for quantifying the time-varying risk of end stage renal disease.

e-mail: **LLi15@mdanderson.org**

105. Challenging Statistical Issues in Imaging

RELATING DEVELOPMENTAL TRANSCRIPTION FACTORS BASED ON DROSOPHILA EMBRYONIC GENE EXPRESSION IMAGES

Siqi Wu*, University of California, Berkeley

TFs play a central role in controlling gene expression. A fundamental problem in systems biology is to understand the interactions between the TFs, or, in other words, to understand the transcription networks. For the first time in any metazoan animal, we have imaged spatiotemporal gene expression of all known and predicted TFs during Drosophila embryogenesis. We used 400+ images of 155 TFs with restricted gene expression during early development and developed novel methods relate these TFs to each other in order to shed light on the TF cascades that trigger transcription. We borrowed the idea of Nonnegative Matrix Factorization (NMF) from computational neuroscience/computer vision and decomposed the expression patterns contained in the images into a group of data-driven subsets, called "principal patterns". The representation of the expression patterns as learned principal patterns allows for a compact and interpretable representation. Based on the learned patterns, we constructed spatially local TF networks. The constructed networks agreed well with known networks such as the gapgene network. More interestingly, our method also identified a number of previously undescribed TFs as possible new candidates regulators of the gapgene network. We are currently validating the candidate TFs in knockout experiments. Thus, our dataset, representation and modeling approach have shown significant potential for modeling and identifying novel components of gene networks during animal development.

e-mail: **binyu@stat.berkeley.edu**

ANALYSIS OF POINT PATTERN IMAGING DATA USING LOG GAUSSIAN COX PROCESSES WITH SPATIALLY VARYING COEFFICIENTS

Timothy D. Johnson*, University of Michigan

Thomas E. Nichols, University of Warwick

Log Gaussian Cox Processes (LGCP) are used extensively to model point pattern data. In a LGCP, the log intensity function is modeled semi-parametrically as a linear combination of spatially varying covariates with scalar coefficients plus a Gaussian process that models the random spatial variation. Almost exclusively, the point pattern data are a single realization from some underlying point process. In contrast, our motivating data are lesion locations from a cohort of Multiple Sclerosis patients with patient specific covariates measuring disease severity. Patient specific covariates enter the model as a linear combination with spatially varying coefficients. Our goal is to correlate disease severity with lesion location within the brain. Estimation of the LGCP intensity function is typically performed in the Bayesian framework using the Metropolis adjusted Langevin algorithm (MALA) and, more recently, Reimannian manifold Hamiltonian Monte Carlo (RMHMC). Due to the extremely large size of our problem---3D data on 240 subjects with 275,000 voxel locations---we show that MALA performs poorly in terms of posterior sampling and that RMHMC is computationally intractable. As a compromise between these two extremes, we show that posterior estimation via Hamiltonian Monte Carlo performs exceptionally well in terms of speed of convergence and mixing.

e-mail: **tdjtdj@umich.edu**

FIBER DIRECTION ESTIMATION IN DIFFUSION MRI

Raymond Wong*, Iowa State University,
Thomas C. M. Lee, University of California, Davis
Debashis Paul, University of California, Davis
Jie Peng, University of California, Davis

Diffusion magnetic resonance imaging is an emerging medical imaging technology to probe anatomical architectures of biological samples in an in vivo and noninvasive manner. It is widely used

to reconstruct white matter ber tracts in brains. In this talk, we first propose a new parametrization of the tensor mixture model and develop a stable numerical procedure for estimating diffusion direction(s) within a voxel. To further improve the estimation of these directions, we then propose a direction smoothing method which is applicable to regions with crossing bers. In addition, we develop a novel tracking algorithm which takes (estimated) diffusion directions as input and allows for multiple directions within a voxel.

e-mail: **raywong@iastate.edu**

FVGWAS: FAST VOXELWISE GENOME WIDE ASSOCIATION ANALYSIS OF LARGE-SCALE IMAGING GENETIC DATA

Hongtu Zhu*, University of North Carolina, Chapel Hill

Meiyang Chen, University of North Carolina, Chapel Hill

Thomas Nichols, University of Warwick

Chao Huang, University of North Carolina, Chapel Hill

Yu Yang, University of North Carolina, Chapel Hill

Zhaohua Lu, University of North Carolina, Chapel Hill

Qianjing Feng, Southern Medical University

Rebecca C. Knickmeyer, University of North Carolina, Chapel Hill

More and more large-scale imaging genetic studies are being widely conducted to collect a rich set of imaging, genetic, and clinical data to detect putative genes for complexly inherited neuropsychiatric and neurodegenerative disorders. Several major big-data challenges arise from testing genome-wide ($N_G>12$ million known variants) associations with signals at millions of locations ($N_V\sim 10^6$) in the brain from thousands of subjects ($n\sim 10^3$). The aim of this paper is to develop a fast statistical method, referred as FVGWAS, to efficiently carry out whole-genome analyses of whole-brain data. FVGWAS consists of three components including a spatially heteroscedastic linear model, a global sure independence screening (GSIS) procedure, and a detection procedure based on wild bootstrap methods. Specifically, for standard linear association, the computational complexity is $O(n^2N_VN_G)$ for the voxelwise genome wide association (VGWAS) method in \citep{Hibar2011} compared with $O((N_G+N_V)n^2)$ for FVGWAS. Simulation studies show that FVGWAS is an efficient method of searching sparse signals in an extremely large search space, while controlling for the family-wise error rate. Finally, we have successfully applied FVGWAS to a subset of ADNI data with 374 subjects, 195,855 voxels, and 503,892 SNPs, and the total processing time was 3997 seconds for a single CPU.

e-mail: **hzhu@bios.unc.edu**

106. Statistical Methods for Predicting Subgroup Level Treatment Response

A REGRESSION TREE APPROACH TO IDENTIFYING SUBGROUPS WITH DIFFERENTIAL TREATMENT EFFECTS

Wei-Yin Loh*, University of Wisconsin, Madison

Regression trees are natural for subgroup identification because they partition the data space. We introduce two new methods that are practically free of selection bias and are applicable to two or more treatment arms, censored response variables, and missing values in the predictor variables. The methods extend the GUIDE approach by using three key ideas: (i) treatment as a linear predictor, (ii) chi-squared tests to detect residual patterns and lack of fit, and (iii) proportional hazards modeling via Poisson regression. Importance scores with thresholds for identifying influential variables are obtained as by-products.

e-mail: **loh@stat.wisc.edu**

INCREASING EFFICIENCY FOR ESTIMATING TREATMENT-BIOMARKER INTERACTIONS WITH HISTORICAL DATA

Jeremy MG Taylor*, University of Michigan

Philip S. Boonstra, University of Michigan

Bhramar Mukherjee, University of Michigan

In a clinical trial an interaction between the treatment and a baseline biomarker in the outcome model indicates that the treatment effect is not the same for all subjects. Efficiently estimating this treatment-biomarker interaction, in a phase II trial, which has a small sample size, is challenging, but important to inform the design of subsequent phase III studies. Two plausibly-available sources of historical data may contain partial information to help estimate the treatment-biomarker interaction parameter in a randomized phase II study. The historical data is either a study of the control group only with the biomarker measured or a study of both the control and treatment group without the biomarker measured. The parameter is not identified in either historical dataset alone; nonetheless, both can provide some information about the parameter and, consequently, increase the precision of its estimate. To illustrate the potential for gains in efficiency and implications for the design of the study, we consider Gaussian outcomes and biomarker data, posit a linear model, and calculate the asymptotic variance using the expected Fisher information matrix. We find that a non-negligible gain in precision is possible, even if the historical and prospective data do not arise from identical underlying models.

e-mail: **jmgt@umich.edu**

FEATURE ELIMINATION FOR REINFORCEMENT LEARNING METHODS

Sayan Dasgupta*, Fred Hutchinson Cancer Research Center

Michael R. Kosorok, University of North Carolina, Chapel Hill

Personalized medicine can be defined as the medical model that can adapt itself to appropriate needs of a patient, with treatments and medical decisions suited to his/her requirements. Discovering tailored therapies for these patients is a very complex issue because effects of must be modeled within the multistage structure. Recently Q-learning (Watkins 1989; Murphy et. al. 2006) has been proposed for maximizing the average survival time of patients in this format. One important problem that we typically face however, is that the information about prognosis is sometimes very rich, and moreover in Q-learning this prognosis information (history) grows with the number of stages in the trial. Hence, overfitting is an issue that needs to be addressed, and feature elimination becomes an importance tool here and this will be the primary focus for this talk. We will discuss a few different methods for feature selection in Q learning, based on the idea of feature screening through ranking in a sequential backward selection scheme. We will discuss the applicability of the methods, partly reasoned on heuristics stemming from our previous work on feature selection in support vector machines and will give results showing their performance in various simulated settings.

e-mail: **sdg.roopkund@gmail.com**

ADAPTIVE DESIGNS FOR DEVELOPING AND VALIDATING PREDICTIVE BIOMARKERS

Noah Simon, University of Washington

Richard M. Simon*, National Cancer Institute, National Institutes of Health

Clinical trials are generally designed to determine whether a treatment provides average benefit for the eligible population. The average benefit is often small and many patients must be treated for each one who benefits. This approach to evaluating treatments is particularly problematic in oncology where the usual diagnostic categories are molecularly heterogeneous and modern molecularly targeted drugs are unlikely to be broadly useful. In this presentation we will present a new paradigm for phase III clinical trials which we believe is better suited to early 21st century oncology. The new paradigm includes two objectives; testing the null hypothesis of uniform ineffectiveness of the test regimen for the eligible population, and prospective development of a predictive biomarker or biomarker classifier that provides internally validated guidance regarding the subset of patients who are most likely to benefit from the test treatment. The first objective is assured in a frequentist framework whereas the development and validation of the "indication classifier" is viewed as a prediction, classification or decision problem, not as a hypothesis testing problem. We will present an adaptive approach to prospective development and validation of an "indication classifier" in a phase III

or phase II/III trial. The approach is suited for settings where the best predictive biomarker is not established by the start of the study but there are a limited number of candidate markers. We describe several types of indication classifiers including a Bayesian method in an otherwise frequentist randomized clinical trial.

e-mail: **rsimon@nih.gov**

107. CONTRIBUTED PAPERS: ROC Curves

IMPROVED ESTIMATION OF DIAGNOSTIC CUT-OFF POINT ASSOCIATED WITH YOUDEN INDEX USING RANKED SET SAMPLING

Jingjing Yin*, Georgia Southern University

Hani Samawi, Georgia Southern University

Chen Mo, Georgia Southern University

Daniel Linder, Georgia Southern University

Diagnostic cut-off point of biomarker measurements is needed for classifying a random subject to be either diseased or healthy. However, such cut-off point is usually unknown and needs to be estimated by some optimization criteria, among which, Youden index has been widely adopted in practice. Youden index, defined as max (sensitivity + specificity -1), directly measures the largest total diagnostic accuracy a biomarker can achieve. Therefore, it is desirable to estimate the optimal cut-off point associated with Youden index. Sometimes, taking the actual measurements of a biomarker is very difficult and expensive, while ranking them without actual measurements can be easy. In such cases, ranked set sampling would give more accurate estimation than simple random sampling, since ranked set samples are more likely to span the full range of population (thus is more representative). In this study, kernel density estimation is utilized to numerically solve for the nonparametric estimate of the optimal cut-off point. Intensive simulations are carried out to compare the proposed method using ranked set samples with the one using simple random samples and the proposed method outperforms universally with much smaller mean squared error (MSE). A real data set is analyzed for illustrating the proposed method.

e-mail: **jyin@georgiasouthern.edu**

A BETTER CONFIDENCE INTERVAL FOR THE SENSITIVITY AT A FIXED LEVEL OF SPECIFICITY FOR DIAGNOSTIC TESTS WITH CONTINUOUS ENDPOINTS

Guogen Shan*, University of Nevada Las Vegas

For a diagnostic test with continuous measurement, it is often important to construct confidence intervals for the sensitivity at a fixed level of specificity. Bootstrap based confidence intervals were shown to have good performance as compared to others, and the one by Zhou and Qin (2005) was recommended as the best existing confidence interval, named the BTII interval. We propose two new confidence intervals based on the profile variance method, and conduct extensive simulation studies to compare the proposed intervals and the BTII intervals under a wide range of conditions. An example from a medical study on severe head trauma is used to illustrate application of the new intervals. The new proposed intervals generally have better performance than the BTII interval.

e-mail: **guogen.shan@unlv.edu**

SIMPSON'S PARADOX IN THE IDI

Jonathan Chipman*, Vanderbilt University

Danielle Braun, Dana-Farber Cancer Institute

The Integrated Discrimination Improvement (IDI) is a commonly used metric to compare two risk prediction models; it summarizes the extent to which a new model increases risk in cases and decreases risk in controls. The IDI averages risks across cases and controls and is therefore susceptible to Simpson's Paradox. In some settings, adding a predictive covariate to a well calibrated model results in an overall negative IDI. However if we stratify by the covariate, the stratum-specific IDIs are non-negative. Meanwhile, the calibration (O/E), AUC, and Brier Score improve overall and for each stratum. We ran extensive simulations to determine which settings lead to paradoxical IDI results. We provide an analytic explanation and suggest a simple modification to be used in these settings. We illustrate the paradox on Cancer Genomics Network data, by calculating predictions based on two versions of BRCAPRO (version 2.08 and version 2.07), a Mendelian risk prediction model for breast and ovarian cancer. Version 2.08 updates contralateral breast cancer (CBC) penetrance. Calibration

(O/E), AUC, and Brier Score improve overall and by CBC stratum; however, the overall IDI is negative while CBC stratum-specific IDIs are non-negative.

e-mail: **jonathan.chipman@vanderbilt.e**

A NONPARAMETRIC TEST BASED ON t-DISTRIBUTION FOR COMPARING TWO CORRELATED C INDICES WITH RIGHT-CENSORED SURVIVAL OUTCOME OR AUCs WITH DICHOTOMOUS OUTCOME

Le Kang*, Virginia Commonwealth University

Shumei Sun, Virginia Commonwealth University

When the reference standard is binary outcome, the area under the receiver operating characteristic (ROC) curve (AUC) is routinely used as a summary measure of diagnostic accuracy. When the reference standard is right-censored time-to-event (survival) outcome, the C index, motivated as an extension of AUC, provides a measure of concordance between a prognostic biomarker and the right-censored survival outcome. Statistical methods for estimating the C index and its confidence interval, as well as for comparing two or more correlated C indices have been investigated extensively. In this work, we propose a nonparametric test based on t-distribution for comparing two correlated C indices (AUCs as a special case). We adopt U-statistics based estimators, both for the C index and for the variance of the difference between two C indices. We show that the resulting test statistic for comparing two C indices follows an approximate t-distribution and we propose to estimate the appropriate degree of freedom for the t-distribution using the Satterthwaite approximation. We validate our proposed test and particularly assess its performance in parallel with the DeLong test for comparing two correlated AUCs via Monte Carlo simulation studies. Simulation results show that the proposed method provides satisfactory type I error rates, even with very small sample sizes.

e-mail: **lekang@live.com**

LATENT MIXTURE MODELS FOR ORDERED ROC CURVES USING THE SCALE MIXTURE OF NORMAL DISTRIBUTIONS

Zhen Chen*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Sungduk Kim, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

In the Physician Reliability Study (PRS), 12 physicians in OB/GYN were invited to diagnose endometriosis of about 150 participants under several settings, each with a different amount of clinical information. To assess the diagnostic accuracy of the physicians under each setting, care has to be taken to address the between-setting dependence of the study outcome rASRM. Moreover, as the clinical information increase with settings, it is desirable to account for the a priori constraint in the estimation of the diagnostic parameters. In this work, we proposed a latent mixture modeling framework for the rASRM scores, using the class of scale mixture of normal distributions within each disease population. The a priori constraint was specified as decreasing variances of the distributions of the outcome. We developed MCMC procedure to implement model inference from a Bayesian perspective and conducted simulation study to evaluate the performance of the proposed approach. By DIC, the model with generalized t distribution for rASRM fits the PRS data the best. Substantively, we observed higher AUC estimate in setting 2 when the constraint was used while higher AUC estimate in setting 1 when the constraint was not used.

e-mail: **chenzhe@mail.nih.gov**

LEAST SQUARES ROC METHOD FOR TESTS WITH THE ABSENCE OF THE GOLD STANDARD

Larry Tang*, George Mason University and National Institutes of Health Clinical Center

Minh Huynh, Department of Labor and National Institutes of Health Clinical Center

Xuan Che, Epidemiology and Biostatistics, National Institutes of Health Clinical Center

Elizabeth K. Rasch, Epidemiology and Biostatistics, National Institutes of Health Clinical Center

Ao Yuan, Georgetown University

The topics on diagnostic accuracy without the gold standard can be classified into several areas, including 1) binary test results with a perfect gold standard,

2) ordinal or continuous test results with a perfect gold standard, 3) binary test results without a gold standard, and 4) ordinal or continuous test results without a gold standard. Extensive literature is available on parametric, semiparametric and nonparametric methods to evaluate the accuracy of diagnostic tests with perfect gold standards. Sensitivities and specificities are commonly used for a binary test. These parameters can be estimated using the proportions when a perfect gold standard is available for every individual in the sample. The receiver operating characteristic (ROC) curve plotting pairs of sensitivities and specificities is a common statistical tool to evaluate the accuracy of ordinal or continuous tests. The ROC curve estimated from data without the gold standard is biased. To correct for the bias, a linear regression method is proposed to estimate the ROC curve from pairs of consistent sensitivity and specificity estimates. The proposed method first applies Hui and Walter's method to estimate a pair of sensitivity and specificity for a given cutoff point. For a set of chosen cutoff points on the continuous data, a number of pairs can be obtained and the estimates in the pairs can be values for the response variable and covariate in the linear regression setting.

e-mail: t**ang7814@yahoo.com**

108. CONTRIBUTED PAPERS: Personalized Medicine and Biomarkers

USING DECISION LISTS TO CONSTRUCT INTERPRETABLE AND PARSIMONIOUS TREATMENT REGIMES

Yichi Zhang*, North Carolina State University

Eric Laber, North Carolina State University

Anastasios Tsiatis, North Carolina State University

Marie Davidian, North Carolina State University

A treatment regime formalizes personalized medicine as a function from individual patient characteristics to a recommended treatment. A high-quality treatment regime can improve patient outcomes while reducing cost, resource consumption, and treatment burden. Thus, there is tremendous interest in estimating treatment regimes from observational and randomized studies. However, the development of treatment regimes for application in clinical practice requires the long-term, joint effort of statisticians and clinical scientists. In this collaborative process, the statistician must integrate clinical science into the statistical models underlying a treatment regime and the clinician must scrutinize the estimated treatment regime for scientific validity. To facilitate meaningful information exchange, it is important that estimated treatment regimes be interpretable in a subject-matter context. We propose a simple, yet flexible class of treatment regimes whose members are representable as a short list of if-then statements. Regimes in this class are immediately interpretable and are therefore an appealing choice for broad application in practice. We derive a robust estimator of the optimal regime within this class and demonstrate its finite sample performance using simulation experiments. The proposed method is illustrated with data from two clinical trials.

e-mail: **yzhang52@ncsu.edu**

SYNTHESIZING GENETIC MARKERS FOR INCORPORATION INTO CLINICAL RISK PREDICTION TOOLS

Sonja Grill*, Technical University Munich, Germany

Donna P. Ankerst, Technical University Munich, Germany and University of Texas Health Science Center at San Antonio

Clinical risk prediction tools built on standard risk factors are important devices for many different diseases. Newly discovered genetic and high-dimensional-omic markers, such as single nucleotide polymorphisms (SNPs) and gene expressions, have the potential to increase the practical utility of clinical risk prediction tools. Typically these markers are not assessed in the original cohorts used to build the existing risk prediction tools, making their incorporation into those tools complicated. We provide an intuitive Bayesian method for updating an existing clinical risk prediction tool with external marker information via the use of likelihood ratios to transform the prior odds of a disease to posterior odds. We illustrate the method with two applications, the

first incorporating SNPs from multiple published genome-wide association studies into the Prostate Cancer Prevention Trial Risk Calculator via a random-effects meta-analysis with an option accounting for linkage disequilibrium between groups of SNPs. The second application is detailed family history of cancer from the nationwide Swedish Family-Cancer Database (the world's largest of its kind). Both markers are independent predictors of prostate cancer to the commonly-used risk factors.

e-mail: **sonja.grill@tum.de**

A PRIM APPROACH TO PREDICTIVE-SIGNATURE DEVELOPMENT FOR PATIENT STRATIFICATION

Gong Chen*, Roche TCRC, Inc.

Hua Zhong, New York University School of Medicine

Anton Belousov, Roche Diagnostics GmbH

Viswanath Devanarayan, AbbVie, Inc.

Patients often respond differently to a treatment due to individual heterogeneity. Failures of clinical trials can be substantially reduced if, prior to an investigational treatment, patients are stratified into responders and non-responders based on biological or demographic characteristics. These characteristics are captured by a predictive signature. In this talk, we introduce a procedure to search for predictive signatures based on the approach of Patient Rule Induction Method (PRIM). Specifically, we discuss selection of a proper objective function for the search, present its algorithm, and describe a resampling scheme that can enhance search performance. Through simulations, we characterize conditions that enable the procedure to work well. To demonstrate practical uses of the procedure, we apply it to two real-world data sets. We also compare the results with those obtained from a recent regression-based approach, Adaptive Index Models, and discuss their respective advantages. We will be focused on oncology applications with survival responses.

e-mail: **gongchen316@gmail.com**

ON ESTIMATION OF OPTIMAL TREATMENT REGIMES FOR MAXIMIZING T-YEAR SURVIVAL PROBABILITY

Runchao Jiang*, North Carolina State University

Wenbin Lu, North Carolina State University

Rui Song, North Carolina State University

Marie Davidian, North Carolina State University

A treatment regime is a deterministic function that dictates personalized treatment based on patients' individual prognostic information. There is increasing interest in finding optimal treatment regimes, which determine treatment at one or more treatment decision points so as to maximize expected long-term clinical outcome, where larger outcomes are preferred. For chronic diseases such as cancer or HIV infection, survival time is often the outcome of interest, and the goal is to select treatment to maximize survival probability. We propose two nonparametric estimators for the survival function of patients following a given treatment regime involving one or more decisions, i.e., the so-called value. Based on data from a clinical or observational study, we estimate an optimal regime by maximizing these estimators for the value over a prespecified class of regimes. Because the value function is very jagged, we introduce kernel smoothing within the estimator to improve performance. Asymptotic properties of the proposed estimators of value functions are established under suitable regularity conditions, and simulations studies evaluate the finite-sample performance of the proposed regime estimators. The methods are illustrated by application to data from an AIDS clinical trial.

e-mail: **rjiang2@ncsu.edu**

EVALUATION OF NOVEL BIOMARKERS WHEN LIMITED BY SMALL SAMPLE SIZE

Bethany J. Wolf*, Medical University of South Carolina

John Christian Spainhour, Medical University of South Carolina

Jim C. Oates, Medical University of South Carolina

Advances in high-throughput biologic methods provide potential for discovery of biomarkers predictive of disease status. Several difficulties in identifying predictive biomarkers include small sample size, weak main effects, marker interactions, and non-linear relationships

between markers and outcomes. Logistic regression is a common approach for modeling binary disease outcomes and while logistic regression can model weak main effects and interactions, it suffers from poor precision of regression estimates, model over-fitting, and failure to meet underlying assumptions. Machine learning methods have many desirable features for evaluating novel biomarkers for prediction of disease outcomes and require fewer assumptions than logistic regression. However these methods may also over fit the data and thus do not necessarily validate well in new data. There is not one statistical method that can provide a "best" model for all data, particularly with small sample size. Thus it is beneficial to evaluate and compare the predictive performance of multiple statistical models when examining biomarkers. We present a strategy for evaluating the predictive capability of a set of biomarkers using different statistical models. We apply this strategy to evaluate prediction performance of a set of novel urine biomarkers of treatment response in patients with lupus nephritis.

e-mail: **wolfb@musc.edu**

CALIBRATE VARIATIONS IN BIOMARKER MEASURES FOR IMPROVING PREDICTION

Cheng Zheng*, University of Wisconsin, Milwaukee

Yingye Zheng, Fred Hutchinson Cancer Research Center

Novel biologic markers have been widely used in predicting important clinical outcome. One specific feature of biomarkers is that they often are ascertained with variations due to the specific process of measurement. The magnitude of such variation may differ when applied to a different targeted population or when the platform for biomarker assaying changes from original platform the prediction algorithm (cutoffs) based upon. Statistical methods have been proposed to characterize the effects of underlying error-free quantity in association with an outcome, yet the impact of measurement errors in terms of prediction has not been well studied. We focus in this manuscript on the settings where biomarkers are used for predicting individual's future risk and propose semiparametric estimators for error-corrected risk, when replicates of the error-prone biomarkers are available. The predictive performance of the proposed estimators is evaluated and compared to alternative approaches with numerical studies under settings with various assumptions on the measurement distributions. We studied the asymptotic properties of the proposed estimator. Application is made in a liver cancer biomarker study to predict risk of 3 and 4 years liver cancer incidence using age and a novel biomarker.

e-mail: **zhengc@uwm.edu**

BUILDING SMALL, ROBUST GENE SIGNATURES TO PREDICT PROGNOSIS

Prasad Patil*, Johns Hopkins University

Jeffrey T. Leek, Johns Hopkins University

We describe a novel approach to building lightweight, robust, and interpretable gene signatures for prediction of prognosis in cancer patients. A bottleneck to building gene signatures is that the feature space of all possible genes is extremely large and noisy. This space is commonly reduced by incorporating knowledge about gene function and regulation to weed out biologically implausible genes, but this may discard genes that offer predictive value. Feature selection methods for microarrays can struggle with the size of the feature set and often incorporate the outcome throughout, which may lead to overfitting. In this work, we focus on pairwise comparisons between genes. These features are robust to the technology used to measure gene expression, and signatures built using these features do not require retraining across platforms and technologies. We present a fast, two-stage filter/wrapper method that can reduce 20,000+ genes to a handful of pairwise comparisons. This method relies on unique properties of predictive pairwise features and on the equivalency of F-statistics when outcome and covariate are flipped in a regression. We then compare gene signatures created using our method to leading signatures that have been validated for use in the clinic for predicting prognosis of breast cancer patients.

e-mail: **ppatil8@jhu.edu**

109. CONTRIBUTED PAPERS: Time Series Analysis and Methods

ROBUST PORTFOLIO OPTIMIZATION UNDER HIGH DIMENSIONAL HEAVY-TAILED TIME SERIES

Huitong Qiu*, Johns Hopkins University

Fang Han, Johns Hopkins University

Han Liu, Princeton University

Brian Caffo, Johns Hopkins University

In this paper, we study a robust portfolio optimization strategy by resorting to quantile-based statistics. Computationally, the method is as efficient as its Gaussian-based alternative. Theoretically, by exploiting the quantile-based statistics, we show that the actual portfolio risk approximates the oracle risk with parametric rate of convergence. The rate is set in a double asymptotic framework where the portfolio size may scale exponentially with the sample size. Moreover, the theory holds under heavy tailed distributions with no moment constraints, and allows for weakly dependent time series. The empirical effectiveness of the method is demonstrated in both synthetic and real data. The experiments demonstrate that the method can significantly stabilize portfolio risk under highly volatile stock returns, and effectively avoid extreme losses.

e-mail: **qht19881226@gmail.com**

CHANGE-POINT DETECTION IN EEG SPECTRA FOR INFORMED FREQUENCY BAND SELECTION

Anna Louise Schroeder*, London School of Economics

Hernando Ombao, University of California, Irvine

The analysis of neural activity in a brain when exposed to an external stimulus is core many neuroscientific research questions, e.g. on Brain-Computer Interfaces or developmental disorders such as dyslexia. In clinical settings applications exist e.g. for the diagnosis of brain diseases, head injuries and sleep disorders. Electroencephalograms measure electrical activity non-invasively and with high temporal resolution. In experiments, data is recorded over multiple trials and at many points distributed over the skull. It is commonly analysed in the spectral domain, where temporal evolution of pre-defined, broad frequency bands are monitored. The a priori definition of frequency bands originated from the analysis of key surface features, such as the average peak frequency. It has been shown that the highest-energy frequency within a band differs from individual to individual and can be related to e.g. age, performance and intelligence. Subject-independent applications therefore typically consider the mean power spectral density and thus risk averaging-out possibly pronounced local changes in power. To avoid this, they require a mechanism to identify most informative frequencies and compare this accounting for individual differences. We present a novel method to detect change points over time in frequencies. Based on these change points we can identify the most informative frequencies which may be located in different points within a frequency band. Our approach takes the high dimensionality of the data over channels into account. Furthermore, we analyse the information content over trial repetitions and subjects.

e-mail: **a.m.schroeder@lse.ac.uk**

TIME SERIES ANALYSIS FOR SYMBOLIC-VALUED DATA

S. Yaser Samadi*, Southern Illinois University

Lynne Billard, University of Georgia

Symbolic values can be lists, intervals, frequency distributions, and so on. Therefore, in comparison with standard classical data, they are more complex and can have structures (especially internal structures) that impose complications that are not evident in classical data. In general, using "classical" analysis approaches directly lead to inaccurate results. As a result of dependency in time series observations, it is more difficult to deal with symbolic (interval) time series data and take into account their complex structure and internal variability. In the literature, the proposed procedures for analyzing interval time series data used either midpoint or radius that are inappropriate surrogates for symbolic interval variables. We develop a theory and methodology to analyze symbolic time series data (interval data) directly. Autocorrelation and partial autocorrelation functions are formulated, maximum likelihood estimators of the parameters of symbolic autoregressive processes are provided.

e-mail: **s.y.samadi@gmail.com**

HIGH DIMENSIONAL STATE SPACE MODEL WITH L-1 AND L-2 PENALTIES

Shaojie Chen*, Johns Hopkins University

Joshua Vogelstein, Johns Hopkins University

Seonjoo Lee, Columbia University

Martin Lindquist, Johns Hopkins University

Brian Caffo, Johns Hopkins University

The time-invariant state space model, also known as the linear dynamical system (LDS) model or linear Gaussian model (LGM), is widely used in time series analysis. A broad class of popular models including factor analysis, principal component analysis (PCA) and independent component analysis (ICA) could be unified as variations of this generative model. Parameters learning in this model is challenging, especially when the dimension is high. In this paper, we generalized the model by penalizing the coefficient matrices with L-1 and L-2 penalties. An Expectation-Maximization algorithm is then designed for parameter learning. At the end the model is applied to explore the motor cortex of human brains.

e-mail: **pzcsj76@gmail.com**

AUTOREGRESSIVE MODELS FOR SPHERICAL DATA WITH APPLICATIONS IN PROTEIN STRUCTURE ANALYSIS

Daniel Hernandez-Stumpfhauser*, University of North Carolina, Chapel Hill

F. Jay Breidt, Colorado State University

Mark van der Woerd, Colorado State University

Proteins consist of sequences of the 21 natural amino acids. There can be tens to hundreds of amino acids in the protein, and hundreds to hundreds of thousands of atoms. A complete model for the protein consists of coordinates for every atom. A class of simplified models is obtained by focusing only on the alpha -carbon sequence, consisting of the primary carbon atom in the backbone of each amino acid. The three-dimensional structure of the alpha -carbon backbone of the protein can be described as a sequence of angle pairs, each consisting of a bond angle and a dihedral angle. These angle pairs lie naturally on a sphere. We consider autoregressive time series models for such spherical data sequences, using extensions of projected normal distributions. Application to protein data and further developments, including regime-switching autoregressive models, are described. This is joint work with F. Jay Breidt, Department of Statistics, Colorado State University, and Mark van der Woerd, Department of Biochemistry and Molecular Biology, Colorado State University.

e-mail: **danielhs@live.unc.edu**

MODELING SERIAL COVARIANCE STRUCTURE IN SEMIPARAMETRIC LINEAR MIXED-EFFECTS REGRESSION FOR LONGITUDINAL DATA

Changming Xia*, University of Rochester Medical Center

Hua Liang, The George Washington University

Sally W. Thurston, University of Rochester Medical Center

Mixed-effects regression accounts for correlation and overdispersion in longitudinal data by introducing random effects of subjects. Any further unexplained correlation and variance structure, such as autoregressive time series and exponential weights, are accounted for by serial covariance structure within subjects after conditioning on random effects. We evaluate the effects of serial covariance structure mis-specification on model fitting and hypothesis testing in semiparametric linear mixed-effects regression for dependent continuous and categorical outcomes fitted by smoothing splines based on reproducing kernel Hilbert space.

e-mail: **c.xia@rochester.edu**

110. Incorporating Biological Information in Statistical Modeling of Genome-Scale Data with Complex Structures

PRIORITIZING GWAS RESULTS BY INTEGRATING PLEIOTROPY AND ANNOTATION

Hongyu Zhao*, Yale School of Public Health

Dongjun Chung, Medical University of South Carolina

Can Yang, Hong Kong Baptist University

Cong Li, Yale University

Qian Wang, Yale University

Joel Gelernter, Yale School of Medicine

Results from Genome-Wide Association Studies (GWAS) have shown that complex diseases are often affected by many genetic variants with small or moderate effects. Identifications of these risk variants remain a very challenging problem. There is a need to develop more powerful statistical methods to leverage available information to improve upon traditional approaches that focus on a single GWAS dataset without incorporating additional data. In this presentation, we will introduce a novel statistical approach, GPA (Genetic analysis incorporating Pleiotropy and Annotation), to increase statistical power to identify disease associated variants because: (1) accumulating evidence suggests that different complex diseases share common risk bases, i.e., pleiotropy; and (2) functionally annotated variants have been consistently demonstrated to be enriched among GWAS hits. GPA performs an integrative analysis of multiple GWAS datasets and functional annotations to seek association signals, as well as hypothesis testing to test the presence of pleiotropy and enrichment of functional annotation. When we applied GPA to analyze jointly five psychiatric disorders with annotation information, not only did GPA identify many weak signals missed by the traditional single phenotype analysis, but it also revealed relationships in the genetic architecture of these disorders. We will also demonstrate the usefulness of GPA using several other examples. This is joint work with Dongjun Chung, Can Yang, Cong Li, Qian Wang, and Joel Gelernter.

e-mail: **hongyu.zhao@yale.edu**

CHALLENGES AND SOLUTIONS FOR WHOLE EXOME SEQUENCE ANALYSIS FOR PEDIGREE AND EXTERNAL CONTROL DATA

Daniel J. Schaid*, Mayo Clinic

Whole exome sequencing (WES) targets protein-coding DNA sequences, a technique we have used to screen for genes associated with familial prostate cancer. The study samples are pedigree members with prostate cancer selected from the International Collaboration of Prostate Cancer Genetics. For cost efficiency, unrelated external controls with WES from prior studies were used. Statistical challenges of analyzing pedigree data with external controls, and proposed solutions, will be presented. Topics include evaluating quality control metrics and comparability of WES data between cases and controls, accounting for pedigree relationships in association analyses, and incorporating biological annotation to weight variants. We will present new statistical methods for case-control comparisons with related subjects, as well as statistical tests for co-segregation of genetic variants with disease, allowing for gene-level analyses that evaluate multiple genetic variants within a gene.

e-mail: **schaid.daniel@mayo.edu**

BIG DATA METHODS FOR DISSECTING VARIATIONS IN HIGH-THROUGHPUT GENOMIC DATA

Fang Du, Johns Hopkins Bloomberg School of Public Health

Bing He, Johns Hopkins Bloomberg School of Public Health

Hongkai Ji*, Johns Hopkins Bloomberg School of Public Health

Variance decomposition (e.g., ANOVA, PCA) is a fundamental tool in statistics to understand data structure. High-throughput genomic data have heterogeneous sources of variation. Some are of biological interest, and others are unwanted (e.g., lab and batch effects). Knowing the relative contribution of each source to the total data variance is crucial for making data-driven discoveries. However, when one has massive amounts of high-dimensional data with heterogeneous origins, analyzing variances is non-trivial. The dimension, size and heterogeneity of the data all pose significant challenges. Big Data Variance Decomposition (BDVD) is a new tool developed to solve this problem. Built upon the recently developed RUV approach, BDVD decomposes data into

biological signals, unwanted systematic variation, and independent random noise. The biological signals can then be further decomposed to study variations among genomic loci or sample types, or correlation between different data types. The algorithm is implemented by incorporating techniques to handle big data. Applying BDVD to ENCODE, we show the variance structure of the ENCODE DNase-seq data and demonstrate that BDVD allows one to develop tools that better separate signals from noise in various applications.

e-mail: **hji@jhsph.edu**

MODEL-BASED APPROACH FOR SPECIES QUANTIFICATION AND DIFFERENTIAL ABUNDANCE ANALYSIS BASED ON SHOTGUN METAGENOMIC DATA

Hongzhe Li*, University of Pennsylvania

The human microbiome, which includes the collective microbial genomes residing in or on the human body, has a profound influence on human health. The DNA sequencing technologies have made large-scale the human microbiome studies possible by using shotgun metagenomic sequencing. It is of great interest to quantify the bacterial abundances based on the sequencing data and to identify the bacteria that are associated with clinical outcomes. We propose a hierarchical Poisson-Gamma regression model and its Empirical Bayes extension to quantify microbial abundances based on species-specific taxonomic markers as well as to identify the covariate-associated bacteria. Our model takes into account the marker-specific effect when normalizing the sequencing count data. Compared to currently available methods on simulated data and real data, our method has demonstrated an improved accuracy in bacterial abundance quantification and better sensitivity and specificity in identifying the covariate-associated species.

e-mail: **hongzhe@upenn.edu**

111. Emerging Issues in Clinical Trials and High Dimensional Data

ASSESSING COVARIATE EFFECTS WITH THE MONOTONE PARTIAL LIKELIHOOD USING JEFFREYS' PRIOR IN THE COX MODEL

Ming-Hui Chen*, University of Connecticut

Mario de Castro, Universidade de Sao Paulo

Jing Wu, University of Connecticut

Elizabeth D. Schifano, University of Connecticut

In clinical trials, the monotone partial likelihood is frequently encountered in the analysis of time-to-event data using the Cox model. When there are zero events in one or more covariate groups, the resulting partial likelihood is monotonic and consequently, the covariate effects are difficult to estimate. In this paper, we develop both Bayesian and frequentist approaches using the Jeffreys' prior to handle the monotone partial likelihood problem. We characterize sufficient and necessary conditions for the propriety of the Jeffreys' prior. We also show that the Jeffreys' prior has finite modes. A modification of the Jeffreys' prior is proposed in order to obtain more robust estimates of covariate effects. We perform extensive simulations to examine the performance of parameter estimates and demonstrate the applicability of our methods by analyzing real data from cancer clinical trials in detail.

e-mail: **ming-hui.chen@uconn.edu**

ASSESSING TEMPORAL AGREEMENT BETWEEN CENTRAL AND LOCAL PROGRESSION-FREE SURVIVAL TIMES

Donglin Zeng*, University of North Carolina, Chapel Hill

Emil Cornea, University of North Carolina, Chapel Hill

Jun Dong, Amgen Inc.

Jean Pan, Amgen Inc.

Joseph Ibrahim, University of North Carolina, Chapel Hill

In oncology clinical trials, progression-free survival (PFS), generally defined as the time from randomization until disease progression (PD) or death, has been a key endpoint to support licensing approval. When PFS is the primary or co-primary endpoint, it is recommended to have tumor assessments verified by an independent review committee (IRC) blinded to study treatments, especially in open-label studies. It is considered reassuring about the lack of reader-evaluation bias if treatment effect estimates from the investigator and IRC's evaluations agree. Agreement between these evaluations may vary for subjects with short or long PFS, while there exist no such statistical

quantities that can completely account for this temporal pattern of agreements. Therefore, in this paper, we propose a new method to assess temporal agreement between two time-to-event endpoints, while the two event times are assumed to have a positive probability of being identical. This method measures agreement in terms of the two event times being identical at a given time or both being greater than a given time. Overall scores of agreement over a period of time are also proposed. We propose maximum likelihood estimation to infer the proposed agreement measures using empirical data, accounting for different censoring mechanisms including reader's censoring. The proposed method is demonstrated to perform well in small-sample via extensive simulation studies and is illustrated through a head and neck cancer trial.

e-mail: **dzeng@email.unc.edu**

STATISTICAL DESIGN OF NON-INFERIORITY MULTIPLE REGION CLINICAL TRIALS TO ASSESS GLOBAL AND CONSISTENT TREATMENT EFFECTS

Guoqing Diao*, George Mason University

Donglin Zeng, University of North Carolina, Chapel Hill

Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Alan Rong, Amgen Inc.

Oliver Lee, Amgen Inc.

Kathy Zhang, Amgen Inc.

Qingxia Chen, Vanderbilt University

Non-inferiority multi-regional clinical trials (MCRTs) have recently received increasing attention in drug development. Two major goals in a MCRT are (1) to estimate the global drug effect and (2) to assess the consistency of drug effects across multiple regions. In this paper, we propose an intuitive definition of consistency of non-inferior drug effects across regions under the random effects modeling framework. Specifically, we quantify the consistency of drug effects by the percentage of regions that meet a pre-defined treatment margin. This new approach enables us to achieve both goals in one modeling framework. We propose to use a signed likelihood ratio test for testing the global drug effect and the consistency of non-inferior drug effects. In addition, we provide guidelines for the allocation rule to achieve optimal power for testing consistency among multiple regions. Extensive simulation studies are conducted to examine the performance of the proposed methodology. An application to a real data example is provided.

e-mail: **gdiao@gmu.edu**

BAYESIAN SHRINKAGE METHODS FOR HIGH DIMENSIONAL DATA

Joseph G. Ibrahim*, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

Zakaria Khondker, Medivation, Inc.

Zhaohua Lu, University of North Carolina, Chapel Hill

Big data presents the overwhelming challenge of estimating a large number of parameters, which is much larger than the sample size. Even for a simple linear model, when the number of predictors is larger than or close to the sample size, such model may be unidentifiable and the least squares estimates of regression coefficients can be unstable. To deal with such issue, we systematically investigate three Bayesian regularization methods with applications in imaging genetics. First, we develop a Bayesian lasso estimator for the covariance matrix and propose a metropolis-based sampling scheme. This development is motivated by functional network exploration for the entire brain from magnetic resonance imaging (MRI) data. Second, we propose a Bayesian generalized low rank regression model (GLRR) for the mean parameter estimation and combine this with factor loading method of covariance estimation to capture the spatial correlation among the responses and jointly estimate the mean and covariance parameters. This development is motivated by performing genome-wide searches for associations between genetic variants and brain imaging phenotypes from data collected by Alzheimer's Disease Neuroimaging Initiative (ADNI). Third, we extend GLRR to longitudinal setting and propose a Bayesian longitudinal low rank regression (L2R2) to account for spatiotemporal correlation among the responses as well as estimation of full-rank coefficient matrix for standard prognostic factors. This development is motivated by genome-wide searches for associations between genetic variants and brain imaging phenotypes observed over time with a primary focus on role of aging and the interaction of age with genotype in affecting brain volume.

e-mail: **ibrahim@bios.unc.edu**

112. Advances in Repeated Measures and Longitudinal Data Analysis

JOINT MODELLING OF DIFFERENT TYPES OF LONGITUDINAL DATA WITH OUTLIERS AND CENSORING

Lang Wu*, University of British Columbia

In multivariate mixed effects models for longitudinal data, the response variables may be of different types, such as continuous and discrete. Moreover, the data may contain outliers, missing values, and censoring. We provide several methods for inference, addressing these data complications. The methods will be illustrated by AIDS datasets and will be evaluated by simulations.

e-mail: **lang@stat.ubc.ca**

A HIDDEN MARKOV MODEL FOR NON-IGNORABLE NON-MONOTONE MISSING LONGITUDINAL DATA FOR MEDICAL STUDIES OF QUALITY OF LIFE

Kaijun Liao, Hisun Pharmaceuticals USA

Qiang Zhang, Radiation Therapy Oncology Group

Andrea B. Troxel*, University of Pennsylvania Perelman School of Medicine

In longitudinal studies, the problem of non-ignorable and non-monotone missing data has gained increasing attention recently. The statistical approach depends on the factorization of the joint likelihood of the data and the missingness mechanism. In this article we adopt a latent process approach for the analysis of longitudinal data with non-ignorable and non-monotone missingness. Hidden Markov models are widely used for applications in pattern reognition including speech recognition, handwriting, bioinformatics, and gene finding and profiling. Multi-state Markov models are widely used to model diease progression and cancer screening. The hidden Markov model is a powerful extension of the multi-state Markov model in longitudinal studies assuming the states are unobserved. Incorporating this approach with selection models and shared parameter models, we can identify differences among disease processes with incomplete data simultaneously in both the state-dependent model and missingness mechanism model. We propose the models in a generalized linear model and generalized linear mixed model framework, using a backward-forward algorithm to provide efficient parameter estimation in the general situation of non-ignorable non-monotone longitudinal missing data. A two-stage pseudo-likelihood method is used to reduce the parameter space to make this model more attractive. We illustrate the approach using data from a clinical trial in brain cancer.

e-mail: **atroxel@mail.med.upenn.edu**

INVERSE WEIGHTED ESTIMATING EQUATIONS FOR REPEATED MEASURES IN TRANFUSION MEDICINE

Richard Cook*, University of Waterloo

Trials in transfusion medicine are routinely designed to assess the effect of experimental platelet products on patients' platelet counts. In such trials patients may receive several transfusions over a period of time, and a response is available from each such administration. It is natural to consider testing for treatment effects based on standard methods for repeated measures data, but naive analyses of the multiple responses can yield biased estimates of the probability of response and associated treatment effects. These biases arise because only subsets of the patients randomized contribute responses to second and subsequent administrations of therapy, and hence the balance between treatment groups is lost with respect to potential confounding factors. We discuss analysis issues in this setting and demonstrate how biases can be reduced by use of inverse probability weighted estimating equations.

e-mail: **rjcook@uwaterloo.ca**

JOINT MODELLING OF NONIGNORABLE MISSING LONGITUDINAL OUTCOMES AND TIME-TO-EVENT DATA

Sanjoy Sinha*, Carleton University

Joint models for longitudinal and time-to-event data has received considerable attention in recent years for analyzing follow-up data. These are typically used when the focus is on survival data and one wishes to investigate the effect of an endogenous time-dependent covariate on the survival times. Often we encounter missing values in the longitudinal data due to a stochastic missing data mechanism. In this work, we investigate methods for jointly analyzing longitudinal and time-to-event data in the presence of nonignorable and nonmonotone miss-

ing responses. We perform sensitivity analyses to study effects of misspecified missing data models as well as random effects distributions on the estimates of the model parameters. The methods will be evaluated using simulations. An application will be presented using actual data from a clinical study.

e-mail: **sinha@math.carleton.ca**

113. Advances in Modeling Zero-Inflated Data

BAYESIAN TWO-PART SPATIAL MODELS FOR SEMICONTINUOUS DATA

Brian Neelon*, Duke University

Li Zhu, University of Pittsburgh

Sara Benjamin, Duke University

In health services research, it is common to encounter semicontinuous data characterized by a point mass at zero and a continuous distribution of positive values. Examples include medical expenditures, in which the zeros represent patients who do not use health services, while the continuous distribution describes the level of expenditures among users. Semicontinuous data are customarily analyzed using two-part mixture models consisting of a Bernoulli distribution for the probability of a nonzero response and a continuous distribution for the positive responses. In the spatial analysis of semicontinuous data, two-part models are especially appealing because they provide a joint picture of how health services utilization and associated expenditures vary across geographic regions. However, when applying these models, careful attention must be paid to distributional choices, as model misspecification can lead to biased and imprecise inferences. This paper introduces a broad class of Bayesian two-part models for the spatial analysis of semicontinuous data. Specific models considered include two-part lognormal, log skew-elliptical, and Bayesian nonparametric models. Multivariate conditionally autoregressive priors are used to link the binary and continuous components and provide spatial smoothing across neighboring regions, resulting in a joint spatial modeling framework for health utilization and expenditures. We develop a fully conjugate Gibbs sampling scheme, leading to efficient posterior computation. We illustrate the approach using data from a recent study of emergency department expenditures.</p>

e-mail: **brian.neelon@duke.edu**

ZERO-INFLATED FRAILTY MODEL FOR RECURRENT EVENT DATA

Lei Liu*, Northwestern University

Xuelin Huang, University of Texas MD Anderson Cancer Center

Alex Yaroshinsky, Vital Systems Inc.

Recurrent event data arise frequently in longitudinal medical studies. In many situations, there are a large portion of subjects without any recurrent events (e.g., tumor recurrences), manifesting the "zero-inflated" nature of the data. Some of the zero events may be due to "cure", while others are due to censoring before any recurrent events. In this paper, we propose a zero-inflated frailty model for this type of data, combining a logistic model for "cure" status (Yes/No) and a frailty proportional hazards model for recurrent event times of those "not cured". The model can be fitted conveniently in SAS Proc NLMIXED. Simulation results show the satisfactory finite sample property of the estimation method. We apply the method to model tumor recurrences in a soft tissue sarcoma study. We find that this model has a better performance than the frailty model alone.

e-mail: **lei.liu@northwestern.edu**

TWO-PART MODELS FOR ROLLING ADMISSION GROUP THERAPY DATA

Lane F. Burgette*, RAND Corporation

Susan M. Paddock, RAND Corporation

Group therapy is a common treatment modality in alcohol and other drug (AOD) treatment programs, wherein multiple clients attend group therapy sessions together. Clients are often admitted into therapy groups on a rolling basis. The analysis of data arising from such studies is complicated by clustering of client outcomes due to joint participation in group therapy sessions. Outcomes are correlated not only for clients attending common sessions but also for clients attending different sessions that are offered as part of the same rolling group. Conditional autoregression has been used to model the correlation of client outcomes that is due to common therapy session attendance among clients, while allowing for the non-independence of random effects for sessions within the same rolling group. Whereas previous research in the area has focused on continuous measures, many AOD treatment outcomes are two-part in nature. For

example, some clients might report no AOD use following treatment while others report some level of AOD use. For two-part outcomes, we model correlations for both parts of the two-part outcome. We propose vector autoregressive and G-Wishart priors to account for correlations in the random effects distribution while taking advantage of the structure of the group therapy design itself.

e-mail: **burgette@rand.org**

A MARGINALIZED TWO-PART MODEL FOR SEMICONTINUOUS DATA

Valerie A. Smith*, Center for Health Services Research in Primary Care, Durham VAMC and University of North Carolina, Chapel Hill

John S. Preisser, University of North Carolina, Chapel Hill

Brian Neelon, Duke University

Matthew L. Maciejewski, Center for Health Services Research in Primary Care, Durham VAMC

In health services research, it is common to encounter semicontinuous data characterized by a degenerate distribution at zero followed by a right-skewed continuous distribution with positive support. Semicontinuous data are typically analyzed using two-part mixtures that separately model the probability of health services use and the distribution of positive responses among users. However, because the second part conditions on a nonzero response, conventional two-part models do not provide a marginal interpretation of covariate effects on the overall population of health service users and non-users, even though this is often of greatest interest to investigators. We propose a marginalized two-part model that yields more interpretable effect estimates by parameterizing the model in terms of the marginal mean. This model maintains many of the important features of conventional two-part models, such as capturing zero-inflation and skewness, but allows investigators to examine covariate effects on the overall marginal mean, a target often of primary interest. Using a simulation study, we examine properties of maximum likelihood estimates from this model. We illustrate the approach by evaluating the effect of a behavioral weight loss intervention on health care expenditures in the Veterans Affairs health care system. Extensions to longitudinal and clustered data are also considered.

e-mail: **vasmith@email.unc.edu**

114. New Developments in Missing Data Analysis: From Theory to Practice

COMPETING RISKS REGRESSION WITH MISSING DATA IN THE PROGNOSTIC FACTORS

Federico Ambrogi*, University of Milan

Thomas H. Scheike, University of Copenhagen

For the medical studies involving competing risks, one often wishes to estimate and model the cumulative incidence probability, the marginal probability of failure for a specific cause. Recently, several new methods have been developed to directly model the cumulative incidence probability of a specific cause of failure. The key issue here is how to deal with incomplete data due to the fact that observations are subject to right-censoring. We refer to a simple problem in which one covariate, say Z, is always observed and the other, say X, is sometimes missing. There has been considerable focus on handling missing covariates and there are several suggestions for dealing with the simpler survival data where there are not several causes of death. For survival data the key suggestions are multiple imputation techniques that typically aim for the modeling of the hazard function. An alternative is the IPCW techniques for survival data. Even though the competing risks framework is very common practice, there are no studies dealing with the problem of missing covariate information in competing risks regression. Here we present some results regarding multiple imputation and IPCW techniques applied to the direct binomial regression model through some simple simulations.

e-mail: **ambrogifederico@gmail.com**

COMPARISON OF MULTIPLE IMPUTATION VIA CHAINED EQUATIONS AND GENERAL LOCATION MODEL FOR ACCELERATED FAILURE TIME MODELS WITH MISSING COVARIATES

Lihong Qi*, University of California, Davis

Yulei He, Centers for Disease Control and Prevention

Rongqi Chen, University of California, Davis

Ying-Fang Wang, University of California, Davis

Xiaowei Yang, University of California, Davis

Missing covariates are common in biomedical studies with survival outcomes. Multiple imputation is a practical strategy for handling this problem with various approaches and software packages available for implementation. In this talk, we compare two important approaches: multiple imputation by chained equation (MICE) and multiple imputation via a general location model (GLM) for accelerated failure time (AFT) models with missing covariates. Through a comprehensive simulation study, we investigate the performance of the two approaches and their robustness toward violation of the GLM assumptions and model misspecifications including misspecifications of the covariance structure and of the joint distribution of continuous covariates. Simulation results show that MICE can be sensitive to model misspecifications and may generate biased results with inflated standard errors while GLM can still yield estimates with reasonable biases and coverages in these situations. MICE is flexible to use but lack of a clear theoretical rationale and suffers from potential incompatibility of the conditional regression models used in imputation. In contrast, GLM is theoretically sound and can be rather robust toward model misspecifications and violations of GLM assumptions. Therefore, we believe that GLM shows the potential for being a competitive and attractive tool for tackling the analysis of AFT models with missing covariates.

e-mail: **lhqi@ucdavis.edu**

THE EFFECT OF DATA CLUSTERING ON THE MULTIPLE IMPUTATION VARIANCE ESTIMATOR

Yulei He*, Centers for Disease Control and Prevention

Iris Shimizu, Centers for Disease Control and Prevention

Susan Schappert, Centers for Disease Control and Prevention

Nathaniel Schenker, Centers for Disease Control and Prevention

Vladislav Beresovsky, Centers for Disease Control and Prevention

Diba Khan, Centers for Disease Control and Prevention

Roberto Valverde, Centers for Disease Control and Prevention

Multiple imputation is a popular approach to statistical analysis with missing data. Although it was originally motivated by survey nonresponse problems, it has been readily applied to other data settings. On the other hand, its general behavior still remains unclear when applied to survey data with complex sample designs including unequal weighting and clustering. Recently, Lewis et al. (2014) compared single and multiple imputation analyses for certain incomplete variables in the 2008 National Ambulatory Medicare Care Survey, which has a nationally representative, multistage, and clustered sample design. Their study results suggested that the increase of the variance estimate due to multiple imputation compared with single imputation largely disappears for estimates with large design effects. We supplement their research by providing a theoretical explanation for this phenomenon. We consider data sampled from an equally weighted, single-stage cluster design and characterize the process using a balanced, one-way normal random-effects model. Assuming that the missingness is completely at random, we derive the analytic expressions of the within and between- multiple imputation variance estimators for the mean estimator and propose an approximation for the fraction of missing information. As hypothesized by Lewis et al. (2014), we show that rate of missingness and intracluster-correlation (i.e., design effect) have opposite effects on the increase of the variance estimate due to multiple imputation. We discuss some generalizations of this research and its practical implications for data release by statistical agencies.

e-mail: **wdq7@cdc.gov**

FRACTIONAL HOT DECK IMPUTATION FOR MULTIVARIATE MISSING DATA IN SURVEY SAMPLING

Jae kwang Kim*, Iowa State University

Wayne A. Fuller, Iowa State University

Hot deck imputation is popular for handling item nonresponse in survey sampling. Fractional hot deck imputation is extended to multivariate missing data. The joint distribution of the study items are nonparametrically estimated using a discrete approximation. The discrete transformation serves to create imputation cells. The fractional imputation procedure first assigns cells to each missing item and then imputes the real observations within each imputed cell. Replication variance estimation is discussed and results from a limited simulation study presented.

e-mail: **jkim@iastate.edu**

115. Environmental Methods with Deterministic and Stochastic Components

HIGH RESOLUTION NONSTATIONARY RANDOM FIELD SIMULATION

William Kleiber*, University of Colorado, Boulder

Stochastic weather generators (SWGs) are used in many scientific studies, including model downscaling, climate impact assessments and seasonal resource planning. The fundamental requirement of a stochastic weather generator is simulated realizations of plausible weather patterns. In recent years, focus has shifted to developing spatially-consistent SWGs. Unless the region of interest is relatively small, or has homogeneous topography, the SWG will necessarily require simulation of nonstationary spatial fields. However, high resolution simulation of a nonstationary process is difficult, typically requiring a Cholesky decomposition of a matrix whose dimension equals that of the desired simulation resolution. We introduce an approach to large, high resolution nonstationary process simulation by exploiting ideas very similar to Sampson and Guttorp (1992), relying on spatially deforming geographical space to achieve approximate stationarity, then using fast stationary simulation algorithms, followed by an inverse transformation back to the nonstationary plane. We illustrate the algorithm on simulated and real datasets.

e-mail: **william.kleiber@colorado.edu**

ESTIMATING PARAMETERS IN DELAY DIFFERENTIAL EQUATION MODELS

Liangliang Wang*, Simon Fraser University

Jiguo Cao, Simon Fraser University

Delay differential equations (DDEs) are widely used in ecology, physiology and many other areas of applied science. Although the form of the DDE model is usually proposed based on scientific understanding of the dynamic system, parameters in the DDE model are often unknown. Thus it is of great interest to estimate DDE parameters from noisy data. Since the DDE model does not usually have an analytic solution, and the numeric solution requires knowing the history of the dynamic process, the traditional likelihood method cannot be directly applied. We propose a semiparametric method to estimate DDE parameters. The key feature of the semiparametric method is the use of a flexible nonparametric function to represent the dynamic process. The nonparametric function is estimated by maximizing the DDE-defined penalized likelihood function. Simulation studies show that the semiparametric method gives satisfactory estimates of DDE parameters. The semiparametric method is demonstrated by estimating a DDE model from Nicholson's blowfly population data.

e-mail: **lwa68@sfu.ca**

ZERO-INFLATED SPATIAL TEMPORAL MODELS FOR EXPLORING TREND IN COMANDRA BLISTER RUST INFECTION IN LODGE POLE PINE TREES

Cindy Feng*, University of Saskatchewan

Environmental and ecological counts data are often characterized by an excess of zeroes, spatial and temporal dependence. Motivated by a forestry study of Comandra blister rust (CBR) infection of lodge pole pine trees from British Columbia, Canada, we develop a class of zero-inflated models for analyzing zero inflated count data. The model consists of two components to compare the abundance of trees that are resistant to CBR infection and the right skewed count of lesions on each tree from CBR infection. The model incorporates a series of predictors, as well as spatially and temporally correlated random effects for each model component. The ran-

dom effect terms are linked to induce the dependence of the two components and also to provide spatial and temporal smoothing. Modeling and inference use the fully Bayesian approach via Markov Chain Monte Carlo (MCMC) simulation approaches.

e-mail: **cindy.feng@usask.ca**

A SPATIO-TEMPORAL APPROACH TO MODELING SPATIAL COVARIANCE

Ephraim M. Hanks*, The Pennsylvania State University

Spatially-correlated data can often be viewed as being generated by a spatio-temporal process. We illustrate how a potential spatio-temporal generating process can motivate spatial statistical models, providing a broad framework for modeling spatial covariance functions. We link spatio-temporal generating processes to the Matern class of spatial covariance functions, intrinsic conditional autoregressive (ICAR) models, and others. We present a continuous-time Markov process on a spatial graph thathas a spatial random fields with ICAR structure as its stationary distribution, and consider generalizations that allow for principled specification of ICAR precision matrices based on existing knowledge of the system. We illustrate the utility of this approach through an example of spatial modeling on a stream network.

e-mail: **hanks@psu.edu**

INCORPORATING COVARIATES IN DETERMINISTIC ENVIRONMENTAL MODELS

Edward L. Boone*, Virginia Commonwealth University

Ben Stewart-Koster, Australian Rivers Institute at Griffith University

Environmental models have been developed by both the statistical and mathematical communities, which are very good at capturing the complex behavior found in nature. While statistics has ventured into estimating parameters in deterministic models, not much work has been done to combine the two approaches in a more unified way. In this talk we present a Bayesian method to incorporate covariates into deterministic models. To estimate the model parameters MCMC techniques will be used. The method will be illustrated using the simple Lotka-Volterra predator-prey model. We will also present how to incorporate spatial correlation into these models.

e-mail: **elboone@vcu.edu**

116. Bayesian and Non-Parametric Bayesian Approaches to Causal Inference

A FRAMEWORK FOR BAYESIAN NONPARAMETRIC INFERENCE FOR CAUSAL EFFECTS OF MEDIATION

Chanmin Kim, Harvard University

Michael J. Daniels*, University of Texas, Austin

Jason Roy, University of Pennsylvania

Except in very simple situations, untestable (from the observed data) assumptions need to be made for drawing causal inferences. Similar to missing data problems, the problem can be partitioned into two components: 1) a model for the observed data; 2) a set of (reasonable) assumptions that allow identification and estimation of causal estimands given the observed data. Given that the second component is not checkable from the observed data, uncertainty about these assumptions is essential for a fair characterization of the uncertainty. We contend that these two components can be handled most naturally in the Bayesian paradigm using flexible Bayesian nonparametric (BNP) models for the observed data and assumptions with sensitivity parameters that can be identified with informative priors. BNP models will provide similar robustness to semi-parametric approaches. We provide an illustration of this approach in the setting of the causal effect of mediation.

e-mail: **mjdaniels@austin.utexas.edu**

A BAYESIAN NONPARAMETRIC CAUSAL MODEL FOR REGRESSION DISCONTINUITY DESIGNS

George Karabatsos*, University of Illinois, Chicago

Stephen G. Walker, University of Texas, Austin

A regression discontinuity design (RDD) is a non-randomized design where treatment (versus non-treatment) assignment to a subject depends on whether or not her/his value of the assignment vari-

able crosses a known threshold. Under relatively mild conditions, the RDD can identify and estimate causal effects for the subgroup of subjects located in a neighborhood around the threshold, as if treatments are randomly assigned to those subjects. However, the accurate estimation of causal effects still relies on a correctly-specified statistical model. Also, in applications, it may be of interest to infer causal effects in terms of general features of the outcome variable distribution, not only the mean. For RDDs, we propose a flexible Bayesian nonparametric regression model that can provide accurate estimates of causal effects, in terms of the predictive mean, variance, quantile, probability density, distribution function, or any other chosen function of the outcome variable. The model allows the entire distribution of the outcome variable to change flexibly as a function of predictors, and can be extended to handle multivariate assignment variables. We illustrate the model through the analysis of two real data sets, involving (resp.) a sharp RDD and a fuzzy RDD. Free user-friendly software is available for the model.

e-mail: **gkarabatsos1@gmail.com**

EVALUATING THE EFFECT OF UNIVERSITY GRANTS ON STUDENT DROPOUT: EVIDENCE FROM A REGRESSION DISCONTINUITY DESIGN USING BAYESIAN PRINCIPAL STRATIFICATION ANALYSIS

Fan Li*, Duke University

Alessandra Mattei, University of Florence

Fabrizia Mealli, University of Florence

Regression discontinuity (RD) designs are often interpreted as local randomized experiments: a RD design can be considered as a randomized experiment for units with a realized value of a so-called forcing variable falling around a pre-fixed threshold. Motivated by the evaluation of Italian university grants, we consider a fuzzy RD design where the receipt of the treatment is based on both eligibility criteria and a voluntary application status. Resting on the fact that grant application and grant receipt statuses are post-assignment (post-eligibility) intermediate variables, we use the principal stratification framework to define causal estimands within the Rubin Causal Model. We propose a probabilistic formulation of the assignment mechanism underlying RD designs, by re-formulating the Stable Unit Treatment Value Assumption (SUTVA) and making an explicit local overlap assumption for a subpopulation around the threshold. A local randomization assumption is invoked instead of standard continuity assumptions. We also develop a model-based Bayesian approach to select the target subpopulation(s) with adjustment for multiple comparisons, and to draw inference for the target causal estimands in this framework. Applying the method to the data from two Italian universities, we find evidence that university grants are effective in preventing students from low-income families from dropping out of higher eduction.

e-mail: **fli@stat.duke.edu**

BAYESIAN NONPARAMETRIC ESTIMATION FOR DYNAMIC TREATMENT REGIMES WITH SEQUENTIAL TRANSITION TIMES

Yanxun Xu*, University of Texas, Austin

Peter Mueller, University of Texas, Austin

Abdus S. Wahed, University of Pittsburgh

Peter F. Thall, University of Texas MD Anderson Cancer Center

Dynamic treatment regimes in oncology and other disease areas are often characterized by an alternating sequence of treatments or other actions and transition times between disease states. The sequence of transition states may vary substantially from patient to patient, depending on how the regime plays out, and in practice there often are many possible counterfactual outcome sequences. For evaluating the regimes, the mean final overall time may be expressed as a weighted average of the means of all possible sums of successive transitions times. A common example arises in cancer therapies where the transition times between various sequences of treatments, disease remission, disease progression, and death characterize overall survival time. For the general setting, I propose estimating mean overall outcome time by assuming a nonparametric Bayesian survival regression for the transition times. I construct a dependent Dirichlet process prior with Gaussian process base measure (DDP-GP). I summarize the joint posterior distribution by Markov chain Monte Carlo (MCMC) posterior simulation. Then I use likelihood-based G-estimation under the DDP-GP model to estimate causal inference by accounting for all possible outcome paths, the transition times

between successive states, and effects of covariates and previous outcomes, on each transition time. The Bayesian paradigm works very well, and the simulation studies suggest that our DDP-GP method yields more reliable estimates than inverse probability of treatment weighted (IPTW) method.

e-mail: **yanxunxu.stat@gmail.com**

117. Design of Multiregional Clinical Trials: Theory and Practice

RANDOM EFFECTS MODELS FOR MULTIREGIONAL CLINICAL TRIAL DESIGN AND ANALYSIS

Gordon Lan*, Janssen Research & Development

In recent years, developing pharmaceutical products via a multiregional clinical trial (MRCT) has become more popular. Many studies with proposals on design and evaluation of MRCTs under the assumption of a common treatment effect across regions have been reported in the literature. However, heterogeneity among regions causes concern that the fixed effects model for combining information may not be appropriate for MRCT. In this presentation, we discuss the use of a continuous random effects model, and a discrete random effects model for the design and evaluation of MRCTs. Many numerical examples will be provided to illustrate the fundamental differences between these two random effects approaches.

e-mail: **glan@its.jnj.com**

CONSISTENCY OF TREATMENT EFFECT IN MULTIREGIONAL CLINICAL TRIALS

Joshua Chen*, Sanofi Pasteur

Global clinical development strategy utilizing multi-regional clinical trials (MRCTs) plays a crucial role in developing innovative medicines. It is readily accepted that studying patients from many different regions within a single trial under a single protocol is an efficient method of trial design. The prevalence of these trials has been growing over the last few decades. MRCTs are most often conducted as a single trial focusing on the overall results, but when such trials are submitted to health authorities, the scope and concern often broaden to include the "local" results. In this presentation, I will discuss specific MRCT concerns at the design stage, methods for assessing consistency of treatment effect across regions and sample size planning. Case studies will be presented and the methods introduced will be applied to those case studies.

e-mail: **josh.chen@sanofipasteur.com**

118. CONTRIBUTED PAPERS: Multivariate Survival Analysis

A SIEVE SEMIPARAMETRIC MAXIMUM LIKELIHOOD APPROACH FOR REGRESSION ANALYSIS OF BIVARIATE INTERVAL-CENSORED FAILURE TIME DATA

Qingning Zhou*, University of Missouri

Tao Hu, Capital Normal University

Jianguo Sun, University of Missouri

Interval-censored failure time data arise in a number of fields and many authors have discussed various issues related to their analysis. However, most of the existing methods are for univariate data and there exists only limited research on bivariate data, especially on regression analysis of bivariate interval-censored data. We present a class of semiparametric transformation models for the problem and for inference, a sieve maximum likelihood approach is developed. The model provides a great flexibility, in particular including the commonly used proportional hazards model as a special case, and in the approach, Bernstein polynomials are employed. The strong consistency and asymptotic normality of the resulting estimators of regression parameters are established and furthermore, the estimators are shown to be asymptotically efficient. Extensive simulation studies are conducted and indicate that the proposed method works well for practical situations. Also an illustrative example with data from an AIDS study is provided.

e-mail: **qz4z3@mail.missouri.edu**

METHODS FOR CONTRASTING GAP TIME HAZARD FUNCTIONS

Xu Shu*, University of Michigan

Douglas E. Schaubel, University of Michigan

Times between successive events (i.e., gap times) are often of interest in clinical and epidemiologic studies. While many methods exist for estimating the effect of covariates on each gap time, relatively few methods have targeted comparisons between the gap times themselves. Motivated by the comparison of primary

and repeat organ transplantation, our interest is specifically in comparing the gap-time-specific hazard functions. We propose a two-stage procedure, wherein the first stage involves a Cox regression model on the first gap time. Weighted estimating equations are then solved at the second stage to compare the first and second gap time hazard functions. Large-sample properties are derived, with simulation studies carried out to evaluate finite-sample performance. We apply the proposed methods to kidney transplant data obtained from a national organ transplant registry.

e-mail: **shuxu@umich.edu**

USING FULL COHORT INFORMATION TO IMPROVE THE EFFCIENCY OF MULTIVARIATE MARGINAL HAZARD MODEL FOR CASE-COHORT STUDIES

Hongtao Zhang*, University of North Carolina, Chapel Hill

Jianwen Cai, University of North Carolina, Chapel Hill

Haibo Zhou, University of North Carolina, Chapel Hill

David Couper, University of North Carolina, Chapel Hill

The case-cohort design is widely used in large cohort studies when it is prohibitively costly to measure some exposures for all subjects in the full cohort, especially in studies where the disease rate is low. To investigate the effect of a risk factor on different diseases, multiple case-cohort studies using the same subcohort are usually conducted. To compare the effect of a risk factor on different types of diseases, times to different disease events need to be modeled simultaneously. Existing case-cohort estimators for multiple disease outcomes utilize only the relevant covariate information in cases and subcohort controls, though many covariates are measured for everyone in the full cohort. Intuitively, making full use of the relevant covariate information can improve efficiency. To this end, we consider a class of doubly-weighted estimators for both regular and generalized case-cohort studies with multiple disease outcomes. The asymptotic properties of the proposed estimators are derived and our simulation studies show that a gain in efficiency can be achieved with a properly chosen weight function. We illustrate the proposed method with a data set from Atherosclerosis Risk in Communities (ARIC) study.

e-mail: **squallteo@gmail.com**

MARGINAL MODELS FOR RESTRICTED MEAN SURVIVAL WITH CLUSTERED TIME TO EVENT DATA USING PSEUDO-VALUES

Brent R. Logan*, Medical College of Wisconsin

Kwang Woo Ahn, Medical College of Wisconsin

Many time-to-event studies are complicated by the nesting of individuals within a cluster, such as patients in the same center in a multi-center study. These clustered data can be readily handled within the Cox model framework; however, when the proportional hazards assumption is violated, the hazard ratio is not easily interpretable. Restricted mean survival is an alternative summary measure that has a useful clinical interpretation as the expected life years up to a pre-specified time point. While a number of techniques have been described for modeling restricted mean survival with independent observations, little work has been done for clustered data. We apply pseudo-value regression to the clustered data framework, and show that it provides a marginal model for the restricted mean survival parameter. We compute leave one out pseudo-observations from estimates of the restricted mean survival. These are used in a generalized estimating equation to model the marginal restricted mean survival, and obtain consistent estimates of the model parameters. The method is easy to implement using standard software once the pseudo-values are obtained, and simulation studies show that the method has good operating characteristics. We illustrate the method using a bone marrow transplantation example.

e-mail: **blogan@mcw.edu**

SEMI-PARAMETRIC MODELING OF BIVARIATE RECURRENT EVENTS

Jing Yang*, Emory University

Limin Peng, Emory University

Recurrent events are frequently observed in biomedical studies, and often they consist of more than one type of events of interests. Marginal analysis of each type of recurrent event is useful but cannot address questions on the relationship between different types of recurrent events. In this work, we study a dynamic association model that extends a recently developed quantile association model. Our estimating equations are constructed based on the stochastic processes

embedded with bivariate recurrent events data. The proposed estimation can be implemented by an efficient and stable algorithm. We investigate the asymptotic properties of the proposed estimator, and develop proper inference procedures. Our proposals are illustrated via simulation studies and an application to a registry dataset.

e-mail: **jyang89@emory.edu**

ANALYSIS OF A COMPOSITE ENDPOINT UNDER DIFFERENT CENSORING SCHEMES FOR COMPONENT EVENTS VIA MULTIPLE IMPUTATION

Yuqi Chen*, University of California, Santa Barbara

Chunlei Ke, Amgen Inc.

Jianming Wang, Celgene Corporation

A composite endpoint is often used as the endpoint of primary interest for various reasons in clinical trials. It may happen that the component events are monitored or collected in different ways, thus potentially leading to different censoring schemes among the components. It becomes challenging to define the time variable for the composite endpoint to be used for analysis. Some ad-hoc methods are used by imputing or defining the time variable based on that of the component events, which may be inefficient or involve some assumptions. In this article, we propose three multiple imputation based methods under a monotone censoring scheme: to impute the event time marginally using Kaplan-Meier estimates; to impute based on a Cox proportional hazard model; and to impute the event time of one component event based on a Kaplan-Meier estimate conditional on the other event. Inference procedures are developed for estimating survival distribution, quantile survival time, comparing survival distributions, and estimating covariate effects. We compare the proposed methods with some ad-hoc methods through simulations. Simulation results show that these multiple imputation methods perform well consistently while the performance of ad-hoc methods depends on simulation settings. We also apply the proposed methods to a real data example.

e-mail: **ychen@pstat.ucsb.edu**

QUANTILE REGRESSION FOR SURVIVAL DATA WITH DELAYED ENTRY

Boqin Sun*, University of Massachusetts, Amherst

Jing Qian, University of Massachusetts, Amherst

Delayed entry arises frequently in follow-up studies for survival outcomes, where additional study subjects enter during the study period. We propose a quantile regression model to analyze survival data subject to delayed entry and right-censoring. Such a model offers flexibility in assessing covariate effects on survival outcome and the regression coefficients are interpretable as direct effects on the event time. Under the conditional independent censoring assumption, we proposed a weighted martingale-based estimating equation, and formulated the solution finding as a L1-type convex optimization problem, which was solved through a linear programming algorithm. We established uniform consistency and weak convergence of the resultant estimators. We developed and justified a resampling inference procedure for variance and covariance estimation. The finite-sample performance of the proposed method was demonstrated via simulation studies. The proposed method was illustrated through an application to a clinical study.

e-mail: **boqinsun@gmail.com**

119. CONTRIBUTED PAPERS: Constrained Inference

ORDER STATISTICS FROM LINDLEY DISTRIBUTION AND THEIR APPLICATIONS

Khalaf S. Sultan*, College of Science King Saud University, Saudi Arabia

Wafaa S. AL-Thubyani, College of Science King Saud University, Saudi Arabia

We derive the exact expressions for the single and product moments of order statistics from Lindley distribution. Then, we use these moments to obtain the best linear unbiased estimates of the location and scale parameters (BLUEs) based on Type-II censoring. Also, we use utilize the single and product moments to develop the correlation goodness-of-fit test of the Lindley distribution. In addition, we calculate the power of the test based on some alterative distributions. In order to show the usefulness of the findings of the paper we carry out some Monte Carlo simulations. Finally, we discuss some applications based on real data sets.

e-mail: **ksultan@ksu.edu.sa**

CLME: A TOOL FOR INFERENCE IN LINEAR MIXED EFFECTS MODELS UNDER INEQUALITY CONSTRAINTS

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

In many applications researchers are interested in testing for inequality constraints in the context of linear fixed effects and mixed effects models. For example, a researcher may wish to test for an increasing response over increasing dose levels. Popular procedures such as ANOVA only test for differences; estimation subject to linear inequality constraints can often yield greater power to detect an effect. Furthermore, while there exists a large body of literature for performing statistical inference for linear models subject to inequality constraints, user friendly statistical software for implementing such methods is lacking. We develop a package in the R language, CLME, which can be used for testing a broad collection of inequality constraints. It uses residual bootstrap based methodology which is reasonably robust to non-normality as well as heteroscedasticity. The package contains a graphical interface built using the shiny package, enabling a researcher with minimal knowledge of R to easily take advantage of the methods implemented in CLME. We illustrate the package using real-world datasets, and demonstrate the graphical interface.

e-mail: **casey.jelsema@nih.gov**

ORDER-CONSTRAINED BAYESIAN NONPARAMETRIC MODELING OF CORRELATED THREE-WAY ROC SURFACES

Beomseuk Hwang*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Zhen Chen, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

In many multiple diagnostic tests, some a priori constraints may exist. Also, the true disease often has more than two levels, e.g., stages of endometriosis. In this case, traditional binary diagnostic measures such as ROC curve and AUC need to be extended to handle three- or more- classification problem assuming the constraints. We propose a nonparametric Bayesian joint modeling framework for three-dimensional ROC surfaces that accounts for stochastic and variability orders. The stochastic order constrains the distributional centers of the three disease populations within each test, while the variability order constrains the distributional spreads of the multiple tests within each of three populations. We demonstrate the performance of the proposed approach using data from the Physician Reliability Study that investigated the accuracy of diagnosing endometriosis. To address the issue of no gold standard in the real data, we use a sensitivity analysis approach that exploited diagnostic results from a panel of experts.

e-mail: **beomseuk.hwang@nih.gov**

PARTIAL LIKELIHOOD ESTIMATION OF ISOTONIC PROPORTIONAL HAZARDS MODELS

Yunro Chung*, University of North Carolina, Chapel Hill

Anastasia Ivanova, University of North Carolina, Chapel Hill

Michael Hudgens, University of North Carolina, Chapel Hill

Jason Fine, University of North Carolina, Chapel Hill

We consider estimation of the semiparametric proportional hazards model with a completely unspecified baseline hazard function where the effect of a continuous covariate is assumed monotone but otherwise unspecified. Previous work on full nonparametric maximum likelihood estimation for isotonic Cox proportional hazard regression with right censored data is computationally intensive, lacking theoretical justification, and may be prohibitive in large samples. We study partial likelihood estimation. An iterative quadratic programming method (IQM) is proposed, with theoretically justified convergence properties. However, unlike with likelihoods for isotonic parametric regression models, the IQM for the partial likelihood cannot be implemented using standard pool adjacent violators techniques, increasing the computational burden. An alternative pseudo iterative convex minorant algorithm (PICM) is

presented which exploits such techniques and is also theoretically justified. The algorithms are extended to models with time-dependent covariates. Analysis of real data illustrates the practical utility of the isotonic methodology in estimating nonlinear covariate effects.

email: **yunro.roy@gmail.com**

NONPARAMETRIC TESTS OF UNIFORM STOCHASTIC ORDERING

Chuan-Fa Tang*, University of South Carolina

Joshua M. Tebbs, University of South Carolina

Dewei Wang, University of South Carolina

We derive nonparametric procedures for testing for and against uniform stochastic ordering in the two-population setting with continuous distributions. We account for this ordering by examining the least star-shaped majorant of the ordinal dominance curve formed from the nonparametric maximum likelihood estimators of the continuous distribution functions F and G. In particular, we focus on testing equality of F and G versus uniform stochastic ordering and testing for a violation of uniform stochastic ordering. For both testing problems, we propose a family of Lp norm statistics, derive appropriate limiting distributions, and provide simulation results that characterize the performance of our procedures. We illustrate our methods using data from a study involving premature infants and the occurrence of necrotizing enterocolitis.

email: **tang9@email.sc.edu**

COVARIATE BALANCED RESTRICTED RANDOMIZATION: OPTIMAL DESIGNS, EXACT TESTS, AND ASYMPTOTIC PROPERTIES

Jingjing Zou*, Columbia University

Jose R. Zubizarreta, Columbia University

In randomized experiments, the act of randomly assigning units to treatment (i) physically induces a distribution that can be used for exact testing and (ii) ensures that both observed and unobserved covariates are balanced across the treatment groups in expectation. However, in a given realization of the random assignment mechanism, the covariates may exhibit considerable imbalances due to chance, especially if the experiment has a small number of units. To address this limitation while explicitly using the randomization distribution for inference, in this paper we propose a new method for achieving strong forms of balance on the observed covariates and develop a procedure for conducting exact inferences based on a covariate balanced restricted randomization distribution. We derive asymptotic results for the procedure and illustrate its use on an important randomized experiment that was used for targeting the poor in Indonesia. It is demonstrated through both theoretical results and simulation studies that the proposed method outperforms unrestricted methods with higher power.

email: **jz2335@columbia.edu**

120. CONTRIBUTED PAPERS: Nonparametric Methods

NONPARAMETRIC AND SEMIPARAMETRIC ESTIMATION IN MULTIPLE COVARIATES

Richard Charnigo*, University of Kentucky

Limin Feng, Intel Corporation

Cidambi Srinivasan, University of Kentucky

We consider the problem of simultaneously estimating a mean response function and its partial derivatives, when the mean response function depends nonparametrically on two or more covariates. To address this problem, we propose a "compound estimation" approach, in which differentation and estimation are interchangeable: an estimated partial derivative is exactly equal to the corresponding partial derivative of the estimated mean response function. Compound estimation yields essentially optimal convergence rates and exhibits substantially smaller squared error in finite samples compared to local regression. We also explain how to employ compound estimation under more general circumstances, when the mean response function depends parametrically on some additional covariates and the observations are not statistically independent. In a case study, we apply compound estimation to examine how the progression of Parkinson's disease may relate to a subject's age and the

signal fractal scaling exponent of the subject's recorded voice. Especially among those intermediate in age, an abnormal signal fractal scaling exponent may portend greater symptom progression.

email: **RJCharn2@aol.com**

NONPARAMETRIC EMPIRICAL BAYES VIA MAXIMUM LIKELIHOOD FOR HIGH-DIMENSIONAL CLASSIFICATION

Lee H. Dicker, Rutgers University

Sihai D. Zhao, University of Illinois, Urbana-Champaign

Long Feng*, Rutgers University

Nonparametric empirical Bayes methods are naturally suited for many problems in high-dimensional statistics. The Kiefer-Wolfowitz (1956) nonparametric maximum likelihood estimator (NPMLE) for mixture models provides an elegant approach to some of these problems. However, implementation and theoretical analysis of the Kiefer-Wolfowitz NPMLE are notoriously difficult. Recently, Koenker and Mizera (2013) proposed a fast method for approximately computing the Kiefer-Wolfowitz NPMLE based on convex optimization. This computational breakthrough has greatly simplified the application of NPMLE-based empirical Bayes methods. In this talk, we propose some novel applications of the Kiefer-Wolfowitz NPMLE to high-dimensional classification problems. In simulated data, these methods dramatically outperform well-known alternative methods (by an order of magnitude, in some instances); in real genomic data analyses, the NPMLE methods appear to be very competitive. Theoretical results will also be discussed. This is joint work with Sihai Dave Zhao and Long Feng.

email: **felixfeng.2009@gmail.com**

NONPARAMETRIC INFERENCE FOR AN INVERSE-PROBABILITY-WEIGHTED ESTIMATOR WITH DOUBLY TRUNCATED DATA

Xu Zhang*, University of Mississippi Medical Center

Efron and Petrosian (1999) formulated the problem of double truncation and proposed nonparametric methods on testing and estimation. An alternative estimation method was proposed by Shen (2010), utilizing the inverse-probability-weighting technique. One aim of this paper was to assess the computational complexity of the existing estimation methods. Through a simulation study we found that these two estimation methods have the same level of computational efficiency. The other aim was to study the non-iterative IPW estimator under the condition that the truncation variables are independent. The IPW estimator and the interval estimatoin was proved satisfactory in the simulation study.

email: **xzhang2@umc.edu**

A TEST FOR DIRECTIONAL DEPARTURE FROM LOEWE ADDITIVITY

Mingyu Xi*, University of Maryland, Baltimore County

In assessing the impact of exposure to chemical mixtures, scientists often assume simplified models where the combined effect of chemicals is the sum of individual effects. One such assumption is Loewe additivity. However, in practice this is often violated due to positive interaction (synergistic interaction) or negative interaction (antagonistic interaction). If the combined effect of mixture is more potent than the simple sum of individual effects, the interaction is positive; less potent, the interaction is negative. Only parametric model based tests are available in the literature for testing directional interaction. However, models for less analyzed mixtures may not be readily available and hence be grossly mis-specified, compromising the power of the test. Based on the observed contour profiles, we propose a novel nonparametric test for directional interaction. The test is shown to be robust and is applied to the motivating example of testing for directional interaction among common battery waste chemicals such as Nickel, Cadmium and Chromium.

email: **mxi1@umbc.edu**

ESTIMATION AND CONFIDENCE BANDS FOR NONPARAMETRIC REGRESSION WITH FUNCTIONAL RESPONSES AND MULTIPLE SCALAR COVARIATES

Andrada E. Ivanescu*, Montclair State University

The proposed nonparametric functional regression methodology accounts for a nonlinear dependence involving several scalar predictors in a modeling approach that explains changes in responses that consist of functional sampled data. The goal is to achieve adaptive inference and a pathway is to establish thresholded estimators and construct nonparametric confidence bands through a sparse estimation strategy developed using a data-supplied approach to determine the threshold levels. Applications to data analysis and simulations display results that show optimal implementations.

email: **ivanescua@mail.montclair.edu**

NONPARAMETERIC BAYESIAN ANALYSIS OF THE 2 SAMPLE PROBLEM WITH CENSORING

Kan Shang*, University of Minnesota

Cavan Sheerin Reilly, University of Minnesota

Testing for differences between 2 groups is a fundamental problem in statistics and due to developments in Bayesian nonparametrics and semiparametrics there has been renewed interest in approaches to this problem. Here we describe a new approach to developing such tests and introduce a class of such tests that takes advantage of developments in Bayesian nonparametric computing. This class of tests use the connection between the Dirichlet process prior and the Wilcoxon rank sum test but extends this idea to the mixture of Dirichlet process model. Given consistency results for this class of models we develop tests that have appropriate frequentist sampling procedures for large samples but have the potential to outperform the usual frequentist tests. Extensions to interval and right censoring are considered and an application to a high dimensional data set obtained from a RNA-Seq investigation demonstrates the practical utility of the method.

email: **shang050@umn.edu**